PI Data Sharing Infrastructure

Wout Hofman
TNO, the Hague, the Netherlands
(wout.hofman@tno.nl)

Abstract: Data sharing is key to realizing the Physical Internet (PI) as its roadmap illustrates. Examples are for full visibility of nodes and their business services, adjustment and assignment of routing to reduce empty miles, and protocols and services for operational efficiency of logistics networks. These examples and many more existing ones are use cases for a PI Data Sharing Infrastructure that can support current and future data sharing requirements. One aspect of the PI is its organizational structure as a network, where logistics is a common resource used dynamically. These imply that any data sharing infrastructure must be open, flexible, and extendible for innovative use cases. Semantics and data sharing functionality is the core for such an infrastructure. By applying semantic web standard and – technology, the objectives can be reached, and a large variety of use cases can be supported. A set of agreements, or protocol stack, is proposed including an approach for governance.

Keywords: *data sharing, data spaces, Physical Internet, federation, open and neutral data sharing infrastructure.*

Physical Internet Roadmap: PI Networks, System of Logistics Networks, Governance.

1 Introduction

The Physical Internet roadmap (Alice, 2022) proposes five phases for constructing the Physical Internet where each of the phases has its own roadmap. Seamless, interconnected transport networks adaptive to change need to be constructed, including governance. One of the main aspects is full visibility, accessibility, use of business services for optimization, and situational awareness for optimal routing (De Juncker, 2023). Data sharing is a prerequisite to realize the Physical Internet, where data is not always public available but needs to be validated against certain criteria (Eckartz, Hofman, & Veenstra, 2014) to address data sovereignty (Dalmolen, et al., 2019). An open, neutral data sharing infrastructure is required where all logistics stakeholders can share data in a controlled way, without prior (bilateral) agreements (Digital Transport and Logistics Forum (DTLF) Subgroup 2: Corridor Information Systems, 2018). It must be flexible and extendible to support current and future data sharing requirements.

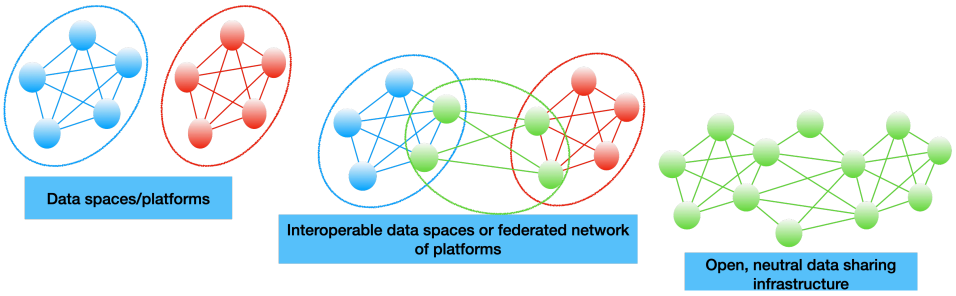
Data spaces/platforms
Interoperable data spaces or federated network of platforms
Open, neutral data sharing infrastructure

Figure 1 – From data spaces/platforms towards an open, neutral data sharing infrastructure

A so-called mobility data space needs to be constructed, according to agreed principles (Nagel & Lycklama, April 2021). Such principles result in functionality that must be supported, where this functionality is data agnostic (Nagel & Lycklama, April 2021). So-called Data Domain Standards can be implemented by a data space, leading to data spaces that are not necessarily interoperable as argued by the Digital Transport and Logistics Forum (The Digital Transport and Logistics Forum (DTLF), 2017). The Physical Internet however requires preferably a global data space with potentially local and/or mode specific sub-spaces (Figure 1). A data sharing infrastructure like the Internet, energy network, or road network is required, where this infrastructure can take different technical appearances, but still be open.

To achieve the objective shown in the previous figure, we consider interoperability models (Wang, Tolk, & Wang, 2009), (European Commission, Belgium) that can be related as shown below. The figure shows a requirement for a legal basis across different national domains and governance of the results as part of the EIF (European Interoperability Framework) that is lacking in the other model, whereas the latter one takes a more detailed approach by addressing conceptual interoperability. The latter is required for constructing a PI Data Sharing Infrastructure.

