TANGENT

DATA-SHARING GOVERNANCE MODEL

D2.2

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 955273

Deliverable administrative information

Deliverable number	D 2.2
Deliverable title	Data-sharing governance model
Dissemination level	Public
Submission deadline	31/10/2022
Version number	1.0
Authors	Marco Comerio (Cefriel) Antonia Azzini (Cefriel) Mario Scrocca (Cefriel) Andrea Fiano (Cefriel) Sabino Metta (Cefriel)
Internal reviewers	Tiago Dias (A-to-Be) Athina Tympakianaki (Aimsun)
Document approval	Antonio David Masegosa Arredondo (Deusto) Ynte Vanderhoydonc (IMEC) Eleni Mantouka (NTUA)

Legal Disclaimer

https://twitter.com/TANGENT_H2020

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

https://www.youtube.com/channel/UCjhz4kwEm_sTHj7fE4zXToA

For further information please visit http://www.tangent-h2020.eu/

Executive summary

The sharing of data needed to enable the development and testing of the TANGENT services for multimodal transport network management must be performed in accordance with a structured and effective data governance model.

This deliverable provides the **TANGENT Data-Sharing Governance model** for managing data-sharing in the TANGENT architecture. The model identifies key processes to be addressed with the roles of the different stakeholders involved and the rules to apply.

The model is the result of three propaedeutic activities described in this deliverable:

- Analysis of the literature related to data governance to identify the necessary structures and processes to handle data-sharing;
- Analysis of the literature related to data governance in the mobility and transport domain to identify guidelines and best practices;
- Analysis of the FAIR data principles to identify rules and behaviors aimed at improving and maximizing the findability, accessibility, interoperability, and re-use of the shared data.

To support the adoption of the TANGENT Data-sharing Governance model, TANGENT will define a data catalogue, i.e., a digital platform supporting the sharing, findability, and accessibility of distributed digital assets. This deliverable discusses the **TANGENT Metadata Profile** for describing datasets in the TANGENT Data Catalogue. The profile, defined as an extension of DCAT-AP and compliant with the FAIR data principles, aims to increase and simplify data-sharing, discovery and reusability.

Keywords

Data-sharing, data governance, metadata, FAIR principles, data catalogue

Table of contents

List of figures

List of tables

1 Introduction

This chapter introduces the objectives of the deliverable and provides an overview of the results presented.

1.1 Attainment of the objectives and explanation of deviations

This deliverable aims to provide the governance model (including the necessary structures and processes) for the strategic and operational management of data-sharing within the TANGENT project. The model focuses on the tasks/processes that are under the responsibility of TANGENT WP2:

- definition of the TANGENT Data Catalogue;
- provision of access to data sources needed by the technical components of the project;
- development of the TANGENT solution for data harmonisation and fusion.

Three propaedeutic activities, described in this deliverable, have been carried out with the scope of formalizing the TANGENT Data-Sharing Governance Model:

- analysis of the literature related to data governance to identify a reference schema covering the necessary structures needed to handle data-sharing properly;
- analysis of the literature related to data governance in the mobility and transport domain to identify best practices to be adopted in the TANGENT project;
- analysis of the FAIR data principles [35] to identify rules and behaviors aimed at improving and maximizing the findability, accessibility, interoperability, and re-use of the data shared within the TANGENT project.

This deliverable contributes to the TANGENT project by providing:

- the **TANGENT Metadata Profile** for the description of datasets in the TANGENT Data Catalogue. The profile, defined as an extension of DCAT-AP [11] and compliant with the FAIR data principles, aims to increase and simplify data-sharing, discovery and reusability.
- the **TANGENT Data-Sharing Governance model** for managing data-sharing in the TANGENT architecture. The model identifies key processes to be managed together with the involved roles for actors, and the rules to be applied. The TANGENT Data-Sharing Governance model
  & & & & & & & & & &J?4&J &S &V .

1.2 Intended audience

This deliverable is mainly intended for project partners, project officers, and the public interested in data-sharing governance processes for the mobility and transport domain.

1.3 Structure of the deliverable and links with other work packages/deliverables

Figure 1 describes this deliverable s contents, contributions, and structure.

Section 2 provides an extensive analysis of the literature related to data governance (DG). The section starts by extracting a **Data Governance Reference Schema** from relevant papers in data governance literature covering the proper structures to handle data-sharing. The reference schema is then used to guide the analysis, comparison, and discussion of literature related to data governance in the mobility and transport domain and to identify **best practices** to be adopted in the TANGENT project.

Section 3 summarizes the FAIR data principles and related initiatives to improve the findability, accessibility, interoperability, and re-use of the data.

Section 4 reports the analysis of initiatives and best practices for defining metadata specification. Considering the results of the analysis, the section presents the defined **TANGENT Metadata Profile** for describing datasets in the TANGENT Data Catalogue.

Section 5 describes the activities performed to collect inputs from stakeholders during the multi-actor cooperation workshops organized by WP1 with each TANGENT case study city . &J 749&Multi-actor cooperation models for NTM. First Release). The input from stakeholders, together with the results of the analyses described in the previous sections, are used to define the main contribution of this deliverable: the **TANGENT Data-Sharing Governance model**.

Concluding remarks are reported in Section 6.

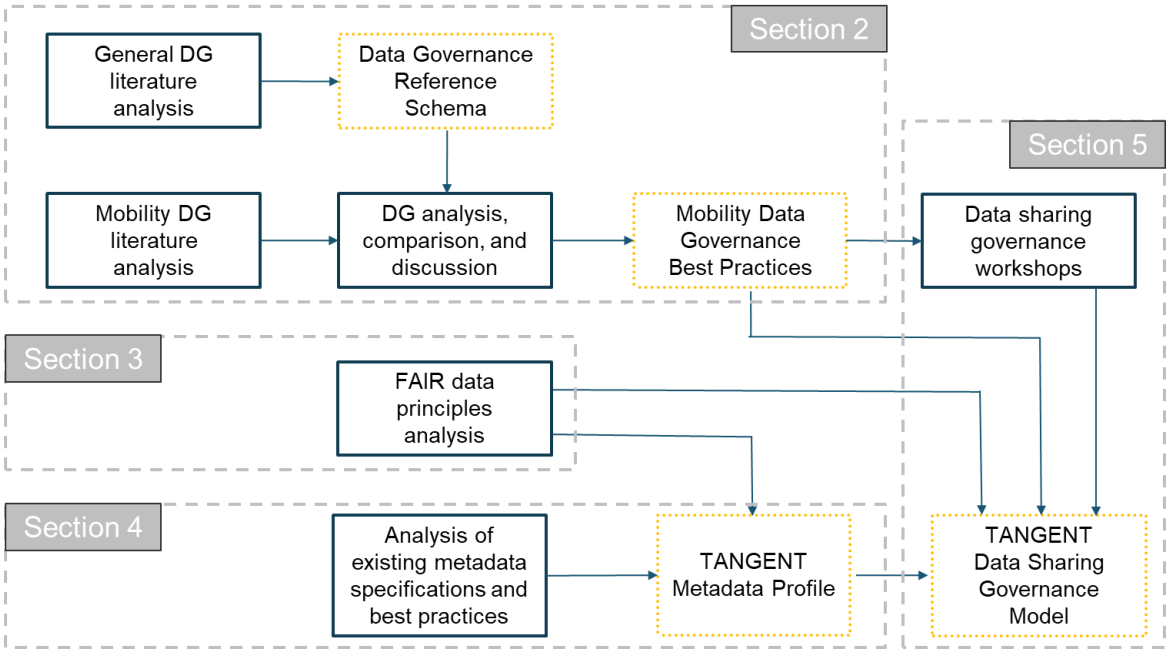

Figure 1: Contents and contributions of this deliverable

Figure 2 shows the activities described in this deliverable and the connections with other TANGENT WPs. The activities are related to Task 2.2 Definition of the data governance framework for data sharing and reusability & &the TANGENT WP2 Data Gathering, Harmonisation & Fusion 4&Task 2.2 has received inputs from WP1 (i.e., results of the analysis of system requirements within the multi-actor cooperation workshops), WP6 (i.e., technical specifications of the TANGENT solution), and its activities have been based on the & &Z &847& & &J 847&Data Requirements and Available Data Y .

The output of Task 2.2 (i.e., this deliverable) provides guidelines to WP6 for the implementation of the data governance in the TANGENT architecture, and will guide the activities of Task 2.3. Moreover, it & & & & & & & & & & &J ?4&J &S &V 4

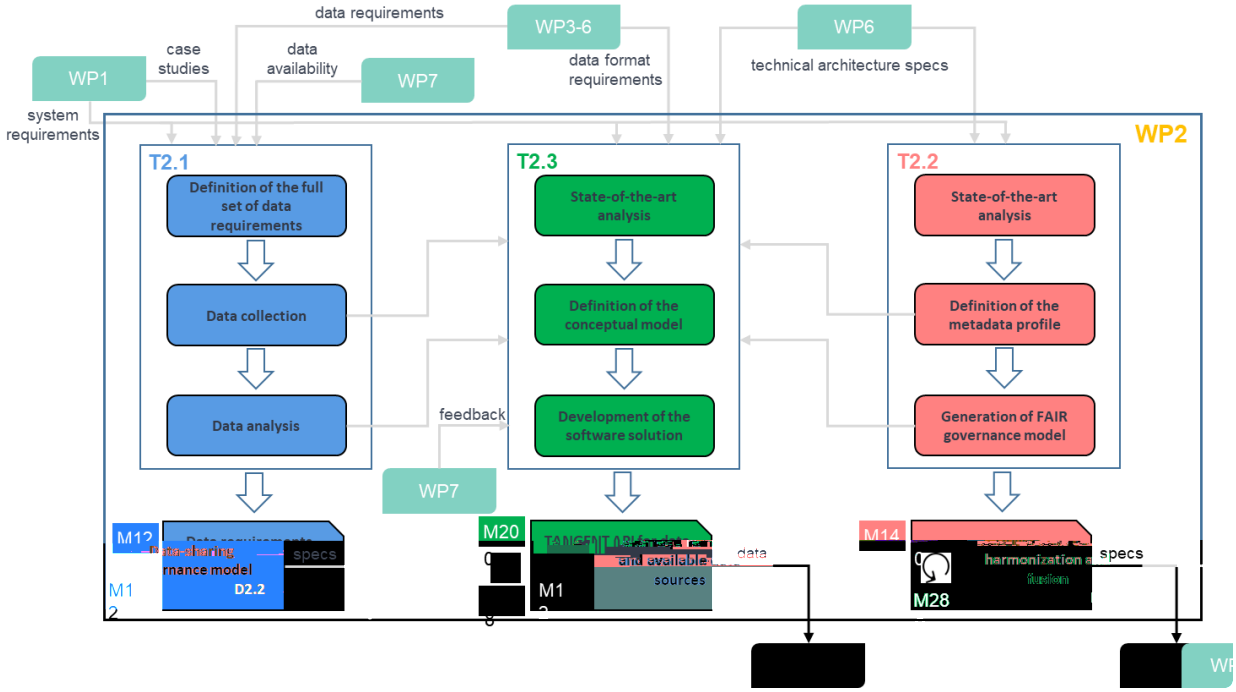

Figure 2: Connections between WP2 and other WPs

2 Data Governance: state-of-the-art and best practices

The European Strategy for Data [18] defines *data-sharing* as the process of making the same data resource accessible to numerous users or applications, and it provides indications for improving data availability, integration, and reuse. The ever-increasing volume of data held by public or private entities, organizations and stakeholders, in general, makes it necessary to define all the regulations to grant data-sharing. With this objective, the Data Governance Act (DGA) [21] has been stipulated, and by such a plan, the EU Parliament and Council signed an agreement concerning the law for information sharing. The DGA aims to increase all the benefits from data for European citizens and businesses by leading efficient public services, good quality and secure data handling, and more transparent data governance.

The continuous growth, especially in recent years, in the availability of data and the role they play underlines how they affect strategic and operational decisions. Consequently, it is crucial to define data governance that emphasizes the role of data and complies with current regulations, such as those on data privacy (e.g., the General Data Protection Regulation (GDPR[1])). At the same time, data governance is also crucial to fulfilling EU regulations related to mobility and transport data, such as Directive 2010/40 and all its supplementary delegated regulations (EU Delegated Regulations 886/2013 [14], 2015/962 [15], 1926/2017 [17]).

Several definitions of Data Governance (DG) are available in the literature. According to the work carried out by Abraham and colleagues [3], the DG specifies a cross-functional framework for managing data as a strategic enterprise asset. Such a definition allows Data Governance to specify rights and accountabilities for an organ & -making about its data. Furthermore, DG allows formalizing data policies, standards, procedures, and monitoring compliance.

From a general perspective, the goals of DG can therefore be summarized as follows:

- ensure that data meets the needs of the domain in which they are managed;
- protect and manage data as a critical asset;
- decrease the cost of managing data.

This section aims to collect, analyse and elaborate DG literature and provide guidelines and principles that will represent the pillars to follow for granting data collection, sharing, harmonisation, and usage within the TANGENT project.

2.1 A review of the data governance literature

Taking up the definition of DG defined by Abraham and colleagues [3], a first consideration concerns the fact that the authors divided DG into six essential aspects: (i) DG represents a cross-functional aspect, as it enables collaboration across different functional areas and data themes; (ii) DG provides the structure and formalization for data management; (iii) DG considers data a strategic enterprise resource; (iv) DG specifies decision rights and & & & & -making about its data; (v) DG develops policies, standards, and procedures regarding data; (vi) DG monitors compliance.

[1] https://gdpr-info.eu

Furthermore, the authors developed the conceptual framework for Data Governance into six dimensions:

1. The & & & & - & & & & 2&
   procedural, and relational mechanisms;
2. T & & & & & & & &J &M &
   and corresponds approximately to the unit of analysis;
3. T &J & & & & & & & & & ;
4. T & J & & & & & & & & & & & & & &
   mechanisms are applied. They include, for example, data quality, data security, data architecture, data lifecycle, metadata, and data storage and infrastructure;
5. T & namely the factors (e.g., strategic, organizational, and cultural aspects, regulatory requirements, etc.) that influence the adoption and implementation of Data Governance;
6. T & & & & & &J &M 4&

In the contribution presented by Panian [44], the focus has been initially posted on some critical issues arising from various organizations' management of data. Indeed, in some cases, data are managed in the context of how they serve a particular application rather than how they serve the strategic interests of the overall organization. As a result, governance practices are designed primarily around an & 2& & & & & 4&O& 2& & & & & & &
governance structure in place for their various data assets. This is compounded by continued data growth. Therefore, there is an increasing need to manage data as a corporate asset to be shared and reused across multiple applications, users, systems, and processes. This aspect comprises the need to define standards for data management and share policies and processes for the use, development, and management of such data by implementing an appropriate organizational structure.

Throughout the contribution presented by Panian [44], the principles of DG are defined and divided into four main aspects: they consider, respectively, (i) standards of the data schema, (ii) processes to be executed in compliance with policies, (iii) organizational aspects, and definition of roles for different actors involved in DG processes at different levels, and, finally, (iv) tools implementation by the emerging technology. The building blocks of the proposed DG model are based on the four main components, summarizing, respectively, the importance of (i) establishing a standard, (ii) enforcing policies and processes for the creation, development, control, sharing and management of data, (iii) defining an organizational structure and (iv) choosing the most appropriate technology. All these aspects emphasize a set of organizational and technological success factors that, according to the author, should be considered during a DG modelling process. For example, a clear definition of roles and specific responsibilities to individuals involved in Data Governance is helpful in enforcing accountability and the involvement of data ownership with the management staff and other partners. Further, the definition of a data integration infrastructure allows for making data accessible, available, of high quality, consistent, auditable, and secure.

Other important aspects presented by Panian [44] that can be considered in the definition of a data integration infrastructure may regard: (i) integrating data across different systems (e.g., managing different data formats in different data schemas), (ii) sharing data and checking data availability, (iii) managing data harmonisation, fusion, integration, and the processes useful to ensure reliability, performance, scalability, and (iv) the continuous monitoring of data quality through the implementation of dedicated metrics.

Through his contribution [41], Boris Otto provides a morphological picture of the DG organization. The author aims to propose a morphology as a guideline for adopting and implementing DG concepts whenever it is deemed useful in data management. Such a defined morphology subdivides DG by considering two main dimensions related to objectives and structure. The objectives are measurable and serve to assess the effectiveness of Data Governance, while the structure aims to set the main instances of responsibility for DG. The contribution also shows cases in which a shared responsibility represents the most suitable solution. The organizational form, defined through the organizational dimension, specifies the centralised or decentralised DG structure settings, while roles and committees (as the data owner, data stewards, sponsor, etc.) are defined by the third structure component. Critical issues arising from the considered use cases report that, although DG is considered to have a significant business impact on companies, it is usually not directly related to the business goals, but only to data quality and Service Level Agreements (SLAs). Similarly, more frequently, DG is considered only internally within the individual organization, rather than concerning external entities.

