TNO innovation for life

CoE DSC

Orchestrating Privacy-Enhancing Technologies in a Data Space

TNO Public ⟩ TNO 2024 R10910
13 mei 2024

ICT, Strategy & Policy
www.tno.nl

TNO 2024 R10910 – 13 mei 2024

Orchestrating Privacy-Enhancing Technologies in a Data Space

Auteurs	P.N. (Peter) Langenkamp, S.H.M.(Stefan) van den Berg, A.M. (Abhishek Mahadevan) Raju, T.R. (Thomas) Nijman, M.A.N.E. (Marie-Beth) van Egmond

Rubricering rapport	TNO Public
Titel	TNO Public
Rapporttekst	TNO Public
Aantal pagina's	4 (excl. voor- en achterblad)
Aantal bijlagen	0
Projectnaam	Centre of Excellence for Data Sharing and Cloud
Projectnummer	060.54832

Inhoudsopgave

1 Management Summary

Combining data from different sources enables better analyses, which could offer enormous economical and societal value to sectors like health care, public administration and in finance. However, the sharing of data is (justly) not always allowed nor desirable e.g. due to privacy legislation. Recent advances in Privacy-Enhancing Technologies offer powerful techniques to address these challenges. These techniques make it possible to perform the required analyses on combined data whilst preserving a high privacy level.

In recent years, there has been a lot of attention for PET technologies, both in academics as in industry. In applied scientific research, many use cases have been thought of and algorithms have been adjusted to use in a privacy-enhancing way. However, in order to be able to really use PETs effectively in the future, a new direction of research is needed. The focus should not only be on the technology itself, but also on how parties actually collaborate. A promising option to do this is to integrate PETs into a Data Space.

In this report we show what such an integration could look like, using concepts from the research on Data Spaces. This gives a view on the use of PETs beyond merely the technology and should encourage potential future users of PETs to think about which party can fulfill what role in the orchestration of PETs.
We expect the use of PETs within Data Spaces will be a major topic in R&D projects in the future. This topic will therefore stay on the agenda of TNO and in the Center of Excellence for Data Sharing in Cloud in the coming years.

2 PETs & Data Spaces

In this report the use of PETs in a data space is explored. It is meant for anyone who is interested in the use of PETs in a real-world setting. Throughout the document, a collaboration between banks is used as illustration for the concepts explained. Note that this is merely an example and the concepts can be applied to many other domains, such as health care.

Note that the focus is the orchestration of the use of PETs in a data space. The basic processes such as onboarding are out of scope. These can for example be found in the joint white paper by the BDVA and CoE DSC. **Error! Reference source not found.**

This report has the following structure. In chapter 2, the concepts of PETs and Data Spaces will be explained. Next, we will look at a role model presented by the Netherlands AI Coalition and adjust this model to the use of PETs. We conclude with some open research questions.
Before we explain the synergy between PETs and Data Spaces, we will first explain the concepts separately.

2.1 Privacy-Enhancing Technologies (PETs)

Sharing and analyzing data are essential for achieving economic growth and addressing societal challenges. Big data has brought many new insights, however, a growing awareness about privacy and commercial and/or legal obstacles has resulted in new challenges to sharing data (legal, ethical) [1] . To adhere to the law and our privacy values, Privacy Enhancing Technologies (PETs) have become an important field. [16]

PETs encompass a diverse range of tools and methodologies geared towards ensuring that data can be used for its intended purpose without compromising the privacy of individuals. These technologies come into play across various domains, from securing communication channels to enabling confidential data analysis. The tools are especially useful in the arsenal of a data analyst. Allowing the analyst to gain the insights, without learning all the data. Some examples of PETs are Federated Learning, Zero Knowledge Proofs and Secure Multi-Party Computation (MPC).

TNO researches the potential use of PETs in many different domains, such as the health care sector, government or the financial sector. In these domains the sharing of data could improve the analysis performed. However the data involved is sensitive and cannot be shared due to privacy concerns. PETs can offer a solution to get the insights without the need to share the data.