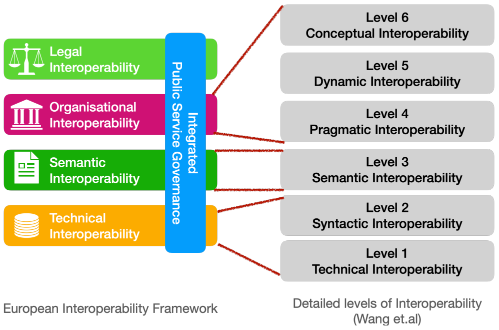
Legal Interoperability
Organisational Interoperability
Semantic Interoperability
Technical Interoperability
Integrated Public Service Governance
Level 6 Conceptual Interoperability
Level 5 Dynamic Interoperability
Level 4 Pragmatic Interoperability
Level 3 Semantic Interoperability
Level 2 Syntactic Interoperability
Level 1 Technical Interoperability
European Interoperability Framework
Detailed levels of Interoperability (Wang et.al)

Figure 2 – Interoperability models

The core of this paper is a proposal for a protocol stack to achieve conceptual interoperability and addresses legal aspects and governance as indicated by the EIF. First, the protocol stack will be discussed, secondly governance and legal aspects are presented. The proposed protocol stack supports business collaboration and compliance, but can also be applied for other types of use cases.

2 Protocol stack

To meet the objective, a set of agreements must be constructed. This is called the 'protocol stack'. This section presents the protocol stacks and focusses on gaps that are not addressed by other initiatives.

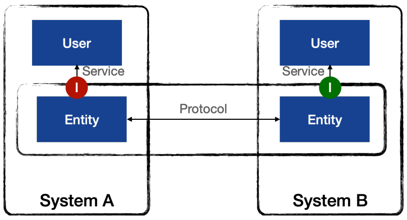

Figure 3 protocol, service, and interface

2.1 Protocol, service, interface

The concepts of protocol, service, and interface date back to the Open Systems Interconnection model that is the basis for the Internet (Tanenbaum, 1996). A **protocol** is a structured sequencing of interactions between two peer entities by different systems (or organizations). These 'entities' that are software components implementing a protocol, provide a **service** via an **interface** to a user. Protocol and service must be identical for any two implementations; each implementation can have a different interface.

Protocol specifications must be concise, consistent, coherent, and complete to enable implementations of a protocol by different providers. Separation of concerns is crucial for modularization. A protocol uses a lower layer service. Any software component must at least implement one protocol layer but can implement more. For instance, an endpoint of an openAPI is an interface of a service that uses Internet protocols for lower layer protocols. A messaging client implements the messaging protocol over Internet protocols.

2.2 Protocol stack

The protocol stack (Figure 4) consists of protocol layers specifying behavior implemented by two roles as defined by the EU Data Act: 'data holder' and 'data user'.

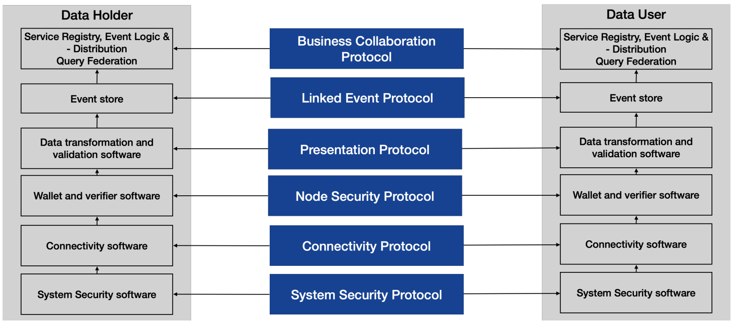
Data Holder
Service Registry, Event Logic & - Distribution Query Federation
Event store
Data transformation and validation software
Wallet and verifier software
Connectivity software
System Security software
Business Collaboration Protocol
Linked Event Protocol
Presentation Protocol
Node Security Protocol
Connectivity Protocol
System Security Protocol
Data User
Service Registry, Event Logic & - Distribution Query Federation
Event store
Data transformation and validation software
Wallet and verifier software
Connectivity software
System Security software

Figure 4 PI protocol stack

The protocol layers are:

- **Business collaboration protocol** – the capability to discover business services and assess organizational profiles of peers with their data capabilities and – requirements and share data for a business activity. Each peer entity of this protocol must implement at least a business activity and its interaction patterns as specified by the design and its configuration of that activity. These will be elaborated in this paper.
- **Linked Event protocol (pull)** – each interaction in a business collaboration protocol is implemented by sharing only links to additional data as specified by subtypes of 'event' in a design. Each link can be evaluated by a (standardized) query. Additionally, each user of the service of this protocol can formulate its own queries according to the multimodal ontology.
- **Presentation protocol(s)** – the syntax and technology (messaging, (open/webhook) APIs (Application Programming Interfaces) with JSON(-LD) (Java Script Object Notation – Linked Data), semantic web protocols (SPARQL (Standard Protocol and RDF Query Language), RDF (Resource Description Framework))) used for sharing data. These are technical aspects of the implementation of the upper layer protocols.
- **Node Security protocol** – it is about identification and authentication: the capability of nodes to verify each other's credentials applying open standards like OAUTH2.1, JWT (JSON Web Tokens), and/or Verifiable Credentials (VCs) and Decentralized Identifiers (DIDs).
- **Connectivity protocol(s)** – the technical capability for reliable, safe, and secure data sharing using a System Security Protocol. Current list of connectivity protocols: FENIX connector protocol, IDSA connector protocol, EDS (Eclipse Data Space) connector of GAIA-X, a large variety of blockchain protocols (e.g. Corda, Hyperledger Fabric, and Baseline protocol), and AS4 implemented by CEF eDelivery.

- **System Security protocol(s)** – the safe and secure sharing of data with PKI certificates, utilizing standard protocols (e.g. https, TLS).

These layers can be mapped against the interoperability models as follows:

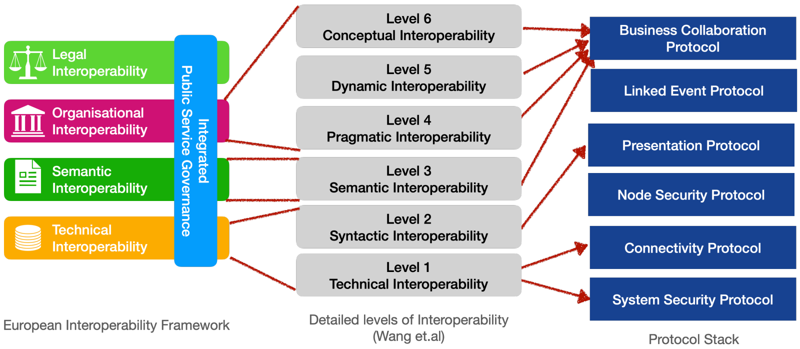
Legal Interoperability
Organisational Interoperability
Semantic Interoperability
Technical Interoperability
Integrated Public Service Governance
European Interoperability Framework
Level 6 Conceptual Interoperability
Level 5 Dynamic Interoperability
Level 4 Pragmatic Interoperability
Level 3 Semantic Interoperability
Level 2 Syntactic Interoperability
Level 1 Technical Interoperability
Detailed levels of Interoperability (Wang et.al)
Business Collaboration Protocol
Linked Event Protocol
Presentation Protocol
Node Security Protocol
Connectivity Protocol
System Security Protocol
Protocol Stack

Figure 5 protocol stack and interoperability layers

Note that two protocol layers are additional to the interoperability layers. The Node Security Protocol is required in an open, neutral data sharing infrastructure providing identity and authentication. Some implementations of the connectivity protocol implement their proprietary node security protocol, e.g. Corda. Other implementations support combinations of the various protocols, but data agnostic and potential in a different manner (e.g. IDSA – or Eclipse Data Space connectors). Differences in implementation could give an indication of the quality of a protocol design: there are too many degrees of freedom. The Linked Event Protocol is about optimizing data storage and contributes to data quality. There is always a single source of truth.

2.3 Business collaboration protocol

As the previous figure shows, this protocol implements four interoperability levels as identified by (Wang, Tolk, & Wang, 2009) and two of the EIF (European Commission, Belgium). This is achieved by modeling interaction patterns for business activities as constraints to a multimodal data sharing model. The latter is an alignment of existing mode and/or cargo specific models as specified by for instance industry associations and regulators.