On the other hand, Brous and colleagues [5] pay particular attention to four main principles (organization, alignment, compliance, and common understanding) for DG to develop effective strategies and approaches. Taking up the objectives presented by Otto in [41], the definition of the **organization** in this contribution indicates: (i) the structure in which responsibilities are specified and assigned, and (ii) the structure of the DG process as the set of roles, rules, objectives, decision-making areas, and model definition for data handling. From an **alignment** perspective, this contribution indicates that a DG program should be able to demonstrate the value represented by the data it holds and shares, as an assurance of its usefulness is a good metric for establishing the appropriateness of specific policies. In this sense, a good DG also provides the framework for addressing complex issues, such as improving data quality dimensions (e.g., accuracy, completeness, consistency, and timeliness) or processes to be considered at the strategic level. The authors reiterate, in the **compliance** principle, the relevance of establishing data policies and procedures that can define the measure of importance associated with the data as such. However, such policies are also considered necessary to preserve and protect personal data and privacy in their own rights. Finally, the fourth and last principle suggested by the authors refers to a concept of **common understanding** of data and corresponding requirements, encouraging the definition of a proper data model. Such activity is, indeed, considered fundamental in the DG definition process, necessary to avoid misinterpretation of information and, at the same time, encourage the use of standard data schemas. Brous and colleagues conclude by indicating that there & & & 2& & & & & & & &J M& & 4& -making bodies need to be identified, and the model should have a formal organizational structure that fits specific needs.

This concept is also included in the contribution from Otto and colleagues [42], which is based on the study of governance aspects related to data quality and corresponding management. Indeed, the objective of this work is to present the main elements of DG in the context of Data Quality Management (DQM). The paper summarizes the main DQM approaches and, for each of them, DG roles, decision-making areas, and processes are reported, as well as the accountabilities indicated by the authors with & & & 4 The authors emphasize how the DG model addresses DQM on three horizontal layers: strategy, organization, and information systems. Finally, the assignment of responsibilities to roles in the Data Governance model follows the idea of a Responsibility Assignment Matrix (RAM) [54]. The author also reports one of the most popular types of RAM [53]: the RACI chart, used to define responsibilities. RACI is the acronym for the four types of interaction: Responsible, Accountable, Consulted, and Informed, which in [42] have been mapped into the DQM. The DG contingency model is also presented, explaining which contingencies (e.g., organizational strategy, size, structure) affect the fit between an organization-specific DG model and a successful DQM. In this case, the analysed contingency factors correspond to performance strategy, organization structure, competitive strategy, diversification width, process harmonisation, market regulation, and decision-making style.

2.1.1 Data governance reference schema

Considering the literature analysis, a reference schema of the main components that define a DG structure has been defined. Figure 3 provides a graphical representation, at a general level, of the main features.

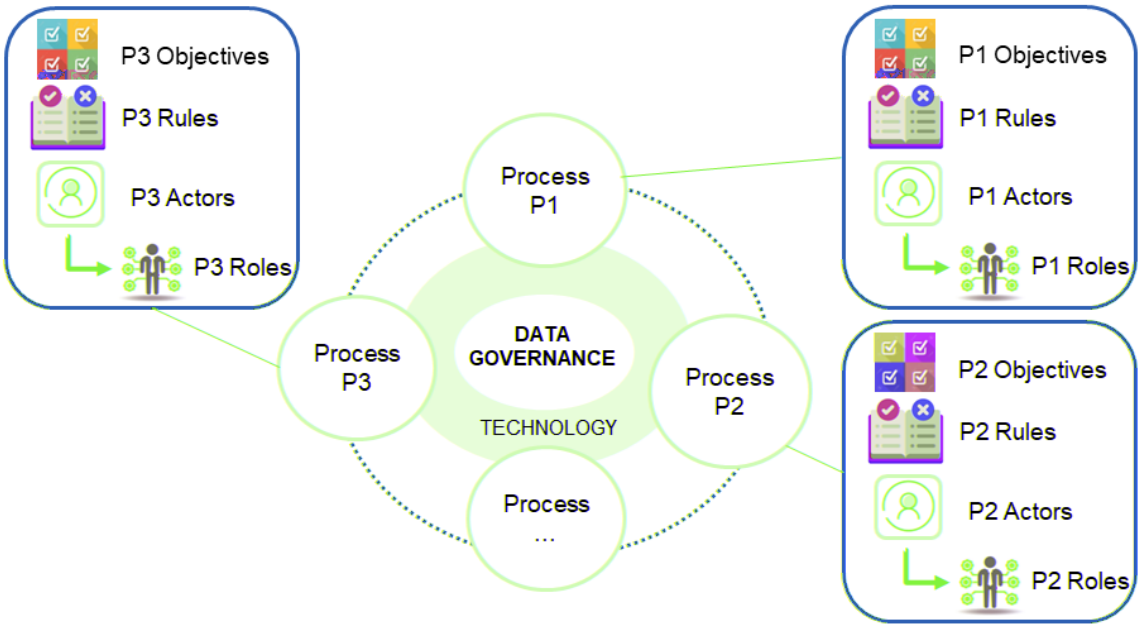

Figure 3: The data governance reference schema

Data governance involves different processes such as data collection, sharing, harmonisation, usage, and so on. For each considered process, different objectives are defined concerning, for example, the assurance of data quality, data accessibility, data security, ownership, compliance with European directives, regulations, standards, and others.

The different rules that must be considered in each process to achieve the objectives are also defined through the DG model. An example is represented by the set of rules that must be applied and observed for the data-sharing process, such as the acquisition of metadata, the filling of a descriptor, its validation, and so on. Rules can be categorized as (i) internal, when they refer to the regulation of activities specific to a given process or project, and (ii) external, when they refer to external regulations such as, for example, European Community regulations and policies that cover data issues.

The persons involved in the execution of the different processes represent all the significant actors that are part of a DG model. However, they may be different according to the domain and context in which DG is applied. Each actor involved may assume one or more roles (e.g., data publisher, data controller, data owner). A committee or an authority body may be defined to cover the role of the central DG authority. Then, more general roles often considered by a DG model are the data steering committee, the DG officer, the data owner, the data steward (who oversees aspects related in general to data management, e.g., format, accessibility, description, quality, etc.), and the technology steward (who may be in charge of technology availability, security and IT architecture availability). This last actor is defined to regulate the use of the different technological solutions and tools needed to perform the processes involved in a DG model.

2.2 Mobility Data Governance

Following the data governance reference schema (Figure 3), it is possible to identify key processes related to mobility DG, such as mobility data collection, sharing, and harmonisation. As previously reported, each process identifies the objectives it intends to achieve, the rules that must be compliant with, the actors involved in the process, and their roles. Transport service providers (TSP), municipalities, and local private mobility management organisations are examples of the actors of the mobility domain potentially involved in data governance. They cover roles related, for example, to data publisher, data controller, Data Protection Officer, responsible for the quality and security of the data shared, and others.

Table 1 reports examples of mobility data governance processes described according to the reference schema.

PROCESS	OBJECTIVES	RULES		ACTORS	ROLES
		Internal	External		
Data collection	Accessibility and acquisition of multimodal data, mobility sensor handling	Metadata acquisition and validation	European directives and guidelines	Transport authorities, operators, infrastructure managers	Data owner, controller
Data-sharing	Data security, accessibility, metadata quality	Metadata and data acquisition, validation, sharing	European directives and guidelines	Transport authorities, operators, infrastructure managers	Data publisher, controller, owner, consumer
Data quality management	Data quality, pre-processing	Data quality management, data validation and update	Mobility data standards, European guidelines	Transport authorities, operators, infrastructure managers	Data controller, publisher
Data conversion	Data harmonisation and integration, standardization, compliance	Reference model	European directives and guidelines, standards	Transport authorities, technical organizations	Data controller

Table 1: Examples of mobility data governance processes

A review of recent relevant projects covering the data governance topic is provided in the following section to acquire more information on how the DG processes mentioned in Table 1 could be realised.

2.2.1 A review of projects covering mobility data governance

The partners of the TANGENT project have been asked to share their knowledge about projects, initiatives, and publications related to data governance in the mobility and transport domain. An analysis of the collected digital ecosystem initiatives and European-funded research projects has highlighted interesting best practices and guidelines that will guide the definition of the TANGENT Data-sharing Governance model.

E015 Digital Ecosystem

The E015 Digital Ecosystem [13] project is funded by the Italian Lombardy Region. Started in 2021, it will end in 2023. It is developed for both public and private partnerships. The Digital Ecosystem is an API Economy initiative based upon a digita& & & & & &GVO& K67; &GVO /& & & & & K67; & /2& & & 4&

At the regional level, the E015 Digital Ecosystem has been included as an official way of promoting data transparency and data exchange between public and private players. Enabling APIs requires the definition of agreements on shared data models that must be commonly chosen, defining common architectural choices, and determining the service components that can be reused. The appropriate access to the APIs is supervised by a management board for preventing the exposure of individual data about citizens and businesses. The Governance structure helps to ensure standardization, address and manage risks, encourage interoperability, ensure adherence to broader government policy principles, and contribute to the rewiring of interactions among different actors.

The Digital Ecosystem definition refers to important aspects. In fact, it requires many different legacy systems to exchange data, and this is made possible by Application Programming Interfaces (APIs) that all parties adopt as a grounding rule. The most important role within the toolkit (composed of common rules, processes, and roles) is the governance exercised by the Ecosystem Management Board. In E015, the Board takes care of several aspects, like those listed below:

- operational governance, to manage processes and the online environment to describe, share and request enabling APIs;
- technical governance, including the technical interoperability standards to exchange data (for example, REST APIs);
- onboarding governance to promote sharing and usage of the e-API, to create valuable business scenarios;
- strategical governance to define the trajectory and main areas of interest for ecosystem development.

From a general point of view, in a digital ecosystem, participants can describe and publish their data assets, in terms of both functionalities and usage policies. This is done to let other participants discover and leverage them, in agreement with the respective usage policies, for obtaining deeper insight in the driving business KPI. The increased ability to combine information by exploiting individual data assets can improve the services offered to the consumers by different players.

MOBILITY DATA SPACE

The European strategy for data (2020) [17] aims to create a single data market & & &K & global competitiveness and data-sharing. For this reason, it will define interoperable data spaces in strategic economic sectors and domains of public interest, including a common European mobility data space (EMDS)[2]. The EMDS will facilitate access, pooling, and sharing of data from existing and future transport and mobility databases, thereby ensuring data accessibility and security among users and data owners.

[2] https://digital-strategy.ec.europa.eu/en/policies/mobility-data

The overall goal of the common European mobility data space will be to accelerate the digital transformation of the European transport sector by allowing availability, accessibility, and re-usage of mobility data. The Data Space will be defined on the interaction of different platforms and digital ecosystems, and it will enhance interoperability, supporting the EU directives and guidelines in Mobility.

The Mobility Data Space[3] of Germany is an example of existing Mobility Data Spaces. It is defined as an open, decentralised ecosystem of data providers, data users, and platforms allowing a sovereign exchange of mobility data and offering a central directory of data resources and services. It is operated by Acatech and funded by the German government. The partners in the mobility sector can share data based on equal rights by supporting innovative, environmentally sustainable, and user-friendly mobility concepts.

All users have access to the Mobility Data Space under the same terms and conditions. Data are handled under specific agreements between providers and users. The Mobility Data Space does not support central data storage, and the data provider can always manage data accessibility.

The International Data Spaces Reference Architecture Model (IDS-RAM)[4] is defined in such a way as to be interoperable and compliant with other data spaces. It also sets the conceptual basis for the exchange of data between organisations and includes mobility data provided by National Access Points[5]. Shared data will provide, for example, information on hazards, roadworks and available truck parking spaces along motorways and highways. All such data, provided both by private or public companies and authorities, must meet European guidelines on open data.

All the data are organized into a data catalogue, and the Data Space collects different kinds of data. Each dataset is described by relevant metadata such as the data description, the provider, the usage policy or the direct link (Access URL). All data are categorized in pre-defined standardized data categories (e.g., real-time information on traffic jams, roadworks and road conditions, public transport information). Such a categorization also helps a user in searching for a specific dataset in the catalogue and offers its users easy access to datasets by following transparent rules. Moreover, in accordance & & & &Y & &J 2& & & & & & & 2& & & & Mobility Data Space has been designed in a decentralized manner, and data is held by its users.

Mobility Data Space complies with the European GDPR as well as the European directives of the Data Governance Act (DGA), Digital Service Act (DSA)[6] and Digital Markets Act (DMA)[7]. Moreover, in order to ensure complete compliance with data protection requirements, the German Federal Commissioner for Data Protection and Freedom of Information (BfDI)[8] has been consulted.

TRANSFORMING TRANSPORT

The TRANSFORMING TRANSPORT (TT) Project [51] is a European Project, funded under the Program H2020 - EU.2.1.1. - INDUSTRIAL LEADERSHIP - Leadership in enabling and industrial

[3] https://mobility-dataspace.eu/
[4] https://internationaldataspaces.org/ids-ram-3-0/
[5] https://transport.ec.europa.eu/transport-themes/intelligent-transport-systems/road/action-plan-and-directive/national-access-points_en
[6] https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:52020PC0825
[7] https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?lang=en&reference=2020/0374(COD)
[8] https://www.bfdi.bund.de/EN/Home/home_node.html

technologies - Information and Communication Technologies (ICT). It was coordinated by the Spanish INDRA SISTEMAS SA, with Grant Agreement ID 731932.

This project represents an interesting example of work reporting experiences on interoperability and direct data-sharing. It was aimed to demonstrate the transformations that a huge amount of data can bring to the mobility and logistics market in a realistic, measurable, and replicable way.

To this end, Transforming-Transport validated the technical and economic viability of big data to reshape transport processes and services to significantly increase operational efficiency, deliver improved customer experience, and foster new business models. Transforming Transport addressed different pilot domains for the mobility and logistics sector in Europe. Examples were represented by smart highways, sustainable vehicle fleets, proactive rail infrastructures, multimodal urban mobility, etc.

The project was built on a Digital Ecosystem that allowed the definition of key aspects such as interoperability, coopetition and no central infrastructure. Interoperability referred to shared guidelines and rules for different actors (in private and public sectors), to the selection of standards and technologies to ensure interoperability via APIs. Coopetition referred to the definition of digital relationships among the actors by considering their strategies. Finally, the data were exchanged without a central infrastructure.

The TT Data Assets Ecosystem led the pilots to describe and publish their data assets, both in terms of functionalities and usage policies, so that the other participants can discover and exploit them, according to their own usage policies, to get a deeper insight into the KPIs of the driving business. The roles were defined according to the RACI model. Examples of emerging tasks for the roles defined above, covered the data committee, data governance leader, ownership and management of data resources, and others. The project addressed the coordination and guidelines of the pilot. At the data asset level, the direct link to the data assets and metadata management was defined. Starting from the experience of the partner Cefriel with the E015 ecosystem, this project took a similar approach customising it in the Transforming Transport domain to obtain a set of best practices for the actors according to the digital ecosystem guidelines.

Data quality management was part of a wide activity of the project, aimed at obtaining a set of guidelines for managing the data asset from the point of view of quality, homogenization, and integration. Moreover, particular attention was given to data management of different kinds of data, including access, storage, backup solutions, and standards (INSPIRE, ISO standards, OGC services, DIA, ...).

The DG guideline defined in this project also included aspects related to information privacy, security, and compliance, based on European data privacy policies.

The project also managed the sharing of the datasets collected by the various pilots through a solution based on the Open Data Portal (ODP) [52]. The Transforming Transport ODP gathers all the datasets used in the context of the pilots. Such datasets came from other open data portals, municipalities, or other public institutions. If data were available through APIs, URLs of such data access were also made available as part of the metadata, with data stored in different data storage systems. Only basic metadata was made available for data that cannot be disclosed.

HARMONY

HARMONY [26] has received & & &K &[&N &8686&X & & Innovation Programme under grant agreement N. 815269. HARMONY is a project under the CIVITAS initiative, and it aims to develop a solution to support authorities in metropolitan areas by guiding them through a sustainable transition to a new era of low-carbon mobility. The project idea is to harmonise spatial and multimodal transport planning tools by modelling change in the transport sector and spatial organisation.

In detail, the HARMONY model suite enables end-users such as planners, decision-makers, researchers, and transport operators/providers to link independent models and analyse a portfolio of regional and urban interventions for passenger and freight mobility. The main objective is to evaluate such interventions from different points of view. Examples concern their impact on land use, economic growth, transportation networks, energy, vehicular noise, and emissions.