2.2 Data Spaces

Before delving into specific applications, it's crucial to understand the fundamentals of data spaces and the challenges they address. Many national and international (EU) initiatives are working on Data Space architectures. Most notably, The Data Spaces Support Centre (DSSC), backed by the European Commission, strives to develop sovereign, interoperable data spaces, enhancing data sharing across sectors in alignment with EU values and economic and societal goals. For a more in-depth view on data space architectures, we refer to the DSSC blueprint [3].
We offer a practical overview, focusing on the technical aspects of data spaces, rather than a comprehensive discussion. The goal is not to provide the complete picture for the motivation of data spaces. For this, refer to the EU Data Strategy.

Data sharing between businesses is often complex. Companies must negotiate legal terms, align business processes, and determine technical details (like API specifications). Traditionally, this is a bilateral, process, specific for those processes and organisations involved. As a result, it does not scale well when more organisations or different processes are involved. The alternative is platform-based data sharing, where a single IT corporation (e.g. a Trusted Third Party) manages the entire process, including governance and technical aspects. While this method offers convenience and less complexity, it often requires clients to partially surrender their data control, compromising data sovereignty.

Data spaces represent a significant evolution in data sharing. They are decentralized networks operating under a unified legal and governance framework for sovereign data exchange. This prevents participants having to make separate agreements with each other party. In data spaces, data always moves directly from the provider to the consumer, maintaining data sovereignty and flexibility. A data space has a standardized control plane for the discovery of data, the negotiation of data usage terms, and the management of the state of a data exchange.

2.2.1 Data Spaces: From building blocks to components

The DSSC Blueprint[3] outlines the essential components for developing a functional data space, categorizing them into two groups:

- Organizational and Business Building Blocks
  These address the organizational, business, and legal aspects, such as:
  - Defining participant behaviour.
  - Establishing data space rules.
  - Creating a governance structure.

- Technical Building Blocks
  These concern the technical elements facilitating data exchange, detailing how data spaces can be implemented on IT infrastructure.

Both categories are illustrated in Figure 2. It's important to note that not all building blocks are mandatory for a data space's operation, and various combinations can be tailored to specific needs.

[1] Here, secure execution refers to performing an analysis such that minimum information is exposed about the input data, for example through the use of MPC.
[2] Context at a minimum refers to a given group of parties that has agreed to work together, making data available for joint analysis. In addition, separate contexts may exists for different use cases, e.g. KYC or AML.

While these building blocks depict a comprehensive overview of the required functionality of data spaces, it does not directly translate to a set of technical components that can be deployed on an It infrastructure.

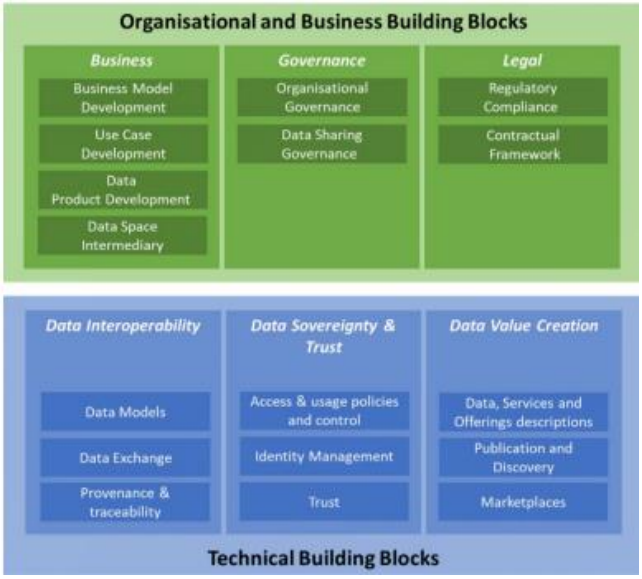

Figure 1: DSSC building block model for data spaces

.