The multimodal data sharing model is a so-called upper ontology. Mode -, cargo -, and/or infrastructure specific ontologies are lower ontologies. Alignment between those lower ontologies is via the upper ontology. Of course, individual lower ontologies can also be aligned (Euzenat & Shvaiko, 2010). Having a single upper ontology for alignment allows individual users to select required functionality of one or more lower ontologies.

Alignment is on concepts representing the physical world, like a taxonomy of Digital Twins with subtypes like container, truck, and vessel, and infrastructural aspects like locations, hubs, and road infrastructure, complemented with actors like a legal entity or a person. These all have associations in place and time, represented by event. Other concepts are the data sharing concepts: business activity their interaction patterns. A pattern reflects a Business Process Modelling (BPMn2.0) choreography (Object Management Group, 2011) that consists of states and state transitions triggered by interactions. These concepts are all represented as subtype of ‘event’. An interaction or a business document is for instance a subtype of event associating Digital Twins, locations, and organizations for a business activity. The data sharing concepts are a separate module of the ontology. A choreography is represented by states and state transitions triggered by events with links to data. States specify access policies, implying that a minimal and maximal data set can be retrieved given a state for interactions. Events are actually shared between a data holder and -user to synchronize their states.

The following figure visualizes the multimodal data sharing ontology (data sharing concepts represented as blue circles; physical and administrative concepts represented as yellow circles). Not only each interaction pattern has a start and end state, also a business activity must have these states. In case of a sequence of interaction patterns, like booking, ordering, and visibility, each interaction patterns adds state data to a business activity start state resulting in its end state.

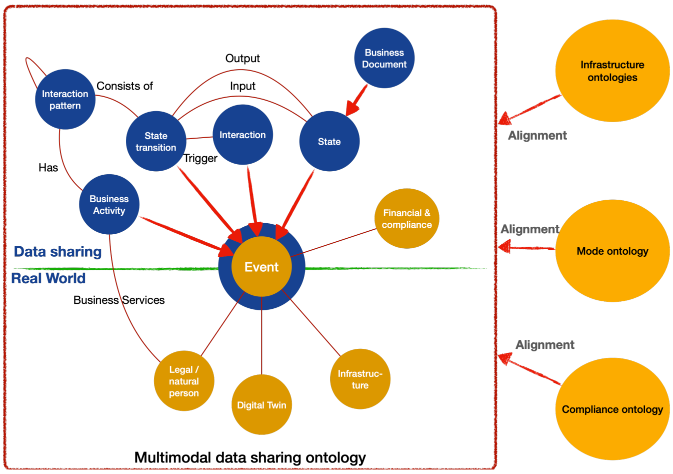

Figure 6 – the multimodal ontology with details of business data sharing concepts

Semantics itself must be machine-readable with open standards. Those are the semantic web standards like Ontology Web Language (OWL), Resource Description Framework (RDF), and SHApe Constraint Language (SHACL) (Berners-Lee, 2006). Modelling data sharing concepts representing process aspects (i.e. business process collaboration) allows the specification of minimal data requirements for states and events represented by SHACL for data quality validation (correctness and completeness); state transitions must be modelled by pre- and post-conditions and firing rules (Hee, 1994) resulting in executable event logic.

2.4 Configuring the business collaboration protocol – Service Registry

Each logistics stakeholder and authority will have its own data capabilities and requirements. These are based on the concepts of the multimodal ontology, for instance a regulation for risk assessment and taxation of incoming goods (transport) into the EU by customs or the transport of cargo by a shipping line. These specific capabilities and requirements must be specified, published, and discoverable. This is supported by a distributed Service Registry. Each organization thus has its own Service Registry. A Service Registry enables any organization:

- to specify its data requirements and
- to define and publish its business services for discoverability.

The data structure of the Service Registry enables a user to formulate interaction patterns for business activities; the blue circles shown in Figure 6. The Service Registry can be applied in two ways:

- **Design** – to specify business activities and their interaction patterns. Industry associations, communities, and regulators can do design. Subtypes of business activities and interaction patterns can be provided as an ad hoc standard and be applied for design to improve alignment.
- **Configuration** – to specify an organizational profile by searching and:
  - selecting those parts of a design that are relevant to an organization,

- specifying its business services and electing the various lower layer protocols it supports (including endpoints). Business activities and their interaction patterns define constraints to the multimodal ontology.