HARMONY includes prototype methods and tools that facilitate the integration of data from many heterogeneous sources in an automated and standard way. Data Governance is therefore implemented in the processes of ensuring data quality and ensuring how large data streams are managed efficiently. Much of the data considered belongs to the big-data category: large quantities of data, captured every day in intervals ranging from hours to seconds, varying from real-time to simulated traffic data, floating-car and GPS data, weather and traffic forecasts, and historical data among several others.

The DG is also involved in data collection, acquisition and pre-processing activities to make them homogeneous and interoperable. The approach for developing HARMONY relies on open standards to the largest possible extent. This applies to the data import and access interfaces and communication protocols. For this reason, the project is aimed to ensure data interoperability through well-known standards, such as Geography Markup Language (GML) [40], GeoJSON [24] and Shapefile for geographical information [46], GTFS for public transportation schedules [25] and DATEX II for traffic information [10]

The project handles different users and groups, having access to the functionalities as well as the stored information. Access rights are required to ensure that sensitive information can be viewed only by the appropriate users, with a Role-based access control (RBAC) approach. This approach establishes permissions based on groups (pre-defined sets of users) and roles (defined sets of actions). Individuals can perform any action that is assigned to their role (or multiple roles if necessary). Users cannot change the level of access control assigned to their role.

MOMENTUM

The MOMENTUM Project [37], Modelling Emerging Transport Solutions for Urban Mobility, is a European project under the Horizon 2020 Program, with Grant Agreement N. 815069. Coordinated by the Municipal Transport Company of Madrid, it started in 2019, and finished in April 2022. The project aimed to develop a set of methods for analysing and exploiting mobility data. It was based on the definition of transport models, planning tools, and decision support tools, able to capture the impact of new transport solutions and ICT-driven behavioural changes on urban mobility. The objective was to support local authorities in designing the right policy solution to exploit the full potential of emerging mobility trends.

Concerning the governance aspects and the measures to be taken, the project indicated that the public authorities are responsible for managing and regulating the data processed and the type of transport considered (urban or interurban). This meant a continuous effort to cope with the multiple actors involved in transport. The project has also shown that policies are not static, as innovations in the

transport supply push towards situations that may require new solutions or modify existing ones. Mobility management is a primary issue for both local authorities and their governmental stakeholders. The need for continuous improvement of modelling and decision support tools was highlighted. The project analysed the application of different policies promoted by public administrations for urban mobility management.

Furthermore, during the first half of the project, different data sources were collected, harmonised and analysed, including public transport supply and demand data for the four project case studies. The analysed data came from both traditional and new sources such as smart card data, telecommunication data, etc. The various data sources were harmonised and evaluated in terms of reliability, temporal and geographical scope, level of granularity, availability and accessibility, relevance and usability.

Authorities have defined regulations for data-sharing and mobility systems by introducing, if necessary, licences for service operation and data access. Some regulations have assumed the form of Service Level Agreements (SLAs). For example, an SLA stipulates specific conditions for the publication and sharing of the data (e.g., frequency, quality, data standards, delivery format, etc.).

NOESIS

The NOESIS project [38], the NOvel Decision Support tool for Evaluating Strategic big data investments in Transport and Intelligent Mobility Services, is a European Project under the Horizon 2020 Program, with Grant Agreement N. 769980. The project started in 2017 and finalized in 2019. Its goal was to identify the critical factors/features which lead to the successful implementation of big data technologies and services in the field of transport and logistics with significant value generation from a socio-economic point of view.

The project presented and analysed how, with respect to digitalization, organizations were facing numerous data-related challenges. In fact, data have been increasingly considered an organizational asset, and several private and public organisations in the transport sector investigated the usage of big data services. Within the implementation of big data services, one of the main challenges was how to manage data in the organization by adopting an effective Data Governance framework. The concept of Data Governance refers to how an enterprise is organized to handle its data and extract value from it. Data Governance contributes to increasing the organizations' value and competitiveness within a sector.

The objective of the NOESIS project was to investigate the transferability of big data-based solutions concerning institutional and governmental constraints. Examples of processes considered are:

- elaboration of the evolving landscape of regulations and directives on data privacy, security and openness concerning big data implementations;
- identification of networks of possible actors who are interested in or contribute to data privacy, security and openness issues;

the stakeholder engagement by creating a formal organizational structure and defining their roles, and the setting of policies and standards to be used. All policies related to privacy, security and openness were taken into account, as well as standards regarding how to manage data formats, accessibility, and data collection principles.

DG aspects were analyzed for each NOESIS case study, referring to real cases for public organizations and institutes, and private companies [39]. Concerning public organizations, four open data initiatives & & & & & /& &T & &Y &U &Data Portal[9] has set an effective data governance framework and created an Open Data Website where, unlike other data governance 2& & & & & & & & & & & ; (ii) the Open Health Data Governance Strategy and Policy by the Irish Public Sector (HSE)[10] has defined detailed policies and standards within the open data strategy; (iii) the Kentucky Department of Education (KDE) Data Governance[11] provides an example of how data within an organization could be shared across multiple departments to improve accuracy and timeless access to data; (iv) the Data Governance Framework in the Australian Institute of Health and Welfare (AIHW)[12] proposes interesting solutions for data storage and a web-based application for data access. On the other hand, several private companies from the financial, health, and insurance sectors are adopting data governance frameworks. The UniCredit Bank (UCB) in Germany and the KPMG represent two examples. The first use case may be of interest to those transport companies that manage massive data and intend to improve its quality to be effectively used by all departments. The latter use case is interesting for identifying the main DG aspects and how to implement DG within an organization.

LEMO

LEMO**,** Leveraging big data to Manage Transport Operation [31] is a Horizon 2020 European project started in 2017 and finalized in 2020, with Grant Agreement N. 770038. The project explored the implications of the utilisation of big data to enhance the economic sustainability and competitiveness of the European transport sector. It analysed big data in the European transport domain.

One of the main objectives regarded the definition of a roadmap for research and policy towards open data, data collection, data use, and data-sharing to support European transport stakeholders in defining and addressing Data Governance issues, such as data privacy and security.

A second objective of the project concerned the involvement of European transport stakeholders to identify and analyse concrete opportunities, criticalities and limitations of transport systems, to exploit the opportunities provided by the use of big data. Finally, the activity aimed at disseminating the results obtained, together with a series of recommendations, improving knowledge of the potential of big data in the transport sector.

The project provided a synthesis of public and private policies on big data and identified which ones facilitate and support the access, connection and (re)use of big data, and which ones may create critical issues. A key aspect emerging from the project concerned open access to big data. Open access to a

[9] https://data.ny.gov/
[10] https://data.ehealthireland.ie/about
[11] https://strategy.data.gov/overview/
[12] https://www.aihw.gov.au/

wide variety of information contributed to high levels of data-sharing and reuse, which aligned with decision-making to meet scalability and performance.

Another key issue in big data management concerned privacy and the balance between privacy and access to data. To provide an overview of big data policies at the European level, the project summarised, in a series of tables, policies (including legislation, legislative proposals and governance) active at the European level. The tables provided information on the collection, usability, sharing, and linking of big data that impacted the transport sector.

The project finally considered directives and regulations common to different countries of the European Union and grouped them into four main areas, respectively, intelligent transport systems, open data, automated driving, and smart mobility [12].

SPRINT

The SPRINT project [47] (Semantics for PerfoRmant and scalable INteroperability of multimodal Z /& & & & & &K &[&N &8686&X & & Innovation Programme (IP) under Grant Agreement N. 826172. The project led developments for fostering the adoption of the Shift2Rail Interoperability Framework (IF) by addressing the specific challenges arising from the Shift2Rail Innovation Programme 4 (IP4). The SPRINT project promoted a modal shift towards green modes of transport, such as rail, by improving the intermodal transport. The main objectives were summarized as follows:

- improve Interoperability Framework performance and scalability to sustain a large deployment;
- simplify and automate all the necessary steps needed to integrate new services and sub-systems in the IP4 ecosystem.

In particular, regarding the vision for an interoperability framework, the project focused on the following aspects:

- it supports transport operators to be interoperable (e.g., to implement message exchange or data conversion). The activity is done by providing pre-configured tools that transport operators can modify and adapt to their systems by posing emphasis on «sharing» and «discovering» distributed resources;
- each transport operator describes its system/API and publishes such descriptions in catalogues;
- it provided an "open" framework in which multiple interoperability solutions were coexisting, along with recognition of external, trusted sources of data and metadata.

Z & & & & & & & & & & & & & of data and metadata. Such an ecosystem allowed the definition of data catalogues, which was leveraged as sources of information. These catalogues were maintained by external sources (as well as standardization bodies), subjected to EU regulations, and referred to trusted data/metadata sources. Examples included Open Data Portals, National Access Points, or even other ecosystems defined using the Interoperability Framework.

European regulations (for example, related to National Access Points) defined the stakeholders (transport authorities, transport operators and infrastructure managers) as well as the data formats (e.g., NeTEx, SIRI and DATEX II). In most cases, transport stakeholders used their data formats or adopted simple models (e.g., GTFS), while the required formats are rarely used (in some cases due to their complexity). In this sense, the objective of the project was to make the data compliant with the EU regulation, which required converting data and managing the complexity of standards.

Moreover, the ecosystem implemented in this project provided building blocks and tools to define the interoperability-oriented model. The Data Governance structure was divided between "Strategic governance" and "Technical governance", where the latter was authorized to control the quality of the Interoperability Framework assets and to approve their publication on an Asset Manager. The role of the Asset Manager in the entire ecosystem implemented in the project was to provide a catalogue of pointers to resources that can directly or indirectly facilitate the construction of an interoperability solution. These resources were divided into different "types" and each type was described by a different set of metadata.

In this sense, the two main actors involved corresponded, respectively, to the asset contributor and the Technical Management Board (TMB). The TMB had the rights to perform quality checks and decided whether to approve or reject the release of an asset. Once the decision was made, the asset was "locked" to prevent unauthorized changes.

RUDI

RUDI, (Rennes Urban Data Interface) [45] is a European project part of the Urban Innovative Actions (UIA) program, in the Digital Transition field. Started in 2019, it will finish in 2022. It defines an [illegible] based on the creation of a meta-catalogue of data and offers features directed towards individuals to enhance their knowledge of, and control over their data, and towards the project holders to facilitate the management of data rights and the implementation of innovative economic models.

One of the main challenges of the RUDI project is to put citizens at the centre of the approach of this public data service. More broadly, the team has to define the role of citizens in the implementation of RUDI. The co-construction of the platform requires including the citizens of the metropolis from the design phase and throughout the implementation phase.

RUDI aims to meet the expectations of citizens in terms of transparency, security and services to guarantee trust between the different actors. The partners' main activities consist of recruiting and leading a panel made up of citizens, experts, associations, companies, producers and users of data. In detail, the team organizes the actions needed to create a catalogue of shared territorial data. The definition and implementation of this catalogue are one of RUDI's challenges. The implementation of such a catalogue involves other transversal activities which draw the basic framework of the decentralized architecture of the RUDI portal. This team is in charge of writing reports and documentation to lead partners' reflection on these modules, the development of tools and the conditions of access to data.

A second objective is to propose and define all the elements necessary for effective implementation, in compliance with the regulations, by guaranteeing a quality service for end-to-end data processing in a federation of IT systems and shared governance. The partners involved in the catalogue definition are mainly data producers.

Another objective of the platform regards the constitution of a territorial catalogue with defined rules of governance.

The Rudi partners have set up a working group to draw up rules concerning anonymisation on the future Rudi platform. The aim is to specify the responsibilities of the various players involved and to build trust, particularly with the residents concerned by the personal data. The GDPR and the Data Protection Act

(LIL) do not include a general obligation to anonymise. Anonymisation process is one solution among others to be able to use personal data while respecting the rights and freedoms of individuals. Indeed, it opens up a wide potential for re-use of data and thus allows actors to exploit and share their data without infringing on the privacy of individuals. This process also allows data to be retained beyond their original retention period. With anonymised data, data protection legislation no longer applies and the dissemination or re-use of anonymised data has no impact on the privacy of the data subjects.

The DG information has undergone several iterations and is subject to input and validation by Rudi partners before release. The project handles the data anonymisation by also defining a set of principles about objectives, actors, and processes belonging to the Data Governance aspects. The following Table 2 summarizes some of the most important ones.

ID	OBJECTIVE	PROCESSES	ACTORS
1	Supporting producer to anonymise data, and monitoring data provided.	Rudi monitors technical, legal and scientific developments and organises the sharing of experience between producers. This monitoring takes into account other sources of data that may allow the anonymity of information to be lifted. It publishes recommendations on effective anonymization techniques. Rudi encourages the provision of anonymization tools implemented by producers. Rudi does not currently have a legal personality or resources of its own to implement anonymization. The responsibility for the processing of personal data and consequently their anonymization cannot be transferred to the project at this time.	Rudi actors have to invest time in a state of the art analysis and documentation. Data producers have to participate in the knowledge extraction process. Rudi partners and third parties (more particularly the citizen) are involved in the publication of recommendations. The producers of the Rudi Federation, in their role as data controllers, are responsible for the implementation of anonymisation techniques.
2	Allowing transparency to define public confidence	Rudi's anonymisation recommendations are public. Each dataset is accompanied by a description of the method used, following guidelines on the publication of methods.	Rudi partners publish and keep up to date on the recommendations. Data producers define how the data has been anonymised.
3	Allowing access to public anonymised data	Rudi grants a comprehension of the data protection issues and provides access to up-to-date educational resources on the anonymisation of personal data. Each year, a RUDI project's annual report is produced as an assessment of the effectiveness of the anonymisation actions.	Rudi administrators redact and publish the annual report, relaying content on the subject.

ID	OBJECTIVE	PROCESSES	ACTORS
4	Guarantee right access and trust	Rudi's governance can decide to refuse publication if the protection of personal data is not assured. RUDI's governance does not refrain from informing the relevant authorities if a violation may create a risk to the rights and freedoms of individuals. Finally, if this risk is high, RUDI's governance informs the persons concerned.	Governance authorities control and may refuse/unpublish data in case of violation of rights. Data producers may be refused publication of their data.
5	Preserving the usefulness of the dataset	The sequence of anonymisation techniques to be implemented: identify which information should be retained according to its relevance; remove direct identifiers and rare values that could allow easy re-identification of individuals; distinguish important from secondary or useless information; define the ideal and acceptable fineness for each information stored.	The producers of the Rudi Federation, as data controllers, are responsible for the choice of the relevant anonymisation process but must follow common principles in making this choice.

Table 2: Summary of DG concepts for the RUDI project

2.2.2 Mobility Data Governance: analysis and identification of best practices

The common objective of all projects described in the previous section is the definition of a DG framework for managing data as a strategic enterprise asset.

This sub-section highlights the main contributions of the projects and identifies a list of best practices and guidelines to be considered for the definition of the TANGENT Data-sharing Governance Model.

European projects and digital ecosystems initiatives: a comparative analysis

The projects analyzed in the previous section have been compared w.r.t. the contributions they provide to the following main components of the data governance reference schema proposed in Section 2.1.1:

- **PROCESSES**: different types of activities, such as operational, technical, onboarding, and strategical, carried out to implement the DG model. Relevant processes to be managed through the data governance regard (i) the definition of a data catalogue for data collection and categorisation, adopting a metadata profile to describe each gathered dataset, (ii) the regulation of data-sharing and data exchange between persons with different roles, and (iii) the pre-processing and harmonisation activities carried out on data.
- **OBJECTIVES**: they can be summarised in the three main objectives: (i) meet the domain needs, (ii) protect and manage data, and (iii) reduce costs of data handling. The objectives are also expressed in relation to processes and actors. For example, data quality management is part of a broader activity aimed at obtaining a set of guidelines for data management from the quality point of view. Another example regards providing support to the local authorities and final users to enhance data availability, accessibility, and (re) usability.

- **RULES**: rules among the actors involved in accessing, sharing, and handling mobility data. European guidelines and laws regarding the different aspects, such as data privacy, data quality, data harmonization and integration need to be considered in the definition of the rules.
- **ACTORS:** are identified according to their roles and positions in the mobility and transport domain. Again, the processes implemented, the considered tools, and the rules are directly related. Examples of actors are transport operators/providers, transport authorities, planners, decision-makers, researchers, etc.
- **ROLES:** the actors involved in the DG processes may cover different roles and have different permissions over the data. Each role may refer to a single person or a group responsible for handling and managing specific procedures to reach defined objectives.
- **TOOLS**: software solutions selected to enable the DG processes. Examples are data catalogues, data pre-processing, and integration tools.

Table 3 shows the results of the comparative analysis.