The DSSC building blocks serve as guidelines rather than exact blueprints for technical components, meaning there isn't a straightforward one-to-one correlation between them. Their role is to outline the essentials for operating a functional data space. The DSSC blueprint does offer a mapping from these building blocks to specific technical elements for practical implementation. To clarify this relationship, a concise mapping here would be beneficial.

In a data space, two layers can generally be distinguished:

- Control plane: This layer is responsible for managing and orchestrating the process of data exchange. It includes various technical components and interactions that enable and facilitate the sharing of data between different participants. The control plane is typically highly standardized across various data spaces and includes all technical building blocks except those directly involved in the actual data exchange. The control plane covers all the technical building blocks from Figure 2, except Data Exchange.
- Data plane: In contrast to the control plane, the data plane is concerned with the execution of the data exchange itself. This layer deals with the actual movement, storage, and retrieval of data. It is specific to each data space and corresponds to the "Data Exchange" building block, focusing on the technical aspects of transmitting data.

Communication between participants usually happens through a component referred to as a "Connector" a connector usually implements most, if not all of the control plane logic, but might also contain data plane logic. It provides an endpoint through which participants can communicate with each other. The SIMPL initiative aims to facilitate implementations for the different technical components, as defined by the DSSC blueprint.

3 Orchestrating PETs in a Data Space

In the previous chapter, the use for combining PETs and Data Spaces was illustrated. But is the model as described in Section 2.2 applicable for a PET setting? In this chapter some limitations of the model with respect to PETs will be identified and an adjusted model will be presented.

3.1 Orchestrating data collaboration

For this report, previous work from the Netherlands AI Coalition's (NLAIC) working group Data Sharing is taken as a reference. They developed a model for data sharing using AI data spaces in their reference guide for intra AI data space interoperability [2]. The authors claim their model to be suitable also for specific subsets of PET interactions, providing a suitable starting point for the PET model proposed in this report.
This model contains a role model (Figure 3 in [2]) for AI data sharing, as well as a building block model that can be attributed to the role model. This reference guide is based on earlier work by the IDSA [1] on data spaces.

Since the role model was primarily designed around AI, it is interesting to see what limitations it has for application to PETs. The NLAIC model for AI data spaces introduced a building block model that can be attributed to the role model from Figure 2.2 .

3.1.1 Use AI business role model for PETs

The nature of PET interactions like those for MPC and FL, where data is no longer collected and processed by a single party, requires some modifications to the model described in Section 1.3. Note that PETs differ from AI algorithms in the sense that AI is used to learn new/other things from a given data set, whereas PETs are basically used to more securely implement an algorithm that might also be implemented without PETs (protecting the input, but ideally having little to no bearing on the outcome).

Three important limitations were identified for the current AI business role model for the application of PET analyses:

- In the current model the AI orchestrator provides the AI result to the initiator. However, when using PETs, not every party should have insight into the results. The orchestrator might not be amongst the parties that are allowed to get insight into the result.
- The original task of the Beneficiary is to both initiate and receive the result. Since the party initiating might not be the party that gets insights into the results, these roles should be split.
- A PET execution usually takes place at multiple parties that sometimes need to communicate encrypted data with each other in order to perform the computation.

The data space core roles for PET interactions are largely identical or very similar to those defined for intra AI data space operability as defined in [2].
A PET role model is presented in Figure 2. The descriptions and roles for PET empowered interactions can be found in Table 1. Note that these are slightly adjusted from the original model and roles. One role has been added. The PET Initiator will cover cases where an interaction takes place for someone else's benefit. Furthermore, the result is not being shared to the Beneficiary through the PET Orchestrator, but directly from the PET Operator.
Note that multiple parties can have the same role and also multiple roles can be fulfilled by one party.