In addition, a user selects the constraints applicable for its organization i.e., selecting the relevant logistic Digital Twins applicable to its organization.

Additionally, Industry Associations, and regulators can align their ontology with the upper ontology, applying standard alignment tools. Alignment requires using the same open standards for representing data schemes, which is not always the case. Alignment is also on all concepts of the upper ontology; for instance, a data scheme for an electronic business document only aligns with that concept of the upper ontology and potentially a business activity (e.g. transport for an eCMR).

The configurations also specify access control: each organization is only able to provide access to data that is stored. Access can of course only be given to data of which links are shared by events.

Discoverability is implemented at technical - and business level, based on known SPARQL endpoints of Service Registries. Technical level is about re-use of a design to construct a configuration with business services, resulting in an organizational profile.

Any two stakeholders can share data for those parts of their organizational profile that are common. They can do business digitally (and be compliant) if goals and business services can be matched, which is established as part of the business collaboration protocol. The more stakeholders implement of a maximal design for a business activity in terms of interaction patterns, the more their business can be supported digital and seamless data sharing is achieved.

To enable migration, the business collaboration protocol can be implemented by both APIs and semantic technology. Therefore, the Service Registry will produce openAPIs and SHACL for implementation by an index. First examples of such generated APIs and SHACL are available.

2.5 Index functionality

The business collaboration protocol utilizes the Service Registry for configuring peer-to-peer data sharing; index functionality is about the events with links to data that are shared between a data holder and -user. An index of an organization contains all events (with links to data) send as data holder with other organizations and received as data user from data holders.

The functionality consists of the following components:

- data quality validation (correctness and completeness of event data and query (results)),
- event logic (validating the sequence of events),
- event storage (storing shared events)
- event distribution (sharing an event with the proper data holder(s)),
- enable access for replying to data users queries (link-based access control), and
- query federation (data provenance).

Having an event with a link to additional data implies that a data user is authorized by a data holder to access data of that link. The access policies specified by states specify the data that will be made available. The data that will be made available upon a query depends on the present state of interactions as stored by a data holder. A data user will not necessarily have the same state, a customs authority for instance may only be aware that a transport movement was started and does not know the latest state of that movement.

An index supports event distribution (sharing an event with the proper data holder(s)) based on input of a data holder initiating a commercial relation, the existing of a commercial relation (previous events are stored by an Index), or for legal compliance.

An index must support data quality validation (correctness and completeness of event data and query (results)) and either in its internal IT systems or by its index. Data quality is specified by an organization profile (see Service Registry).

The functionality mentioned here can be accessible via open- and webhookAPIs and semantic technology. Each organization must make a choice, where semantic technology is the preferred choice since it enables a more open way to query for data on states and has flexibility in access policies and authorization.

2.6 Identity, Authentication, and Authorization (IAA)

IAA is about trust in access to (links to) data. The data is business data (e.g., order data), a design, or an organization profile. IAA relates to authorization of users, i.e. employees of a participant, and architectural components (Service Registry and Index) that provide (access to) data. Safe and secure data transfer is addressed separately by connectivity protocols for the Index.

IAA is built upon two pillars[1]:

- **Organizational trust** – each organization that requires to be a node must implement measures that assure trust, for instance cyber security measures and an Identity and Access Management (IAM) registry. Rules for creating this type of trust will be formulated by a legal framework. It also covers authorization of employees to act on behalf of its employer and non-repudiation of actions taken by these employees.
- **Inter-organizational trust** – each organization must share an identity with another organization that can be verified by that other organization when sharing events, queries, and/or query results.

Authorization is internal to each organization and is the basis for access control. Organizations thus do not know authorized users of other organizations; they trust that authorization is properly implemented by others (organizational trust).

Each node must have at least one endpoint with inter-organizational trust (Identity and Authentication); it may have multiple ones (e.g. one for its business services and another one for data sharing). Identity and authentication must be based on a completely distributed solution based on which is provided and governed by:

- a regulator (providing– establishing a legal data sharing framework (e.g. EC) and accreditation of registration authorities,
- a trusted registration authority acting as issuer of verifiable credentials, and
- a certification body for organizational trust.