PROJECT	PROCESSES	OBJECTIVES	RULES		ACTORS	ROLES	TOOLS
			Internal	External			
E015	✓	✓	✓	✓	✓	✓	✓
MOBILITY DATA SPACE	✓	✓	✓	✓			
TRASFORMING TRANSPORT	✓	✓			✓	✓	
HARMONY	✓				✓	✓	✓
MOMENTUM	✓	✓	✓	✓	✓	✓	
NOESIS	✓		✓	✓	✓		
LEMO		✓	✓	✓			
SPRINT	✓				✓	✓	✓
RUDI	✓	✓			✓		

Table 3. Comparative analysis of the DG contributions provided by the analyzed projects

European projects and digital ecosystems initiatives: best practices

Table 4 summarizes relevant contributions for mobility data governance that resulted from the analysis of the European projects and digital ecosystems initiatives.

PROJECT	TOPICS OF BEST PRACTICES	DESCRIPTION
E015 digital ecosystem **E015**	PROCESSES ACTORS ROLES	• Definition of guidelines and rules of the different processes managed by the ecosystem. • Definition of publication processes, by allowing the interaction between the application provider and the ecosystem itself. • Definition and implementation of a digital API ecosystem to simplify the data identification process through which people can share APIs or request them. • Settings of all activities carried out by the Technical Management Board: support and facilitation to the actors involved; publication procedures of APIs; control of the suitability and quality of the publication.
Mobility Data Space **MOBILITY DATA SPACE**	PROCESSES ROLES RULES	• Data catalogue definition for dataset collection and sharing. • Metadata definition for each dataset of the catalogue for identification and description. • Rules definition for sharing and accessing data in the Mobility Data Space. • Roles definition into Data Space structure. • Enabling interoperability, re-usage, portability, and connectivity with other infrastructures.
TRANSFORMING TRANSPORT	PROCESSES RULES ROLES	• Definition of an ecosystem by a set of interrelated components: technical expert team, glossaries, data catalogues, procedures and KPIs for data quality management. • Definition and usage of decentralized infrastructures for data exchange. • Rules to guide data gathering, integration and homogenization. • Roles definition (according to RACI models).
HARMONY **HARMONY**	PROCESSES RULES	• Specification of the required data schemas for storing transport data and for managing heterogeneous data in an automated and standard way. • Definition of procedures and rules for data collection, acquisition, data quality, data integration and security. • Settings of pre-processing steps needed for data harmonisation and fusion. • Definition of access rights and roles for DG in mobility and transport services.

PROJECT	TOPICS OF BEST PRACTICES	DESCRIPTION
MOMENTUM	PROCESSES ROLES RULES	• Definition of rules to guide data collection, sharing, data harmonisation and data security. • Definition of actors involved in data collection and data-sharing processes. • Definition of metadata for dataset description.
NOESIS	OBJECTIVES PROCESSES RULES	• Analysis of the transferability of solutions based on Big Data concerning institutional and governmental constraints of Big Data Governance aspects. • Definition of the infrastructure for Big Data services based on DG policy related to big data optimization and privacy management.
LEMO	PROCESSES RULES	• Definition of policies on big data in public and private transport sectors. • Definition of DG procedures for data collection, data usage and data-sharing to support data privacy and security.
SPRINT	PROCESSES TOOLS	• Usage of open data portals, metadata governance tools and National Access Points as data sources. • Usage of European regulations and standard data schemas. • Definition of an Asset Manager, aimed to provide a catalogue of resources linked to different transport and mobility data, described by a set of metadata. • Definition of a Technical Management Board in charge of performing quality and conformance checking.
RUDI	PROCESSES TOOLS RULES	• Definition of a portal as a federated organization that manages direct communications between data producers. • Rules for data publishing and sharing. • Definition of standards for data and metadata. • Definition of anonymisation rules for data re-use.

Table 4: Summary of relevant contributions from the analysed projects

The analysis of the contributions for mobility data governance provided in Table 4 has allowed us to highlight the following set of best practices:

- establishment of a Technical/Management Board for supporting and facilitating the processes of data collection, publication, access, quality and sharing;
- definition of a metadata catalogue for supporting collection, publication, and sharing processes;
- usage of data sources like open data portals and National Access Points;
- definition of rules to enforce the compliance with European data regulations and standard data schemas;
- definition of a set of rules and actors for publishing, accessing, sharing, harmonising and merging data and metadata;

- definition of a set of rules and actors for ensuring interoperability and re-use of data and metadata.

This set of best practices is considered the starting point for defining the TANGENT Data-sharing governance model described in Section 5.2.

3 FAIR Data Principles

In general, the availability and the reuse of data allow researchers to increase their knowledge about several phenomena, enhance analysis and insights, improve existing models, and create new technologies. Nevertheless, computational analysis of datasets is a crucial phase in any scientific process. So, the reuse of data needs to be adequately managed. For this purpose, a consortium of several stakeholders (from academia, industry, funding agencies, and scholarly publishers) designed in 2016 a set of guidelines [35], the so-called FAIR Data Principles.

FAIR stands for:

- *Findable:* discoverable with metadata, identifiable, and locatable by means of a standard identification mechanism;
- *Accessible:* always available and obtainable; even if the data is restricted, the metadata is open. J & & & & & & & &LGOX2& & & & & & & 2& & & & & & 4&Y & not all data has to be made open. The Open Data Institute (ODI)[13] defines Open Data as those that anyone can access, use and share. According to the ODI, open data must be licenced to make clear that anyone can use the data in any way they want, including transforming, combining, and sharing it with others, even for commercial purposes. On the other hand, Open data may not be FAIR. For example, publicly available data may lack sufficient documentation to meet the FAIR principles, such as licensing for clear reuse;
- *Interoperable:* both syntactically parseable and semantically understandable, allowing data exchange and reuse between researchers, institutions, organizations, or countries;
- *Reusable:* sufficiently described and shared with the least restrictive licences, allowing the widest reuse possible and the least cumbersome integration with other data sources.

FAIR Data principles, see Table 5, precede implementation choice and do not necessarily suggest any specific technology, standard, or implementation/solution. In other words, they are a set of principles and not a standard. FAIR principles are related but technically independent and may be implemented incrementally in any combination as data providers evolve to increasing degrees of FAIRness. Data providers are encouraged to increase the number of FAIR principles they comply with gradually. So, FAIRness can be achieved with a wide range of technologies and implementations.

Findable	Accessible
F1. (meta)data are assigned a globally unique and eternally persistent identifier F2. data are described with rich metadata F3. (meta)data are registered or indexed in a searchable resource F4. metadata specify the data identifier	A1. (meta)data are retrievable by their identifier using a standardized communications protocol A1.1 the protocol is open, free, and universally implementable A1.2 the protocol allows for an authentication and authorization procedure, where necessary A2. metadata are accessible, even when the data are no longer available

[13] https://theodi.org/

Interoperable	Reusable
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation I2. (meta)data use vocabularies that follow FAIR principles I3. (meta)data include qualified references to other (meta)data	R1. meta(data) have a plurality of accurate and relevant attributes R1.1. (meta)data are released with a clear and accessible data usage licence R1.2. (meta)data are associated with their provenance R1.3. (meta)data meet domain-relevant community standards

Table 5: A summary of FAIR principles

3.1 Findable principles

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for the automatic discovery of datasets and services. F1, F2, F3 and F4 *Findable* principles are explored here.

F1. *(Meta)Data are assigned a globally unique and persistent identifier*. Globally unique and persistent identifiers remove ambiguity in the meaning of published data by assigning a unique identifier to every element of metadata and every concept/measurement in datasets. It is hard to achieve other aspects of FAIR without globally unique and persistent identifiers [36]. Identifiers consist of an internet link. Many data repositories will automatically generate globally unique and persistent identifiers to deposited datasets. Identifiers can help other people understand exactly what data mean. Identifiers allow computers to interpret data in a meaningful way (i.e., computers that are searching for data or trying to integrate it automatically). Further, identifiers will help researchers to properly cite a specific work when reusing data. According to F1, an identifier:

- must be globally unique. Globally unique identifiers can be obtained from a registry service (e.g., ORCID[14]) that uses algorithms guaranteeing the uniqueness of newly minted identifiers
- must be persistent. It takes time and money to keep web links active, so links tend to become invalid over time. Registry services guarantee the resolvability of that link into the future, at least to some degree.

Example services that supply globally unique and persistent identifiers are: Digital Object Identifier[15], Identifiers.org[16], Persistent URLs[17], and others.

F2. *Data are described with rich metadata (defined by R1 below)*. In creating FAIR digital resources, metadata can (and should) be generous and extensive, including descriptive information about the context, quality and condition, or characteristics of the data. Rich metadata allow a computer to automatically accomplish routine and tedious sorting and prioritizing tasks that currently demand a lot

[14] https://orcid.org/
[15] http://www.doi.org
[16] http://identifiers.org/
[17] http://www.purlz.org

of attention from researchers. The rationale behind this principle is that someone should be able to find
& & & & & & & & 2& & & & & 4G & 2&
compliance with F2 helps people to locate data, and increase re-use and citations.

<u>F3. *Metadata clearly and explicitly include the identifier of the data it describes*</u>. The metadata and the dataset they describe are usually separate files. The association between a metadata file and the
& & & & & & && & & & & & &&
the metadata. As stated in F1, many repositories will generate globally unique and persistent identifiers for deposited datasets that can be used for this purpose. The connection should be annotated in a formal manner, for example using the *foaf:primaryTopic* predicate in the case of RDF metadata. The *FAIRifier* tool[18] guarantees F3.

<u>F4. *(Meta)Data are registered or indexed in a searchable resource*</u>. Identifiers and rich metadata
& & & & & & & & 4&Good data resources may go unused simply because no one knows they exist. If the availability of a digital resource such as a dataset, service or repository is not known, then nobody (and no machine) can discover it. There are many ways in which digital resources can be made discoverable, including indexing. Principles F1-F3 will provide the core elements for fine-grained indexing by some current repositories and future services.

3.2 Accessible principles

Once the users find the required data, they need to know how it can be accessed, possibly including authentication and authorization. A1 and A2 *Accessible* principles are explored here.

<u>A1. *(Meta)Data are retrievable by their identifier using a standardized communications protocol.*</u> Most internet users & & & & && 4Z & && -level interface to a low-level protocol. Principle A1 states that FAIR data retrieval should be mediated without specialized or proprietary tools or communication methods. This principle focuses on how data and metadata can be retrieved from their identifiers. Barriers to access that should be avoided include protocols that have limited implementations, poor documentation, and components involving manual human intervention. However, note that it may not be possible to provide secure access through a fully mechanized protocol, for example, for highly sensitive data. In such cases, it is perfectly FAIR to provide a contact person with whom access to data can be agreed. This contact protocol must be clear and explicit in the metadata.

<u>A1.1 *The protocol is open, free, and universally implementable*</u>. To maximize data reuse, the protocol should be free (no-cost) and open (-sourced) and thus globally implementable to facilitate data retrieval (e.g., HTTP, FTP, SMTP, etc.). Anyone with a computer and an internet connection can access at least the metadata. Hence, this criterion will impact the choice of the repository where research data will be shared.

<u>A1.2 *The protocol allows for an authentication and authorization procedure, where necessary*</u>. Z &G&
&LGOX& & & & & & & 4&X 2&& & & & & & &
which the data are accessible should be provided. Hence, even heavily protected, and private data can be FAIR. Ideally, accessibility is specified in such a way (e.g., HMAC authentication, HTTPS, FTPS, Telephone) that a machine can automatically understand the requirements and then either

---

[18] https://github.com/FAIRDataTeam/OpenRefine-metadata-extension/

automatically execute the requirements or alert the user to the requirements. It often makes sense to request users to create a user account for a repository. This allows each dataset's owner (or contributor) to set user-specific rights. Hence, this criterion will also affect the choice of the repository where research data will be shared.

A2. *Metadata are accessible, even when the data are no longer available*. Datasets tend to degrade or disappear over time because there is a cost to maintaining an online presence for data resources. When this happens, links become invalid and users waste time hunting for data that might no longer be there. Storing the metadata generally is much easier and cheaper. Metadata are valuable in and of themselves, when planning research, especially replication studies. Even if the original data are missing, tracking down people, institutions or publications associated with the original research can be extremely useful.

3.3 Interoperable principles

Data usually needs to be integrated with other data. In addition, data needs to interoperate with applications or workflows for analysis, storage, and processing. I1, I2 and I3 *Interoperable* principles are explored here.

I1. *(Meta)Data use a formal, accessible, shared, and broadly applicable language for knowledge representation*. Data should be readable for machines without the need for specialized or ad hoc algorithms, translators, or mappings. Interoperability typically means that each computer system at least & & & & & & & hange formats. To ensure automatic findability and interoperability of datasets, it is critical to use (i) commonly used controlled vocabularies, ontologies, thesauri (having resolvable globally unique and persistent identifiers, see F1) and (ii) a good data model (a well-defined framework to describe and structure (meta)data).

I2. *(Meta)Data use vocabularies that follow FAIR principles*. The controlled vocabulary used to describe datasets needs to be documented and resolvable using globally unique and persistent identifiers. This documentation needs to be easily findable and accessible by anyone who uses the dataset. The FAIR Data Point (FDP) software[19] ensures I2.

I3. *(Meta)Data include qualified references to other (meta)data*. A qualified reference is a cross-reference that explains its intent. The goal, therefore, is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data, balanced against the time/energy involved in making a good data model. To be more concrete, if a dataset builds on another dataset (namely, if complementary information is stored in a different dataset) it should be specified. In particular, the scientific links between the datasets need to be described. Furthermore, all datasets need to be properly cited (i.e., including their globally unique and persistent identifiers).

3.4 Reusable principles

The goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings. R1 *Reusable* principle is explored here.

[19] https://specs.fairdatapoint.org/

R1. *Meta(Data) are richly described with a plurality of accurate and relevant attributes.* Principle R1 is related to F2, but R1 focuses on the ability of a user (machine or human) to decide if the data is useful in a particular context. To make this decision, the data publisher should provide not just metadata that allows discovery, but also metadata that richly describes the context under which the data was generated. T & & &s used to indicate that the metadata author should be as generous as possible in providing metadata, even including information that may seem irrelevant (e.g., mentioning any particularities or limitations about the data that other users should be aware of, specifying the date of generation/collection of the data, etc.).

R1.1. *(Meta)Data are released with a clear and accessible data usage licence*. R1.1 is about legal interoperability, regarding usage rights attached to research data. This should be described clearly. Ambiguity could severely limit the reuse of your data by organizations that struggle to comply with licensing restrictions. Clarity of licensing status will become more important with automated searches involving more licensing considerations. The conditions under which the data can be used should be clear to machines and humans.

There are different licences that can be linked to data (e.g., CC BY-SA 2.0[20], MIT licence[21]).

R1.2. *(Meta)Data are associated with detailed provenance*. To reuse data, the users should know where the data came from (i.e., clear story of origin/history, see R1). Metadata should include a description of the workflow that led to the shared data. Ideally, this workflow is described in a machine-readable format.

R1.3. *(Meta)Data meet domain-relevant community standards*. It is easier to reuse data sets if they are similar: the same type of data, data organized in a standardized way, well-established and sustainable file formats, documentation (metadata) following a common template and using common vocabulary. If community standards or best practices for data archiving and sharing exist, they should be followed. FAIR data should at least meet those standards. Publishing (meta)data in a manner that increases its use(ability) for the community is the primary objective of FAIRness.

3.5 Data FAIRness and FAIRification

In 2017 Germany, Netherlands and France agreed to establish an international office to support the FAIR initiative, the GO FAIR International Support and Coordination Office[22]. Since its beginning in early 2018, the GO FAIR community has been working towards the implementation of the FAIR Guiding Principles. This collective effort has resulted in a framework that formulates the essential steps towards the end goal, a global Internet of FAIR Data and Services where data are Findable, Accessible, Interoperable and Reusable (FAIR) for machines.

G& & & & & & o FAIR can be found in the *Three-point FAIRification Framework*[23]. By following this framework, stakeholders can rest assured that their efforts toward FAIRification will be optimally coordinated with the efforts of other stakeholders in the GO FAIR community. The framework

[20] https://creativecommons.org/licenses/by-sa/2.0/
[21] https://opensource.org/licenses/MIT
[22] https://www.go-fair.org/
[23] https://www.go-fair.org/how-to-go-fair/

maximizes reuse of existing resources, maximizes interoperability, and accelerates convergence on standards and technologies supporting FAIR data and services.