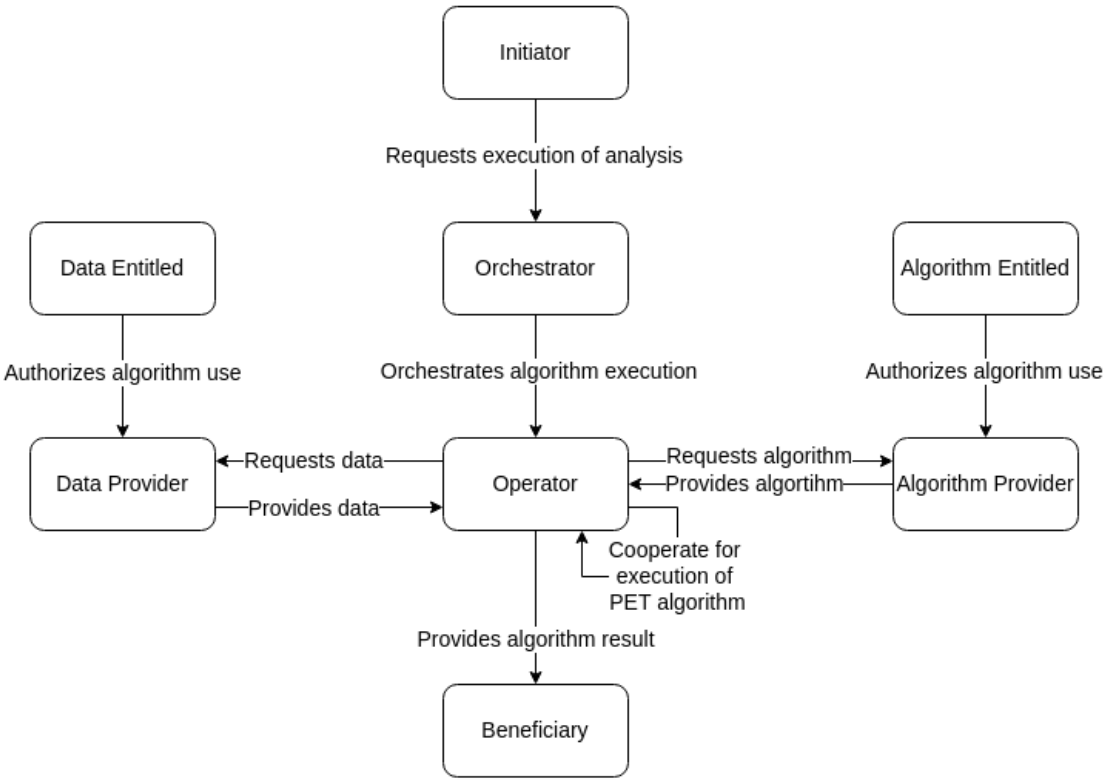

Figure 2: Data space core roles for PETs

Table 1. Data Space Core Roles for PETs.