The implementation of such a distributed solution with Verifiable Credentials (VCs) and Decentralized Identifiers (DIDs) is still under development. A registration authority may for instance have a DID issued by a regulator allowing it to act as issuer after a potential participant is certified. Any two nodes that share data may set up a persistent channel with DIDs after verifying each other's VCs. The latter only needs to be done once, which limits the number of

[1] There is also trust at business level, i.e. the trust in properly executing business activities for customers according to agreements made with them. This trust is outside scope of IAA.

interactions and improves performance. Existing standards, and solutions (like OAUTH2.1) can be applied to create inter-organizational trust (applicable to data of a Service Registry and an Index). This intermediate level requires one or multiple Identity Brokers acting as intermediate Registration Authorities. Preferably, a regulator is a public body and different roles are implemented by different organizations (separation of concern).

3 Governance and legal aspects

The proposed approach has implications for standardization and governance. Configuration of the PI Data Sharing Infrastructure is distributed if these configurations adhere to a meta level standard and are implemented via proposed governance agreements. This requires a legal framework for creating trust. This section briefly elaborates governance and potential legal aspects.

3.1 Governance

Governance is basically on the protocol stack and the upper ontology. The protocol stack consists of various elements that are prone to standardization, are already (based on) open standards, or already have a governance structure. For instance, connectivity -, node security - , and presentation protocols are already based on open (or defacto) standards.

The following elements of the protocol stack are generic and prone to standardization:

- **Multimodal data sharing ontology** – the upper ontology for data sharing supporting business activities and compliance with concepts like Digital Twins, events, states, and state transitions.
- **Interaction patterns** – a set of interaction patterns to support commercial transactions or compliance-based data sharing that specify access control policies.
- **Linked Event Protocol** – the way of sharing events with links to data.

Any other aspects are subject to further governance, which may result in additional standards.

Since the upper ontology creates an open PI data sharing infrastructure, alignment procedures need to be established and implemented. An example of such a procedure is that any concept or data property that is common to two or more designers is part of the upper ontology. Furthermore, particular Industry Associations, Communities, and regulators must be recognized designers, i.e. they must represent the interests of a number of users. Any designer can utilize the upper ontology for innovative applications by creating a lower ontology that can later be proposed as part of the upper ontology.

Governance of these rules requires a governance board, where recognized designers and logistics stakeholders (enterprises and authorities) collaborate. A support organization and advisory board will support the governance board for direction and daily operation.

3.2 Legal aspects

Creating an open PI data sharing infrastructure, which is a complex system, requires regulation relevant for its correct and trusted operation. As IAA identified, there is a separation between organizational – and inter-organizational trust. Additionally, also the behavior of any organization that requires to participate in the open PI data sharing infrastructure must be validated. The approach can be formulated as follows:

- **Organizational trust** – each organization must implement a list of applicable acts, regulations (list to be provided), and functionality for accountability. The latter is internal IAA and non-repudiation functionality like logs and audit trails.

- **Behavior** – the behavior of each organization as specified by its organizational profile must be published and validated or certified. In case of enterprises, it must include all relevant public and private compliance aspects for which data sharing requirements are published (e.g. private rules like The Hague-Visby rules). Behavior also must include data quality aspects (completeness, correctness, timelines, etc.). In case an organization utilizes a third party (e.g. a platform), that third party acts on behalf of its customers.
- **(Continuous) monitoring** – organizational trust and behavior requires monitoring, either continuously or periodically.
- **Change management** – any changes to an organizational profile in terms of its behavior must be validated and/or certified. This enables the extension of the infrastructure with new functionality.

Of course, this is just a simplified outline for which a legal framework can be formulated. A certification authority can perform continuous or periodical monitoring as a basis for a registration authority for issuing a VC/DID. Such a certification authority can be supported by an (online) testing, validation, and certification environment.

4 Conclusions and future work

This paper proposes a set of agreements, the protocol stack, as a means for constructing the PI Data Sharing Infrastructure. The upper layer protocols are given special attention, since these specify semantics and data sharing functionality by alignment of specific developments by Industry Associations, communities, and regulators. To facilitate alignment, semantic web standards are proposed, modeling alignment concepts like Digital Twins, events, states, and state transitions. This so-called upper ontology for multimodal data sharing is the basis for individual organizations to specify their data sharing requirements and – capabilities, implement these by for instance Application Programming Interfaces (APIs) or semantic web technology, and develop data quality validation solutions. The latter is the way forward, since it supports all potential queries of a data user that don't require standardization. Thus, it provides flexibility. Alignment and the support of different technologies contributes to adoption, which requires an adoption and migration strategy.