Typically, the FAIRification process begins when a community of practice considers its domain-relevant metadata requirements and other policy considerations. The FAIRification Framework guides the choice and configuration of FAIR infrastructure, for example the use of FAIR Data Points (FDP)[24] or FAIR Digital Objects (FDO)[25] which contributes to a global Internet of FAIR Data and Services. The framework is an elegant approach that helps a broad spectrum of & & & & &LGOX & & & practice to them, and to insert themselves into the emerging FAIR landscape.

FAIR data can be conceived as a spectrum or continuum ranging from partly to completely FAIR Digital Objects. Different degrees of FAIRness could be conceived that articulate minimal conditions for discovery and reuse to richly documented, functionally linked FAIR data. Each field of research needs to define what it means to be FAIR and decide on appropriate measures to assess this.

The GO FAIR initiative recommends that FAIR data maturity models and metrics should define, across all research areas, a basic minimum standard of FAIR as discovery metadata, persistent identifiers and access to the data or metadata. Several initiatives, for example the *Data Archiving and Networked Services* (DANS)[26], have developed a framework and are piloting self-assessment tools based on their criteria. These approaches make it easy for researchers and data stewards to evaluate the data that they make available and to obtain prompts on how to increase FAIRness. Naturally, such manual self-assessment approaches do not scale but simple, easy-to-understand metrics such as those proposed in these schemes play an important role in engaging and educating the research community to improve practice. Work is underway by various groups to develop metrics and evaluation criteria for FAIR at a data set or digital object level.

The FAIR Metrics group has published a design framework and exemplar metrics [55]. They put forward a template for developing metrics, and the associated GitHub repository provides a core set of quantitative, universally applicable metrics. The intention is that the core set of metrics will be enhanced with additional metrics and qualitative indicators that reflect the needs and practices of different communities. Standardizing the creation of additional metrics in this fashion is recommended.

3.6 FAIR Mobility Data

Open and FAIR data (OFD) represent important opportunities and challenges also in the transport research. The benefits and barriers of applying OFD in the transport field can be identified [29]. Data availability, international data-sharing, and advance research represent some benefits. On the other hand, Big Data poses difficulties when dealing with OFD. Transport research finds a huge challenge due to the nature of transport data (e.g., different sources, types of transports, infrastructures, vehicles, and geographic data). Storing, preserving, compiling, or combining research data represents a crucial issue. Data collected for transport research or by governmental entities tend to be stored in distributed data silos, with different ownerships and data formats, which can cause difficulties when cataloging, finding, accessing, and using research data. Another barrier to the OFD is due the sensitiveness and privacy of the data (see GDPR legal policies). Personal and safety threats can surge from incorrect

[24] https://specs.fairdatapoint.org/
[25] https://fairdigitalobjectframework.org/
[26] https://blog.ukdataservice.ac.uk/fair-data-assessment-tool/

transport data management (e.g., sharing data that it is not meant to be shared). To ensure that members of the transport research community can get benefits from OFD, standard metadata that define data formats are necessary. The use of standards is indispensable for the success of OFD sources for any research data, including the transport sector. Metadata allow automated search engines to catalogue data from different sources.

According to the FAIR Data principles, some considerations on data transportation are mentioned below.

- *Findability*. Transport data should be published so that an entire ecosystem can easily discover them. One directory of data covering all modes of mobility across an entire transport ecosystem is ideally needed.
- *Accessibility.* Transport data should be provided via technologies aimed at allowing persons or systems to access data easily or without unnecessary restrictions (e.g., the data cannot be accessed in a standardised way, the technology is restricted by propriety protocols and/or licences, unnecessary authorisation is required, etc.).
- *Interoperability*. Transport data should be made exchangeable and readable by humans & machines (without needing specialised or clever converters or algorithms) using common entity definitions, standards, or vocabularies across an entire sector or ecosystem.
- *Reusability*. Transport data sources should define how it can or cannot be replicated and/or combined in different ways, along with a clear data usage / re-usage licence. In fact, often it is legally necessary to restrict data usage or prevent freely published data from being subsequently changed or monetised without permission.

4 TANGENT Metadata Profile

To support the adoption of FAIR principles in the TANGENT Data-Sharing Governance Model, WP2 will deliver two tools: the TANGENT Metadata Profile and the TANGENT Data Catalogue. This chapter discusses the definition of a TANGENT Metadata Profile for describing the data sources shared in TANGENT by each case study for the demonstration of services and functionalities. The TANGENT Metadata Profile will support the implementation of the TANGENT Data Catalogue and will be adopted for the description of data shared through the catalogue. The TANGENT catalogue will be an instance of a metadata catalogue, i.e., a digital platform supporting the sharing, findability, and accessibility of distributed digital assets. Both tools will support the definition of the TANGENT solution for data harmonisation and fusion developed by WP2.

The definition of an initial set of metadata fields has been already discussed in the TANGENT &J847&J & & & & & & & & & & & 4& In this deliverable, starting from that preliminary analysis, we consider the best practices for the definition of a metadata profile for data catalogues and the already available initiatives in the transportation domain. Considering these aspects, a conceptual model for the TANGENT Metadata Profile is defined considering a set of requirements for the project. Finally, the steps for its implementation and integration in the TANGENT solution for data harmonisation and fusion are discussed.

4.1 Data catalogue metadata profiles

Metadata are key elements in implementing a data-search functionality for the end user. The main obstacle for cross-portal and cross-border data search is the need for harmonising the set of metadata adopted by each data catalogue. The challenges are:

a) the adoption of different terminology for the same information (e.g., supplier *vs* publisher);
b) the usage of different sets of metadata (e.g., specific information not available in other data catalogues);
c) the usage of different values for metadata fields (e.g., a custom set of string values used as data categories);
d) the implementation of non machine-readable metadata.

In the definition of metadata specifications, Semantic Web technologies offer a valid solution to encode semantics in an interoperable machine-readable format addressing the aforementioned challenges:

- **Challenge (a):** Reference vocabularies (i.e., ontologies) defined in RDF[27] and published online provide a means to ensure a common terminology and semantics of metadata fields;
- **Challenge (b):** Profiles can be defined to improve interoperability by describing constraints and expected combinations on the use of one or more vocabularies as metadata fields;
- **Challenge (c):** Controlled vocabularies (e.g., taxonomies) published online provide a means to ensure the same entity is used to represent the same value for a metadata field;
- **Challenge (d):** Metadata represented in RDF are machine-readable.

[27] http://www.w3.org/TR/rdf11-concepts/

However, an effective application of the mentioned approach relies on the identification of a common set of vocabularies and profiles for the interoperability of metadata shared by different data catalogues.

The Data Catalog Vocabulary (DCAT) [8] is a well-established RDF vocabulary to describe data catalogues and their resources. DCAT Application Profile (DCAT-AP) [11] is the recommended profile for the implementation of DCAT in data catalogues in Europe. DCAT-AP defines stricter guidelines on the usage of DCAT to foster interoperability of metadata.

The intended DCAT-AP scope is cross-border and cross-domain, for this reason, several extensions of the DCAT-AP profile are defined to meet specific requirements. The main types of extensions are custom metadata profiles, extensions for a specific domain, and national extensions adopted as regulation by a certain state.

The definition of metadata specifications as DCAT-AP extensions allows, on one hand, to fulfil specific requirements for the use case considered and, on the other hand, to ensure that a DCAT-AP-compliant export of (a portion of) the metadata is possible. Figure 4 shows how the adoption of DCAT-AP and its extensions supports the interoperability and discoverability of data portals. Data portals exposing metadata in compliance with DCAT-AP enable the possibility for applications to easily consume metadata from multiple sources. The integrated metadata can also be used to feed an overarching data portal, e.g., a European data portal for the findability and discoverability of data.

For these reasons, we decided to define the TANGENT Metadata Profile as an extension of DCAT-AP, namely, TANGENT DCAT-AP. This section introduces DCAT, DCAT-AP and the already available extensions for the application profile. Section 4.2 discusses napDCAT-AP, a relevant initiative defining an extension of DCAT-AP for the transportation domain. Finally, Section 4.3 discusses the approach towards TangentDCAT-AP.

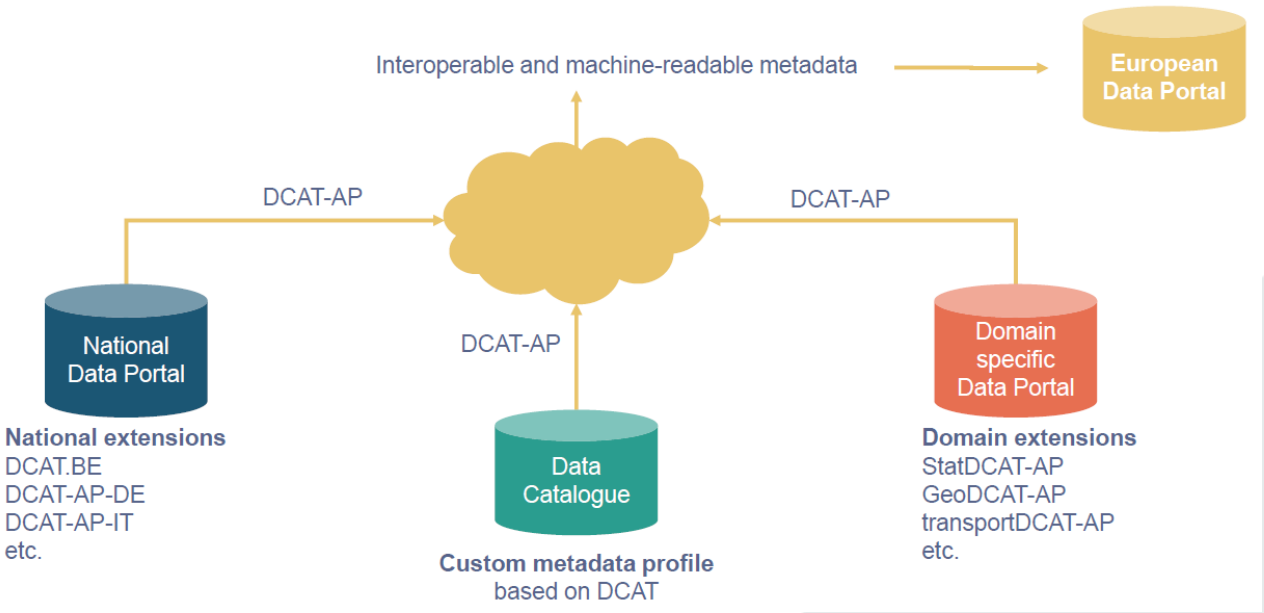

Figure 4: Interoperability of data portals based on DCAT-AP and its extensions

It should be mentioned that other approaches not relying on RDF for the specification of metadata profiles have been defined. As relevant examples, we cite the *Coordinated Metadata Catalogue*

(CMC)[28] initiative for National Access Point (NAP) metadata, not covering the exchange format and the communication protocol that shall be used for (meta)data exchange, and the INSPIRE Implementing Rules for INSPIRE metadata[29] adopting the ISO/TS 19139 based XML format. As discussed in the following sections, both initiatives acknowledge the relevance of DCAT-AP and are currently working on the harmonisation/mapping of their metadata specification to DCAT-AP extensions, e.g., napDCAT-AP for CMC (Section 4.2) and GeoDCAT-AP for INSPIRE metadata (Section 4.1.3).

4.1.1 DCAT

The Data Catalog Vocabulary (DCAT) is an RDF vocabulary designed to describe data catalogues using a standardized set of classes and properties to model data sources.

Figure 5 represents the main concepts defined in DCAT. Each box represents a class that can be assigned to describe an individual entity. An entity of a specific class can be described using a predefined set of properties. Entities can be associated with other entities according to the arrows depicted in green. Yellow arrows represent a *subclass of* relationships.

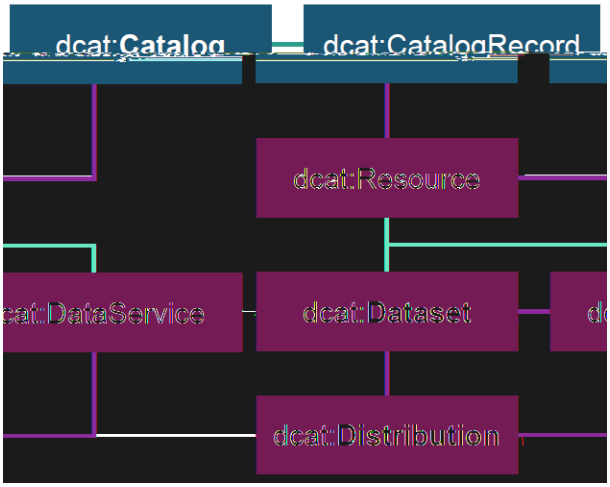

Figure 5: Main concepts and relationships in the DCAT Vocabulary

A *dcat:Catalog* represents a data portal, or a sub-portion of it, if multiple catalogues of data sources, logically distinct, are contained within the same data portal. Metadata associated with a *dcat:Catalog* represents information about the overall catalogue.

A *dcat:Catalog* is associated with a set of *dcat:Resource*s stored in the catalogue, and a set of *dcat:CatalogRecord*s representing the description of a data resource within the catalogue. Metadata associated with a *dcat:Resource* describes the resource itself, whereas metadata for the *dcat:CatalogRecord* are more related to metadata about the metadata (e.g., timestamp of the last update of the resource description in the catalogue).

A *dcat:Resource* can be of two types, a *dcat:Dataset* if it represents an actual dataset or a *dcat:DataService* if it is a web service exposing data. Both types of data sources can be associated with

[28] https://www.its-platform.eu/achievement/monitoring-harmonisation-of-naps/
[29] https://github.com/INSPIRE-MIF/technical-guidelines/tree/main/metadata

one or more *dcat:Distribution* representing a serialization of the data in a specific format. All the *dcat:Distribution* associated with the same *dcat:Dataset* or *dcat:DataService* should contain the same exact information.

4.1.2 DCAT-AP

The DCAT Application Profile (DCAT-AP) specifies, considering DCAT as the base vocabulary, a profile for data portals in Europe to favour the aggregation, exchange, search and automated processing of metadata. DCAT-AP relies on the same concepts as DCAT but describes a profile for its usage through:

- cardinalities and obligations (mandatory, recommended, and optional) for DCAT elements to be provided;
- recommendations for controlled vocabularies[30] to be adopted for metadata values;
- additional properties (not defined in DCAT) that can be used for specific metadata fields.

4.1.3 DCAT-AP extensions

Different extensions of the basic DCAT-AP profile have been defined to support specific requirements, mainly for a nation or a specific domain. A list reporting some of the available extensions is provided in Figure 4.

A public document identifies the guidelines that must be followed for the extension of DCAT-AP[31]. DCAT-AP extensions should ensure compatibility with DCAT-AP, i.e., it should be possible to extract DCAT-AP-compliant metadata from a set of metadata defined according to the DCAT-AP extension. A DCAT-AP extension can: (i) provide additional guidelines and recommendations on the usage of DCAT-AP class and properties, (ii) extend the scope of the metadata profile.

Three main types of modifications are usually defined through an extension:

- changes in cardinalities and obligations (mandatory, recommended, and optional) Example: The property *dct:publisher* (recommended in DCAT-AP) is made mandatory for each *dcat:Dataset* described;
- controlled vocabularies to be adopted for metadata values. Example: The property *dcat:theme* should have as value an element of the controlled vocabulary defined for the extension (e.g., a taxonomy of data categories for transportation datasets);
- additional properties that can be used for metadata. Example: The property *ex:transportMode* is introduced as mandatory for each *dcat:Dataset* described.

In this section, we briefly present GeoDCAT-AP and statDCAT-AP, as widely accepted specifications in the corresponding domain, and transportDCAT-AP, as a relevant extension defined for the transportation domain.

GeoDCAT-AP [23] is the recommended DCAT-AP extension for metadata describing geospatial datasets and data services. The objective of GeoDCAT-AP is not to replace existing recommendations for datasets in this domain, but to specify canonical mappings enabling the representation of metadata

[30] https://op.europa.eu/en/web/eu-vocabularies/controlled-vocabularies
[31] https://joinup.ec.europa.eu/release/dcat-ap-how-extend-dcat-ap

in compliance with DCAT-AP. The binding with DCAT-AP entities enables owners of geospatial information to increase the interoperability of metadata through the RDF format. GeoDCAT-AP describes mappings to DCAT-AP for the union of metadata defined by ISO 19115 (geographic information metadata) and the INSPIRE Directive (2007/2/CE). The GeoDCAT-AP extension includes additional properties with respect to the DCAT-AP standard model and a set of transformation guidelines.

StatDCAT-AP [48] is a DCAT-AP extension supporting the definition of metadata for statistical datasets and enhancing the interoperability of statistical data sources in open data catalogues adopting DCAT-AP. StatDCAT-AP supports the description of statistical datasets in multiple formats. StatDCAT-AP provides a specification that is fully compliant with DCAT-AP and defines a small number of additions to model relevant metadata for the statistical datasets. In particular, StatDCAT-AP allows describing the multidimensional structure of statistical datasets representing the numerical variables (or measurements) on different dimensions, e.g., geographic, temporal or specific dimensions.

TransportDCAT-AP [7] is a profile of the DCAT-AP ontology focused on the public transport domain. The profile takes into account the geospatial nature of public transport data by defining the set of metadata related to this type of information as mandatory. In addition, a specific set of admissible keywords is defined to standardise the range of metadata properties (i.e., the values that can be associated with a property) and to enable domain-related queries, e.g., considering specific transportation modes. The extension is currently not actively maintained and does not have broad adoption in existing transportation data catalogues.

4.2 napDCAT-AP: a metadata specification for data-sharing in the transportation domain

The ITS Directive 2010/40/EU and its Delegated Regulations [6] required each EU Member State to set up a National Access Point (NAP) to publish mobility data. Currently, more than 30 NAPs are in operation in Europe but their implementation shows a high fragmentation with respect to the adopted formats, standards, access interfaces, etc. The NAPCORE project[32] aims at coordinating all NAPs in Europe on an organizational and technical level. The objectives of NAPCORE are:

- to establish common recommendations on data exchange technology, standards, formats, processes used by the NAPs;
- increase interoperability for common data discoverability and accessibility, thus enabling pan-European services;
- define common strategies concerning existing and upcoming developments.

One of the problems addressed by the NAPCORE project is related to the appearance of different metadata vocabularies in European NAPs and the need for interoperability of metadata from different NAPs [1]. The adoption of heterogeneous metadata descriptions is reflected in the availability of different web services and interfaces offering limited search capabilities for the users and ITS service providers. For this reason, the NAPCORE project is currently working on the design, implementation

[32] https://www.napcore.eu

and publication of a napDCAT-AP metadata specification to improve the findability and accessibility of transportation datasets from different data portals [32].

The Coordinated Metadata Catalogue (CMC), defined under the former EU EIP project[33], represented the first elaboration towards NAP metadata harmonisation in the transportation domain. The collection of requirements from stakeholders resulted in the definition of a proper conceptual model for the identification of metadata needed for transportation data. However, the lack of guidelines in the definition of an exchange format for CMC metadata led to several issues in the interoperability of metadata adopting the specification. Therefore, the adoption of an RDF-first metadata specification for transportation data was deemed as a potential solution to increase the interoperability of metadata in NAPs. The CMC specification provides mappings for DCAT-AP, but not all the CMC metadata can be directly mapped to DCAT-AP. For these reasons, starting from the CMC specification, the NAPCORE project is currently working towards napDCAT-AP, an extension of DCAT-AP for the description of data sources in NAPs.

The work towards napDCAT-AP, currently ongoing in the NAPCORE Sub-Working Group (SubWG) 4.4, represents a relevant initiative for the definition of the TANGENT Metadata Profile. Indeed, the best practices in the definition of RDF-based metadata specifications recommend the reuse of already available vocabularies to guarantee interoperability. For this reason, a collaboration between the NAPCORE and TANGENT projects has been established to facilitate the alignment of activities. This section describes the work done and the next steps performed in NAPCORE for the definition of napDCAT-AP. Section 4.3 discusses the definition of the TANGENT Metadata Profile considering both napDCAT-AP and the specific requirements of the TANGENT project.

The first two outcomes of the NAPCORE SubWG 4.4 are described in this section: the definition of a roadmap with guidelines and best practices to design, implement and publish a napDCAT-AP [34]; the elicitation of requirements towards the definition of a DCAT-AP extension for the transportation domain [43]. Finally, the next steps towards napDCAT-AP are discussed.

4.2.1 A roadmap towards napDCAT-AP

To support the definition of a roadmap towards napDCAT-AP the following inputs were considered by the NAPCORE project: documents from the literature describing DCAT-AP and its extensions, artifacts published online for DCAT-AP and its extensions, interviews with DCAT-AP and SEMIC experts[34].

The defined roadmap towards the publication of napDCAT-AP is represented by a diagram describing the roadmap and complemented by a set of guidelines and best practices. As a contribution to this deliverable, a generalised version of the roadmap and the identified guidelines and best practices has & & & &ZGTMKTZ& & & & &G&X & & &J I GZ-AP extension" [33]. This section provides an overview of the general roadmap, Section 4.3 discusses how we plan to apply it within the TANGENT project for the definition of TangentDCAT-AP.

The roadmap is composed of the following five steps:

1. Definition of requirements for the DCAT-AP extension
2. Definition of the domain model for the DCAT-AP extension

[33] https://www.its-platform.eu/
[34] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic

3. RDF implementation of the DCAT-AP extension
4. Documentation and guidelines for the DCAT-AP extension
5. Hosting and publication of the DCAT-AP extension

The first phase of the roadmap is related to the definition of requirements for the DCAT-AP extension. A collection of use cases and functionalities, which should be supported by the extension, must be identified involving the relevant stakeholders and used to elicit a set of requirements. This phase should consider also existing projects and initiatives for metadata to support their reuse in the design and implementation phase. In particular, the already existing DCAT-AP extensions should be considered.

The second step of the roadmap concerns the definition of the conceptual model considering the elicited requirements. Relevant steps to be performed are: (i) analyse the domain and the use cases addressed by the model to be developed, (ii) reach a consensus by developing semantic agreements among different stakeholders, (iii) define and implement a structured and clear model, and (iv) publish the domain model for public revision.

The third step of the roadmap is related to the implementation of the defined conceptual model in RDF. This step should identify and/or define the set of RDF classes and properties to encode the domain model. Moreover, controlled vocabularies may be specified or designed to define constraints on the set of admissible values for specific properties. Finally, this phase identifies cardinality and produces different RDF serialization of the modelled vocabularies.

The fourth step of the roadmap concerns the definition of the documentation and additional artifacts supporting the publication of the DCAT-AP extension. The additional artifacts may specify:

- guidelines for the implementation and usage of the specification;
- canonical mappings with respect to other relevant metadata standards;
- guidelines/tools for the validation of metadata adopting the DCAT-AP extension;
- governance and maintenance structures;
- methodologies for the revision and update of the DCAT-AP extension;
- licence for the DCAT-AP extension and the controlled vocabularies implemented.

The fifth step of the roadmap defines the steps needed for the hosting and publication of the DCAT-AP extension and the controlled vocabularies defined for the extension. A hosting solution should be identified and configured properly to serve the artifacts produced according to the namespace assigned to the extension.

4.2.2 Requirements for napDCAT-AP

The definition of requirements for napDCAT-AP, i.e., the first step of the roadmap, considered a wide analysis of the literature and the direct involvement of relevant stakeholders (NAPCORE partners, NAP developers, NAP metadata provider, NAP metadata user, Metadata standard experts). The references analysed include previous research, policy, documents, and project documentation. The analysed resources can be grouped into 6 categories:

- open data literature;
- European legislature (Public Sector Information Directive, Open Data Directive, INSPIRE Directives, Data Governance Act, EU strategy);
- metadata standards and interoperability;
- DCAT-AP materials (documentation, training webinars, and extensions);
- EU EIP metadata, documents (Coordinated Metadata Catalogue specification, deliverables, initial napDCAT-AP proposal);

- project results (in particular, the SPRINT and LOD projects).

The resources providing relevant inputs for the definition and implementation of napDCAT-AP focused on: FAIR principles for data-sharing, guidelines for the definition and usage of metadata, (transportation) data catalogues and metadata currently adopted, data sources to be described for the transportation domain, guidelines on how to build metadata specifications extending standardized vocabularies, challenges and requirements in the definition of metadata specifications (also considering other domains), tools and existing vocabularies (in particular DCAT-AP and its extensions), and mappings between (meta)data concepts in different specifications.

Considering both the inputs from the literature and the involved stakeholders, NAPCORE identified forty requirements for napDCAT-AP, divided into the following categories: *General* (8), *Existing Vocabularies* (5), *Content* (22), *Implementation* (5). *General* requirements describe high-level requirements for the metadata specification, *Existing Vocabularies* requirements identify already available specifications and initiatives to be considered, *Implementation* requirements address the application of best practices in the development of napDCAT-AP.

Due to the shared context concerning transportation data, *Content* requirements are the most relevant for the definition of the TANGENT Metadata Profile and should be considered for the definition of requirements for TangentDCAT-AP. Table 6 summarizes the collected *Content* requirements with

4.2.3 Implementation and publication of napDCAT-AP

The ongoing steps within NAPCORE SubWG4.4 are related to the development of napDCAT-AP and the definition of structures and processes for the maintenance and governance of napDCAT-AP.

The definition of the first version of the napDCAT-AP conceptual model (second step of the roadmap) has been completed by the working group and a first review of the model by relevant external stakeholders is expected in the following months. In parallel, the work for the implementation (third step) and documentation (fourth step) is starting and will be finalized once the comments received from reviewers will be incorporated into the conceptual model. Finally, the first version of napDCAT-AP will be published (fifth step) in the second trimester of 2023.

4.3 TangentDCAT-AP

The TANGENT Metadata Profile will support the description of the data sources shared by stakeholders for each TANGENT case study and collected within WP2 in the TANGENT Data Catalogue. The TANGENT Metadata Profile is a relevant tool in the definition of the TANGENT Data-Sharing Governance Model, as will be discussed in Section 5, and should facilitate the findability and accessibility of data sources by TANGENT users.

A preliminary version of the TANGENT Metadata Profile was proposed in the TANGENT deliverable D2.1 to support the data collection phase. The definition of this metadata profile considered different metadata specifications to identify a basic set of metadata fields. The fields were selected to support the description of data sources shared in TANGENT by different stakeholders in each case study to match the data requirements. However, the metadata profile defined in D2.1 did not support a machine-readable encoding of the information and did not provide guidelines and cardinalities restrictions for its usage. To enable an effective harmonisation of metadata supporting the implementation of the TANGENT solution for data harmonisation and fusion, this deliverable defines the TANGENT Metadata Profile as an extension of the DCAT Application Profile, following the European recommendations, and considering the specific requirements of the TANGENT project.

Section 4.2 described the NAPCORE initiative for the definition of napDCAT-AP, a metadata specification for the harmonisation of metadata among European NAPs. The initiative is very relevant for the definition of the TANGENT Metadata Profile because it defines an extension of DCAT-AP in the same transportation domain. On one hand, we leverage the work done by NAPCORE to define a roadmap for TangentDCAT-AP, relying on guidelines and best practices from state-of-the-art and metadata experts. On the other hand, we will consider the reuse of the artifacts produced for napDCAT-AP in the different steps of the roadmap.

This section discusses the different steps of the roadmap and how we plan to apply them in the design, implementation, and publication of TangentDCAT-AP.

4.3.1 Requirements for TangentDCAT-AP

The first step of the roadmap is related to the definition of requirements for TangentDCAT-AP. As input for this activity we considered the following inputs:

- the initial set of metadata fields defined in TANGENT D2.1 for data collection;
- issues and feedback from TANGENT partners in the provision of metadata according to the metadata profile adopted in TANGENT D2.1 for data collection;

- list of requirements defined for napDCAT-AP [43], in particular considering the *Content* requirements reported in Table 6;
- set of metadata fields available for data sources collected for TANGENT but already published on other data portals;
- additional inputs from TANGENT partners considering the expected usage of metadata for the findability of data sources (data search process).

Table 7 reports an initial list of requirements for TangentDCAT-AP derived from the metadata profile for data collection defined in the TANGENT deliverable D2.1.

Requirement ID	Description
TangentDCAT-AP-1	The data source should be associated with a unique identifier.
TangentDCAT-AP-2	The data source should be associated with a description of the data according to a specific list of pre-defined categories.
TangentDCAT-AP-3	The data source should have an explicative name.
TangentDCAT-AP-4	The data source should have a description of its content.
TangentDCAT-AP-5	The data source should be associated with a description of the provider(s) of the data source (e.g., transport authorities, operator, etc) and/or the platform (e.g., NAP, open data platform) that stores/captures the data.
TangentDCAT-AP-6	The temporal aggregation of the data (e.g., minutely, hourly, weekly, monthly, or event-based) for a data source should be specified.
TangentDCAT-AP-7	The timeframe of the data (e.g., last year, 31st December 2020, 2017-2021, specific period, etc...) for a data source should be specified.
TangentDCAT-AP-8	The geographical aggregation of the data (e.g., kilometres, metres, etc.) for a data source should be specified.
TangentDCAT-AP-9	The geographical coverage of the data (e.g., group of streets, specific area, city, metropolitan area, region, national) for a data source should be specified.
TangentDCAT-AP-10	The terms for the access and usage of the data (e.g., open data, available only under agreement, released with specific licences) should be specified for a data source.
TangentDCAT-AP-11	How to access the data (e.g., direct download URL, API) should be specified for a data source.
TangentDCAT-AP-12	The data schema/model adopted to represent the data (e.g., GTFS, NeTEx, Datex, SIRI, etc) should be specified for a data source.
TangentDCAT-AP-13	The format used to encode the data (e.g., XML, CSV, JSON) should be specified for a data source.

Table 7: Initial requirements from the metadata profile for data collection defined in D2.1

In the definition of the list of requirements presented in Table 7, we considered the feedback from TANGENT partners in the provision of metadata for the data collection phase described in D2.1, and

the metadata available for data sources already published in external data portals. The main modifications introduced with respect to the TANGENT Metadata Profile presented in D2.1 are:

- we removed the constraint on the *category-subcategory-type* categorization adopted for the data collection. This categorization, based on the results of the H2020 project Momentum, is not aligned with available controlled vocabularies. In particular, the definition of the *type* field is customizable by the user causing heterogeneity issues in the categorization. TangentDCAT-AP will consider available controlled vocabularies to categorize transportation data (in alignment with napDCAT-AP), guaranteeing a machine-readable and interoperable representation of the metadata values;
- the description of the *type of the data* (e.g., static, historical, real-time) was interpreted in different ways by TANGENT partners providing metadata, thus causing heterogeneity issues. We decided to remove this field considering this information as inferable from the temporal granularity and scope;
- the definition of *geographical granularity* in terms of geospatial entities (e.g., streets, stops, neighbourhoods, lanes, city, region) also resulted in very heterogeneous interpretations of the semantic of the metadata field by TANGENT partners. We decided to define this metadata field as done in other data portals, i.e., based on the unit of length;
- the *availability* field is used with a different meaning in DCAT-AP (more related to the planned availability of the data source[35]). For this reason, the associated requirement is now defined specifying the need for providing information regarding both the terms of access and usage of the data source;
- the fields *accessibility* and *data source URL* were merged as a single requirement, indeed in DCAT the description of the data source as dataset or data service requires different metadata fields.

The presented list of data requirements is complemented by additional requirements listed in Table 8 and considers the specific needs of the TANGENT project.

Requirement ID	Requirement Description
TangentDCAT-AP-14	The napDCAT-AP specification should be considered in the definition of TangentDCAT-AP as the reference DCAT-AP extension for the transportation domain.
TangentDCAT-AP-15	The set of guidelines defined in TANGENT (cf. D2.1) for the definition of the data source identifier should be followed.
TangentDCAT-AP-16	The data sources should be categorised considering the case study they are related to.
TangentDCAT-AP-17	The data sources should be described according to the set of data requirements fulfilled.
TangentDCAT-AP-18	The data sources derived from another data source (e.g, harmonisation and/or fusion) should explicitly refer to the original data source.

Table 8: Additional requirements for TangentDCAT-AP

[35] Cf. the controlled vocabulary http://publications.europa.eu/resource/authority/planned-availability

4.3.2 Design and implementation of TangentDCAT-AP

To define the first list of metadata fields for TangentDCAT-AP, we discuss the coverage of requirements with respect to DCAT-AP v2.1.1, which represents the minimum set of metadata for a DCAT-AP extension. For each requirement covered, we mention both the class of the entity associated with the metadata field (e.g., Dataset), the property defined to encode the metadata and the cardinality obligation defined by DCAT-AP (mandatory, recommended, optional).