Adoption requires more attention, combined with governance and legal aspects. Various stakeholders of different domains need to be involved like Industry Associations, communities, regulators, and public (and private) policy makers at national, EU, and global level (e.g. EC DG CNECT/Move/Agri/etc., WEF (World Economic Forum), UN CEFACT, WCC (World Customs Council), and the World Bank). Most of them still take the traditional approach to data sharing based on pushing data or implementing a subscription method. These approaches are supported by new technology like blockchains with NFTs (Non-Fungible Tokens) and VCs (Verifiable Credentials). An overarching approach is required that considers all developments and integrates them into the proposed approach in this paper. Innovation in standardization is required, involving proper standardization bodies. One observation from practice that requires more attention is that private initiatives are competing and do not lead to an open data sharing infrastructure. A public initiative must be taken.

And still more technical work needs to be done like exploring the potential of so-called Large Language Models like chatGPT for alignment, matching, and query formulation. Also a prototype of the infrastructure supporting one or more (artificial) use cases must be developed (first prototypes of components are available via FEDeRATED (federatedplatforms.eu)). At the same time, already existing implementation initiatives must be coordinated and supported to realize the infrastructure.

Acknowledgements

This paper has been made possible by the CEF FEDeRATED Action and the Digital Transport and Logistics Forum.

References

- Digital Transport and Logistics Forum (DTLF) Subgroup 2: Corridor Information Systems. (2018, June). *Enabling organizations to reap the benefits of data sharing in logistics and supply chains - executive summary.* Retrieved from Digital transport and logistics forum: https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeeting&meetingId=4855
- Berners-Lee, T. (2006). *Artificial Intelligence and the Semantic Web: AAAI 2006 Keynote.* Retrieved from World Wide Web consortium: www.w3.org/2006/talks/0718-aaai-tbl/overview.html
- Object Management Group. (2011). *Business Process Model and Notation Specification 2.0.* Retrieved from https://www.omg.org/spec/BPMN/2.0/
- Tanenbaum, A. S. (1996). *Computer Networks (Third Edition).* Prentice Hall.
- Alice. (2022, 01 17). *Roadmap to the Physical Internet.* Retrieved from etp-alice.eu: https://www.etp-logistics.eu/alice-physical-internet-roadmap-released/
- De Juncker, M. (2023). Optimising Routing in an Agent-Centric Synchromodal Network with Shared Information. In *Optimisation in Synchromodal Logistics* (pp. 171-185). Springer.
- Eckartz, S., Hofman, W., & Veenstra, A. F. (2014). A Decision Model for Data Sharing. *eGov conference.*
- Dalmolen, S., Bastiaansen, H., Somers, E., Djafary, S., Kollenstart, M., & Punter, M. (2019). Maintaining control over sensitive data in the Physical Internet: towards an open, service oriented, network-model for infrastructural data sovereignthy. *International Physical Internet Conference (IPIC2019).* Londen.
- The Digital Transport and Logistics Forum (DTLF). (2017). *An outline for a generic concept for an innovative approach to interoperability in supply and logistics chains.* Discussion Paper, EC DG Move, Brussels.
- Nagel, L., & Lycklama, D. (. (April 2021). *Design Principles for Data Spaces - position paper.* Open DEI.
- Wang, W., Tolk, A., & Wang, W. (2009). The levels of conceptual interoperability model: applying systems engineering principles to M&S. *Spring Simulation Multiconference.* Society for Computer Simulation International.
- European Commission. (Belgium). *New European Interoperability Framework - promoting seamless services and data flows for European Public administrations.* 2017: European Union.
- International Data Spaces Association. (April 2019). *Reference Architecture Model - version 3.0.* Berlin, Germaniy: International Data Spaces Association.
- Euzenat, J., & Shvaiko, P. (2010). *Ontology Matching.* Heidelberg: Springer-Verlag.
- Hee, K. (1994). *Information Systems Engineering - a Formal Approach.* Cambridge University Press.