The following prefixes are used in Table 9 to identify RDF vocabularies:

- adms: http://www.w3.org/ns/adms#
- dcat: http://www.w3.org/ns/dcat#
- dcatap: http://data.europa.eu/r5r/
- dct: http://purl.org/dc/terms/

Requirement ID	DCAT-AP Coverage	
TangentDCAT-AP-1	dcat:Dataset	*adms:identifier* (optional)
TangentDCAT-AP-2	dcat:Dataset	*dcat:theme* (recommended) *dct:type* (optional)
TangentDCAT-AP-3	dcat:Dataset	*dct:title* (mandatory)
TangentDCAT-AP-4	dcat:Dataset	*dct:description* (mandatory)
TangentDCAT-AP-5	dcat:Dataset	*dct:publisher* (recommended) *dct:creator* (optional) *dct:provenance* (optional)
TangentDCAT-AP-6	dcat:Dataset	*dcat:temporalResolution* (optional)
TangentDCAT-AP-7	dcat:Dataset	*dcat:temporal* (recommended)
TangentDCAT-AP-8	dcat:Dataset	*dcat:spatialResolutionInMeters* (optional)
TangentDCAT-AP-9	dcat:Dataset	*dcat:spatial* (recommended)
TangentDCAT-AP-10	dcat:Dataset dcat:Distribution	*dct:accessRights* (optional) *dct:license* (recommended)
TangentDCAT-AP-11	dcat:Dataset dcat:DataService	*dcat:accessURL* (mandatory) *dcat:downloadURL* (optional) *dcat:endpointURL* (mandatory) *dcat:endpointDescription* (recommended)
TangentDCAT-AP-12	dcat:Distribution	*dct:conformsTo* (optional)
TangentDCAT-AP-13	dcat:Distribution	*dct:format* (recommended)
TangentDCAT-AP-16	Not mapped in DCAT-AP	
TangentDCAT-AP-17	Not mapped in DCAT-AP	
TangentDCAT-AP-18	dcat:Dataset	*dct:source* (optional)

Table 9: Coverage of TangentDCAT-AP requirements in DCAT-AP

The documentation of TangentDCAT-AP will reference the one produced for napDCAT-AP while adding specific content on:

- additional properties defined by TangentDCAT-AP;
- controlled vocabularies defined by TangentDCAT-AP;
- changes in the cardinality obligations (mandatory, recommended, optional);
- guidelines for the usage of TangentDCAT-AP (e.g., guidelines for identifiers as requested by the requirement TangentDCAT-AP-15).

5 TANGENT Data-Sharing Governance

The objective of data-sharing in TANGENT consists in enabling the TANGENT services and functionalities to be demonstrated in each case study. Data-sharing must be performed in accordance with a structured and effective Data Governance model.

This section provides a short report on the activities performed to collect inputs from stakeholders during the multi-acto & & & & & V7& &J 749&S -actor cooperation models for TZS4&L &X /4&Z & formation collected from stakeholders and the results of the analyses described in the previous sections have been used to define the TANGENT Data-Sharing Governance Model (Section 5.2).

5.1 Data-Sharing Governance Workshops

In the context of the WP1 of the TANGENT project, multi-actor cooperation workshops with the stakeholders involved in the TANGENT case studies have been organized. Each workshop had a dedicated section on data governance and data-sharing aiming at collecting information about directives, regulations, and DG aspects considered and applied by the TANGENT case study cities.

In the first part of each DG section, after the description of the adopted Data Governance Reference Schema (see Section 2.1.1), the following initial set of data-sharing governance processes were presented:

- *Dataset identification and description*: identification of high-quality datasets to fulfill TANGENT data requirements. Description of the datasets in the TANGENT Data Catalogue. Sharing of the catalogue within the TANGENT consortium.
- *Data access*: get access to the content of the dataset without manipulation.
- *Data harmonisation*: data conversion and fusion to support TANGENT service functionalities.
- *Data usage*: data usage to feed the TANGENT services.

Then, in the second part of each DG section, the participants, both case study leaders and stakeholders, were invited to actively join a working session by filling a Miro Board focused on the main aspects of data governance. The Miro Board was defined by four frames, one for each considered process. Figure 6 shows one of the frames. Each frame's header described the process, while the body was divided into three columns. The first column aimed to collect feedback on examples of rules proposed by TANGENT to regulate data-sharing within the project partners. The second column was proposed to collect information on active regulations and directives at the city or national level that TANGENT must be aware of. The last column was proposed to collect information about stakeholders potentially involved in the process.

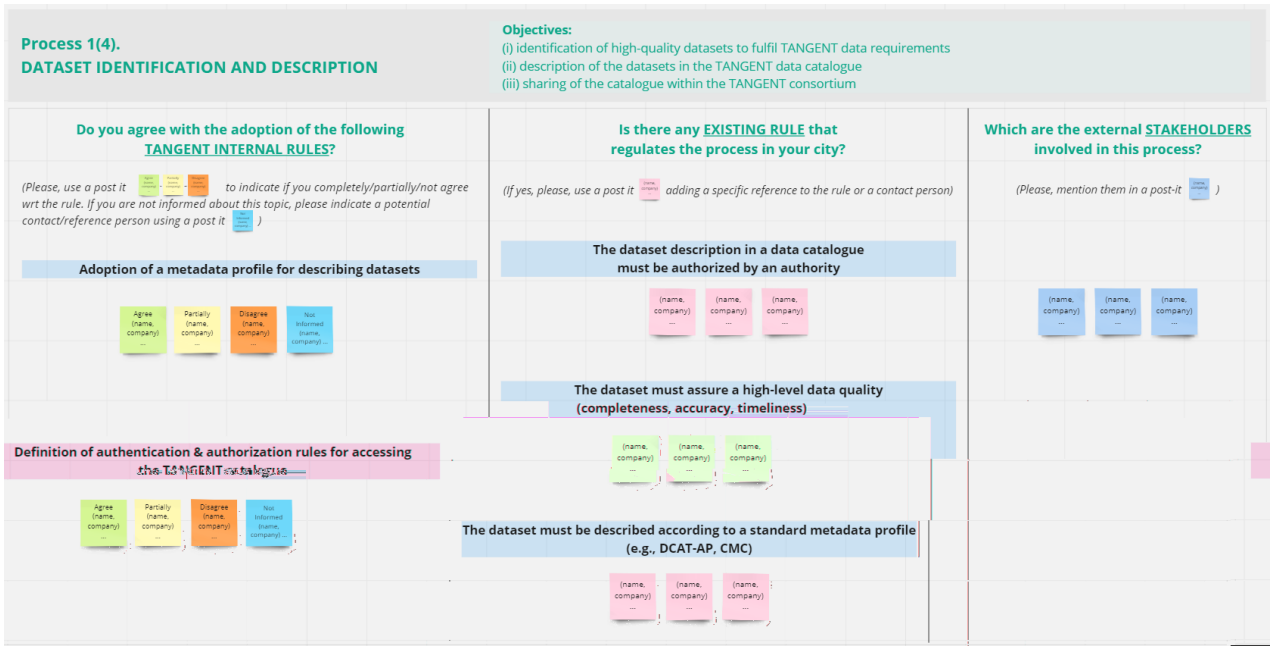

Figure 6: Miro Board frame adopted to collect information from stakeholders

In the working session, the participant was asked to complete the frames of the Miro Board by adding comments, notes, suggestions, or references through different post-it notes. For each case study, the Miro Board was left open to the participants after the workshop to let them finalize their contributions.

All the information and suggestions provided by the participants through the four Miro Boards were analysed. The analysis highlighted the need to identify agreements, roles, processes, and rules in the TANGENT Data-Sharing Governance Model. In particular, all participants agreed on the importance of:

- adopting a metadata profile for describing datasets;
- signing an NDA between TANGENT partners and each external data provider for enabling and regulating the access to private datasets. In case the data provider is a TANGENT partner, the access to private datasets is already regulated by the Consortium Agreement;
- signing an NDA between TANGENT partners and each external data provider for enabling and regulating the usage of private datasets. In case the data provider is a TANGENT partner, the usage of private datasets is already regulated by the Consortium Agreement;
- assuring a high-level of data quality of the dataset;
- ensuring the compliance with existing external rules and licences for handling the access, harmonisation and usage of datasets.

5.2 Data-Sharing Governance Model

From the analysis of the state-of-the-art in Section 2, the following elements resulted as key aspects to be taken into consideration for the definition of a Data-Sharing Governance Model:

data governance is characterised by different **objectives** (e.g., enhance findability and reuse of data);
data governance should consider **different data source types**;
data governance should consider **tools** and technologies adopted for data-sharing;
data governance refers to different **processes** (e.g., data publication, data access);
different **actors with different roles** are involved in the governance of each process;

different **rules** (both external and defined by the project) should be applied to establish effective data governance.

5.2.1 Objectives

The definition of the Data-Sharing Governance Model considered as input the following objectives of the project:

open and proprietary data sources should be collected with the direct involvement of data providers for the TANGENT case studies according to a set of data requirements (as defined in D2.1);
data-sharing among all the TANGENT partners should be fostered, defining a TANGENT Data Catalogue supporting the findability and accessibility of data sources;
WP2 could transform the data sources provided by each case study according to the harmonisation and fusion requirements elicited by TANGENT partners;
the data sources provided by each case study should be made available in raw format or transformed by WP2 to feed the TANGENT components developed by WP3-5 and to support the implementation of the TANGENT solution by WP6 and its testing in WP7.

5.2.2 Data source types and Tools

The TANGENT Data-Sharing Governance Model should consider the types of data that will be handled and the already identified tools for data sharing.

The following types of data sources should be considered:

- **Open Data Sources:** native/raw data sources already available in Open Data Portals and reused for the TANGENT project. Open data sources are published according to the terms and conditions of a licensing framework (e.g., attribution, creation of derivatives, commercial exploitation).
- **Proprietary Data Sources:** native/raw data sources made available for the TANGENT project through a specific agreement. Proprietary data sources can be accessed and used according to specific terms and conditions mutually agreed between the data owner and the data user.
- **Constrained Data Sources**: data sources with specific commercial constraints (e.g., data sources extracted or derived from the calibrated transport models) that will be used only by entitled users within the TANGENT project.
- **Internal Data Sources**: data sources processed or generated by the project (e.g., results of the TANGENT solution for data harmonisation and fusion).

The following tools, developed by WP2 for the TANGENT solution for data harmonisation and fusion, should be considered for the data-sharing governance:

- **TANGENT Data Catalogue**: data catalogue implementing a single solution for the findability and accessibility of all the data sources shared within the TANGENT project.
- **TANGENT Metadata Profile:** specification defining metadata fields for the description of the data sources in the TANGENT Data Catalogue (cf. Sect. 4.3).

5.2.3 Agreements and Licences

The TANGENT Data-Sharing Governance Model must consider and be compliant with the following documents:

- **TANGENT Consortium Agreement**: This document specifies the relationship among the beneficiaries of the TANGENT project, in particular concerning the organisation of the work between them, the management of the project and the rights and obligations of the partners concerning liability, access rights and dispute resolution. The CA supplements binding commitments among the beneficiaries of the TANGENT project in addition to the provisions of the Grant Agreement with the funding authority.
- **TANGENT D9.2 "Data Management Plan" (DMP)**: This document provides an overview of the main elements of the data management policy that are going to be used by the TANGENT consortium with regard to all the datasets that will be generated, harvested and/or used in the project. The DMP will be updated during the lifespan of the project. The first version has been released on M6. The TANGENT Data-Sharing Governance Model, proposed in this deliverable, will be the basis for the release of the second version of the DMP (M18).
- **TANGENT Non-Disclousure Agreement (NDA)**: This document represents a mutual confidentiality agreement between two Parties that desire to exchange and share information under the auspices of the TANGENT project and wish to protect certain confidential information which may be disclosed between them for the purpose of facilitating the development and testing of collaborative scenarios, test cases and pilot applications. The Parties agree to ensure such information is treated as confidential and protected in accordance with the terms of this agreement. TANGENT NDA will be defined and managed within WP7.
- **TANGENT Memorandum of Cooperation (MoC)**: This document represents an agreement of cooperation between stakeholders involved in a TANGENT case study. The MoC specifies that the cooperation could be related also to data-sharing for the development and testing of the TANGENT services. The MoC specifies how confidential information must be managed. TANGENT MoC will be defined and managed within WP7.
- **Data Licence**: licence associated to a specific data source to be managed by the TANGENT project that regulates its access and usage.
- **Metadata Licence**: the licence selected by the TANGENT project to be associated with the metadata collected through the TANGENT catalogue is Creative Commons CC-BY[37].

5.2.4 Data Governance Roles and Structures

Different roles and structures are involved in the TANGENT Data-Sharing Governance. They are categorised as external (i.e., already in place) and internal (i.e., defined by TANGENT). Each role is involved in specific data governance processes and is responsible for the application of one or more data governance rules.

External:

- **Data owner**: person/company that owns the data contained in an external data source. A data owner could be a TANGENT project partner or an external person/company.
- **Data publisher**: person/company responsible for publishing and describing an external data source in an external data catalogue/platform (e.g., a national access point (NAP) for mobility and transport data).

[37] https://creativecommons.org/licenses/by/4.0/

- **Data manager**: person/company (e.g., NAP operator) responsible for managing a (set of) external data source(s) in an external data catalogue/platform.
- **Data Protection Officer**: a person responsible for ensuring compliance with GDPR in managing and processing personal data in a specific company/public administration.

Internal:

- **TANGENT data user:** person accessing and using a data source within the TANGENT project (directly or through a software system).
- **TANGENT data publisher**: person responsible for publishing and describing a data source within the TANGENT Data Catalogue.
- **TANGENT Data Catalogue Management Board (TANGENT DCMB):** group of persons responsible for the management and control of a (set of) data source(s) within the TANGENT Data Catalogue.
- **TANGENT Data Access Control Board (TANGENT DACB)**: group of persons responsible for defining access control rules to a (set of) data source(s) shared in the TANGENT Data Catalogue.

To avoid misunderstanding, we provide two clarifications on roles already defined within the TANGENT project.

The **internal data manager** 2& & &J?48&J &S &V & & 2& & for uploading the materials (e.g., datasets, deliverables, presentations) produced by the project in Zenodo[38] is out-of-scope for the TANGENT Data-Sharing Governance Model.

The **TANGENT Data Protection Officer** role is defined in the TANGENT Consortium Agreement as a reference person appointed by each TANGENT beneficiary to assure compliance with GDPR in managing and processing personal data. The TANGENT Data-Sharing Governance Model does not focus on personal/sensitive data collected by the project (e.g., revealed preferences of users from the] V9& & &M &Z /4&Z & & & & & & & & &J?48&J & Management Plan - & 2&J 7647&N-Requ &T 47 & &he following versions.

5.2.5 Data Governance Processes and Rules

The TANGENT Data-Sharing Governance Model focuses on five different processes identified according to the described objectives, data types and tools:

Data publication: process of describing a data source in the TANGENT Data Catalogue;
Data quality: process of assessing the quality of a TANGENT data source and metadata published;
Data access: process of accessing data contained in/exposed from a data source;
Data storage: process of storing (a portion of) data provided from a data source;
Data usage: process of using and or manipulating data from one or more data sources.

[38] https://zenodo.org/

In this section, each process considered in the TANGENT Data-Sharing Governance Model is described through a set of data governance rules. Each rule is identified with a name and specified considering the involved data types, tools, roles, and structures.

5.2.5.1 Process 1: Data publication

The rules identified for the *data publication* process are:

- **Sharing of external data sources for the case study:** The external data sources (both open and proprietary) required for each case study should be published in the TANGENT Data Catalogue. The TANGENT case study leader assumes the role of TANGENT data publisher for all the external data sources related to a specific case study.
- **Compliance with TANGENT data requirements**: The TANGENT data publisher should publish data sources in compliance with the TANGENT data requirements defined in deliverable D2.1.
- **Sharing of internal data sources for the case study**: Only the internal data sources generated as a result of the TANGENT solution for data harmonisation and fusion should be published in the TANGENT Data Catalogue. The TANGENT WP2 leader assumes the role of TANGENT data publisher for these data sources.
- **Metadata description of data sources:** The TANGENT data publisher should describe a data source in the TANGENT Data Catalogue according to the TANGENT Metadata Profile defined to support the application of the *findability* FAIR principle (Sect. 3.1). The licence selected for metadata in the TANGENT Data Catalogue is Creative Commons CC-BY. Metadata requiring a stricter licence should not be added to the TANGENT Data Catalogue.
- **Compliance with terms and conditions:** The TANGENT data publisher should publish data sources in the TANGENT Data Catalogue following terms and conditions defined by the data owner. For open data sources, the adopted licensing framework should be considered. For proprietary data sources, specific terms and conditions should be agreed upon with the data owner. The TANGENT data publisher should explicitly indicate the terms and conditions for access and usage of the data source in its description.
- **Personal/sensitive data**: The TANGENT data publisher should collaborate with the Data Protection Officer to ensure that data sources containing personal/sensitive data are not published in the TANGENT Data Catalogue.
- **Constrained data**: The TANGENT data publisher should collaborate with the data owner to ensure that data sources with specific commercial constraints are not published in the TANGENT Data Catalogue.
- **Approval of data source publication**: The TANGENT DCMB validates the publication of data sources shared and described by the TANGENT data publisher. In particular, the TANGENT DCMB should ensure the publication of data sources in accordance with signed NDA and data licences.

5.2.5.2 Process 2: Data quality

The rules identified for the *data quality* process are:

Metadata quality management: The TANGENT DCMB, in collaboration with the TANGENT data publisher, should maintain a high quality of metadata in the TANGENT Data Catalogue to support the application of the *reusability* FAIR principle (Sect. 3.4). In particular, the quality

process should consider the following dimensions: accuracy, completeness, consistency, uniqueness, timeliness, and validity.

Data quality issue report: The TANGENT DCMB should inform the data owner and/or the data manager of data quality issues reported by TANGENT data users.

5.2.5.3 Process 3: Data access

The rules identified for the *data access* process are:

- **Metadata accessibility:** The TANGENT DCMB should guarantee access to the metadata description of data sources shared in the TANGENT Data Catalogue supporting the application of the *findability* FAIR principle, i.e., providing access to searchable metadata.
- **Data accessibility:** The TANGENT DCMB should guarantee access to data sources shared in the TANGENT Data Catalogue supporting the application of the *accessibility* FAIR principle, i.e., providing online access to data sources via a web browser and/or API.
- **Metadata access control**: The TANGENT DACB should manage access to TANGENT data users to all metadata of data sources shared in the TANGENT Data Catalogue.
- **Open data access control**: The TANGENT DACB should guarantee access to TANGENT data users to all public data sources shared in the TANGENT Data Catalogue.
- **Proprietary data access control**: The TANGENT DACB should guarantee access to proprietary data sources shared in the TANGENT Data Catalogue as follows:
    - If the data owner is a TANGENT partner, TANGENT data users can access private data sources according to the Consortium Agreement.
    - If the data owner is not a TANGENT partner, only TANGENT data users affiliated with an entity/company that signed a one-to-one agreement with the external data owner (formalised in an NDA with specific terms and conditions) can access the data source.

5.2.5.4 Process 4: Data storage

The rules identified for the *data storage* process are:

Data sources not online: The TANGENT DCMB should manage the storage of data sources not already available online.

Data storage management: The TANGENT DCMB should manage the processes of data storage in compliance with data external rules (e.g., security policies, restrictions on the data location).

Data storage duration: The TANGENT DCMB should ensure an appropriate data storage duration.

5.2.5.5 Process 5: Data usage

The rules identified for the *data usage* process are:

Data reuse: The TANGENT data publisher, in accordance with the data owner, should associate the data source with a data licence supporting the *reusability* FAIR principle.

Open data usage: The TANGENT data user should use open data sources shared in the TANGENT Data Catalogue according to their licensing framework.

Proprietary data usage: The TANGENT data user should use proprietary data sources shared in the TANGENT Data Catalogue as follows:

If the data owner is a TANGENT partner, TANGENT data users can use private data sources according to the Consortium Agreement.

If the data owner is not a TANGENT partner, only TANGENT data users affiliated with an entity/company that signed a one-to-one agreement with the external data owner (formalised in an NDA with specific terms and conditions) can use the data source.

Data transformation: The TANGENT data user should transform the data sources (e.g., change of format/standard to support the application of the *interoperability FAIR* principle) ensuring that the output of the transformation is consistent with the information contained in the input.

Data fusion: The TANGENT data user should merge the data sources ensuring that the final output is compliant with the more restrictive licence associated with the data sources.

- **Usage of constrained data sources**: Constrained data sources cannot be accessed and/or used by TANGENT data users and will be directly handled by the data owner within the TANGENT project.

6 Conclusions

Based on an extensive analysis of the literature related to data governance, this deliverable has provided a **Data Governance Reference Schema** describing the proper structures and processes to handle data-sharing. The reference schema has been used to guide a more tailored analysis of literature related to data governance in the mobility and transport domain. The comparison performed led to the identification of the following main **best practices** to be adopted in the TANGENT project:

- establishment of a technical/management board for supporting and facilitating the processes of data collection, publication, access, and sharing;
- definition of a metadata catalogue for supporting data-sharing;
- definition of a set of rules and actors for publishing, accessing, sharing, harmonizing and merging data and metadata;
- definition of a set of rules and actors for ensuring interoperability and re-use of data and metadata.

The literature analysis related to mobility data governance has been complemented with the study of the FAIR data principles and related initiatives to improve the findability, accessibility, interoperability, and re-use of the data.

Data-sharing in TANGENT will be supported through specific tools defined by WP2 as part of the TANGENT solution for data harmonisation and fusion: the TANGENT Data Catalogue and the TANGENT Metadata Profile. A data catalogue, i.e., a digital platform supporting the sharing, findability, and accessibility of distributed digital assets, will be developed within Task 2.3 and populated considering TANGENT data sources. This deliverable has discussed the design, implementation and publication of a **TANGENT Metadata Profile** for describing data sources in the TANGENT Data Catalogue. The profile, defined as an extension of DCAT-AP and compliant with the FAIR data principles, aims to increase and simplify data-sharing, discovery, and reusability. The profile relies on the reported initiatives and best practices for defining metadata specifications in the transport domain.

These propaedeutic activities and the inputs collected from stakeholders during the multi-actor & & & & V7& &J 749& Multi-actor cooperation models for NTM - First Release /& & & to define the main contribution of this deliverable: the **TANGENT Data-Sharing Governance model**. The model supports the strategic and operational management of data-sharing within the TANGENT project, focusing on the tasks/processes that are under the responsibility of TANGENT WP2: (i) provision of access to data sources needed by the technical components of the project; (ii) development of the TANGENT solution for data harmonisation and fusion.

The model identifies key processes (data publication, data quality, data access, data storage, data usage) to be addressed by TANGENT stakeholders with different roles and following specific rules. Table 10 summarises the defined TANGENT Data-Sharing Governance Model. The rules defined for each process are rearranged considering the identified roles: TANGENT data publisher, TANGENT Data Catalogue Management Board (TANGENT DCMB), TANGENT Data Access Control Board (TANGENT DACB), TANGENT data user.

Role	Process	Rule
TANGENT data publisher	Data publication	Sharing of external data sources for the case study
	Data publication	Compliance with TANGENT data requirements
	Data publication	Sharing of internal data sources for the case study
	Data publication	Metadata description of data sources
	Data publication	Compliance with terms and conditions
	Data publication	Personal/sensitive data
	Data publication	Constrained data
	Data usage	Data reuse
TANGENT DCMB	Data publication	Approval of data source publication
	Data quality	Metadata quality management
	Data quality	Data quality issue report
	Data access	Metadata accessibility
	Data access	Data accessibility
	Data storage	Data sources not online
	Data storage	Data storage management
	Data storage	Data storage duration
TANGENT DACB	Data access	Metadata access control
	Data access	Public data access control
	Data access	Private data access control
TANGENT data user	Data usage	Open data usage
	Data usage	Proprietary data usage
	Data usage	Data transformation
	Data usage	Data fusion
	Data usage	Usage of constrained data sources

Table 10: TANGENT Data-Sharing Governance Model.

7 References

1. G4l 2&G4L 2&S 4&Y 2&S 4l 2& &G4l 2&K &l -Border Travel U &Z &T &G &V &L & &S &N 2& &V & of the 3rd International Workshop Semantics And The Web For Transport, Online, September, Sep. 2021, vol. 2939. Accessed: Jun. 10, 2022. [Online]. Available: http://ceur-ws.org/Vol-2939/#paper6
2. G4&S & &Y4&H 2&YQUY& & & & & 2&. 9l 2&. 9l & recommendation, 2009. [Online]. Available: http://www.w3.org/TR/skos-reference
3. Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data Governance: A conceptual framework, structured review, and research agenda. International Journal of Information Management, 424-438.
4. Alex Ball (2014). How to License Research Data. Retrieved on 16th november 2022 from https://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How_To_License_Research_Data.pdf
5. Brous, P., Janssen, M., & Vilminko-Heikkinen, R. (2016). Coordinating Decision-Making in Data Management Activities: A Systematic Review of Data Governance Principles. International Conference on Electronic Government - EGOV 2016 (pp. 115-125). Springer Verlag.
6. Commission Delegated Regulation (EU) 2017/1926 of 31 May 2017 supplementing Directive 2010/40/EU of the European Parliament and of the Council with regard to the provision of EU-& & & & 4&S &867=4&G &P 4&762&8682. [Online]. Available: http://data.europa.eu/eli/reg_del/2017/1926/oj/eng
7. J 4& 2&P4&G4&X 2&V4& 2&P4& 2& &U 4& 2&G& & & & for representing public & & K & 4& 867; 4& aU 4& G & https://github.com/cef-oasis/DCAT-AP/tree/master/TransportDCAT-AP
8. Data Catalog Vocabulary (DCAT) - Version 2. https://www.w3.org/TR/vocab-dcat-2/ (accessed Jun. 10, 2022)
9. Data Governance Institute (2022) DGI Data Governance Framework Components. Retrieved on 20-10-2022 https://datagovernance.com/the-dgi-data-governance-framework/dgi-data-governance-framework-components/
10. DATEX II Standard (2022). Retrieved on 21-10-2022 https://www.datex2.eu/
11. DCAT Application Profile Version 2.1.1. https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/211 (accessed Nov. 18, 2022).
12. Docherty, I., Marsden, G., & Anable, J. (2018). The governance of smart mobility. Transportation Research Part A: Policy and Practice, 114-125.
13. E015 Digital Ecosystem (2022). Retrieved on 28-10-2022 https://www.e015.regione.lombardia.it
14. European Commission (2013) commission delegated regulation (eu) no 886/2013: Retrieved on 20-10-2022 https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32013R0886
15. European Commission (2015) commission delegated regulation (eu) 2015/962: Retrieved on 20-10-2022 https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32015R0962
16. European Commission (2016). H2020 Programme. Guidelines on FAIR Data Management in Horizon 2020. Retrieved on 16th November 2022 from

https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

17. European Commission (2017) commission delegated regulation (eu) 2017/1926: https://eur-lex.europa.eu/eli/reg_del/2017/1926/oj
18. European Commission (2020) European Data Strategy: EUR-Lex website. Retrieved on 20-10-2022 https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1593073685620&uri=CELEX%3A52020DC0066
19. European Commission (2020). Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions. A European strategy for data: EUR-Lex website. Retrieved on 20-10-2022 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066
20. European Commission (2022). BE OPEN - European forum and oBsErvatory for OPEN science in transport. Retrieved on 16th november 2022 from https://beopen-project.eu/storage/files/beopen-d22-open-fair-data-software-and-infrastructure-ineuropean-transport-research.pdf
21. European Commission (2020). Data Governance Act. Retrieved on 5th December 2022 from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52020PC0767
22. European Commission (2022). Horizon Europe (HORIZON). Programme Guide. Retrieved on 16th november 2022 from https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf
23. GeoDCAT-AP - Version 2.0.0. https://semiceu.github.io/GeoDCAT-AP/releases/2.0.0/ (accessed Nov. 18, 2022)
24. GeoJSON (2022). GeoJSON Open Standard Format. Retrieved on 20-10-2022 https://en.wikipedia.org/wiki/GeoJSON
25. GTFS Standard (2022). Retrieved on 20-10-2022 https://en.wikipedia.org/wiki/General_Transit_Feed_Specification
26. HARMONY (2019). H2020 European Project. Retrieved on 27-10-2022 https://harmony-h2020.eu/
27. Hirschhorn, F., Paulsson, A., Sørensen, C. H., & Veeneman, W. (2019). Public transport regimes and mobility as a service: Governance approaches in Amsterdam, Birmingham, and Helsinki. Transportation Research Part A: Policy and Practice, 178-191.
28. Hodson S, Jones S (2018). Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data. Report, ISBN Number ISBN 978-92-79-96546-3, URL https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf, DOI 10.2777/1524
29. Kovacikova, T., Böhm, M., Jones, S., et al. (2018). Analysis of the state of the art, barriers, needs and opportunities for setting up a transport research cloud : expert group final report. European Commission, Directorate-General for Research and Innovation, Publications Office, https://data.europa.eu/doi/10.2777/77906
30. Leire Serrano, Marco Comerio, Antonia Azzini (2022). D9.2 - Data Management Plan. First release. GRANT AGREEMENT 955273. TANGENT project.
31. LEMO (2017). H2020 European Project. Retrieved on 20-10-2022: https://lemo-h2020.eu/
32. S 4&Y 2&G4&G 2&V4&H 2&S 4& 2& &V4&R 2&Z & JI GZ-AP: Roadmap &X & & &Z &S &Y 2& &V & &V & &

Demo Track and Workshop Track of the 18th International Conference on Semantic Systems, Vienna, Austria, Sep. 2022, vol. 3235. Accessed: Nov. 21, 2022. [Online]. Available: https://ceur-ws.org/Vol-3235/#paper22

33. M. Scrocca, M. Come 2& &G4&G 2& G&X & & &J I GZ-GV&K 4&aU 4& Available: https://doi.org/10.5281/zenodo.7053406

34. S 4&Y 2&S 4& 2& & 2& TGVI UXK& 2&X &Y] M& 4 &] &O &8 4& Approach towards napDCAT-GV& 4&8688 4&aU 4&G &https://napcore.eu/wp-content/uploads/2022/08/NAPCORE-4.4.2.2-Approach-towards-napDCAT-AP-specification_v1.0-Report.pdf

35. Mark D. Wilkinson; Michel Dumontier; IJsbrand Jan Aalbersberg; et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 3 (1): 160018. doi:10.1038/SDATA.2016.18

36. McMurry JA, et al. (2017). Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLOS Biology 15(6): e2001414. https://doi.org/10.1371/journal.pbio.2001414

37. MOMENTUM (2019). H2020 European Project Retrieved on 21-10-2022 https://h2020-momentum.eu/

38. NOESIS (2017). H2020 European Project. Retrieved on 20-10-2022 http://noesis-project.eu/

39. NOESIS (2019). D4.2 Data Governance and institutional issues. Retrieved on 5th December 2022 from https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5c3ad11eb&appId=PPGMS

40. Open Geospatial Consortium (2022) Standard Geographical Markup Language. Retrieved on 20-10-2022 https://www.ogc.org/

41. Otto, B. (2011). A Morphology of the organisation of Data Governance. European Conference on Information Systems.

42. Otto, B., & Osterle, H. (2009). One Size Does Not Fit All - A Contingency Approach to the Data Governance. International Journal of Data and Information Quality.

43. V4&H 2&V4&R 2& & 2& TGVI UXK& 2&X &Y] M& 4 &] &O &8 47& X & G & & & & S & Y 4& 8688 4& aU 4& G & https://napcore.eu/wp-content/uploads/2022/08/NAPCORE-4.4.2.1-Requirement-analysis_v1.0.pdf

44. Panian, Z. (2010). Some Practical Experiences in Data Governance. International Journal of World Academy of Science, Engineering and Technology, 939-946.

45. RUDI Portal (2022). Rennes Urban Data Interface. Retrieved on 20-10-2022 https://rudi.datarennes.fr/

46. Shapefile (2022). Shapefile Data Format. Retrieved on 20-10-2022 https://en.wikipedia.org/wiki/Shapefile

47. SPRINT (2018). UITP European Project: Retrieved on 20-10-2022 http://sprint-transport.eu/Page.aspx?CAT=STANDARD&IdPage=e67a5661-c46a-4fb5-b116-8c6a469d878d

48. StatDCAT-AP - Version 1.0.1. https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/statdcat-application-profile-data-portals-europe/release/101 (accessed Nov. 18, 2022)

49. Swiss National Science Foundation SNSF (2022). Retrieved on 16th November 2022 from https://www.snf.ch/en/FAiWVH4WvpKvohw9/topic/research-policies

50. T. Berners-Lee, J. Hen 2& &U4R 2&Z &Y &] 2&Y &G 2& 48 : 2& no. 5, pp. 34 43, 2001

51. Transforming Transport (2017): H2020 European Project. Retrieved on 28-10-2022 https://cordis.europa.eu/project/id/731932/it

52. TransformingTransport (2019). D3.9 TransformingTransport Open Data Portal. Retrieved on 5th December 2022 from https://transformingtransport.eu/sites/default/files/2017-08/D3.9%20-%20Open_Data_Portal-v1.0.pdf

53. Wende, K. (2007). A Model for Data Governance Organising Accountabilities for Data Quality Management. International Conference on Information Systems (pp. 416-425). Toowoomba : Association for Information Systems, AIS Electronic Library (AISeL).

54. Wende, K., & Otto, B. (2007). A Contingency Approach to Data Governance. International Conference on Information Quality, (pp. 163-176). Cambridge, USA.

55. Wilkinson, M., Sansone, SA., Schultes, E. et al. (2018). A design framework and exemplar metrics for FAIRness. Sci Data 5, 180118. https://doi.org/10.1038/sdata.2018.118