Role	Definition
PET Initiator	The PET Initiator is responsible for initiating a PET interaction via the PET Orchestrator for the benefit of the PET Beneficiary. E.g. an analysis might be kicked off by a bank (based on red flags) for the benefit of a regulatory agency.
PET Beneficiary	The PET Beneficiary is interested in a result of the PET-supported interaction. The Beneficiary receives the results from the PET Operator. Note that this can be done in such a way that the PET Operator(s) do not learn the result themselves (e.g. by combining secret shared results at the Beneficiary, or by having homomorphically encrypted results be encrypted using the beneficiary's public key).
PET Orchestrator	The PET Orchestrator orchestrates the intended interaction and ensures that the PET-supported algorithm yields the intended results for the PET Beneficiary. The PET Orchestrator properly manages the policies for what it orchestrates. The PET Orchestrator understands what core modules for PETs are required and is tasked with bringing these together (i.e. orchestration), e.g. on identifying and bringing together relevant data and PET algorithms. The Orchestrator is also responsible for properly assessing policies that are relevant to the intended analysis result. A main added value of the Orchestrator is in being a single-point-of-contact for the PET Beneficiary in orchestrating and integrating the interactions with all core business roles and the services/building blocks they provide.
PET Operator	The PET Operator is responsible for providing an environment for execution of algorithms on the data. As such, it provides a capability (building block) that is referred to as the 'Application Container Environment (ACE)' in which the security gateway and the PET-supported algorithms are executed with the required data in order to produce the intended results of the algorithm. Moreover, the PET Operator is responsible for properly assessing policies that are relevant during the execution. For PETs like MPC and FL there is a minimum of two PET Operators that also interact with one another. (A PET Operator can but need not be a PET Beneficiary).
Data Entitled Party	Data Entitled Parties have one or more entitlements, e.g. having control over or being the subject of the data as provided by a Data Services Provider. The Data Entitled Party has the right to define the terms and conditions of use of data to which it is entitled.
Data Services Provider	Data Services Providers hold data in the data spaces and makes the data available in a controlled manner for PET Algorithms. The Data Services Provider manages policies for the data it is holding, e.g. it manages and enforces access and usage policies and provides additional policies to the PET Operator. The Data Services Provider also manages the quality and availability of data on behalf of Data Entitled Parties.
Algorithm Provider	Algorithm Providers hold the PET-supported algorithm in the data spaces. The Algorithm Provider properly manages policies for the algorithms it is holding. It manages and enforces access and usage policies and shares the policies with the PET Operator. The Algorithm Provider also manages the quality and availability of algorithms on behalf of Algorithm Entitled Parties.
Algorithm Entitled Party	Algorithm Entitled Parties have one or more entitlements to the PET-supported algorithm as provided by an Algorithm Provider. The Algorithm Entitled Party has the right to define terms and conditions of use of the algorithm to which it is entitled.

3.2 PET Infrastructure

This section discusses the translation from the role model to infrastructure capabilities that should be implemented to facilitate the interactions between the different roles in the model. This section also relates these capabilities to the different technical layers (control plane and data plane) of a data space.

3.2.1 Capabilities required by the business role model

To inform design decisions for the system layer of the PET infrastructure, a comprehensive listing of functionalities and capabilities is required for each of the business roles. The capabilities can be found in Table 2.

To support this categorization in the context of the execution of a PET algorithm, the concept of a PET workflow needs to be defined first. A PET workflow is the instantiation of an analysis for a specific purpose, involving specific participants, datasets, algorithms, execution platforms, beneficiaries and a well-defined sequence of steps.

Table 2. Capabilities for business roles in PET platform.

Role	Capabilities
Data Entitled Party	• Define metadata regarding dataset (details regarding content, access control and usage rights, default participation) e.g. with an associated Data Sharing Agreement • Publish or withdraw dataset to local or common catalogue • Negotiate contract with interacting participants
Algorithm Entitled Party	• Define metadata regarding algorithm (details regarding content, access control and usage rights, default participation) • Publish or withdraw algorithm to local or common catalogue • Negotiate contract with interacting participants
Data Provider	• Publish metadata regarding distribution of dataset • Provide access to dataset (push- or pull-based) • Enforce contracts
Algorithm Provider	• Publish metadata regarding distribution of algorithm • Provide access to algorithm (push- or pull-based) • Enforce contracts
Initiator (High-level orchestrator)	• Initiate PET workflow
Orchestrator (PET-level orchestrator)	• Verify and validate compatibility and platform support of participants for workflow • Verify and validate compatibility of datasets with PET workflow • Coordinate execution of each of the workflow steps with the corresponding participants
Operator	• Execute instructions provided by orchestrator on the execution platform • If client-side, send results of execution to defined destination • If server-side, accept results of execution from other participants
Beneficiary	• Obtain results of the execution of PET workflow

3.2.2 PET-level orchestration/governance components

The functionalities defined above can be best supported through the following technical components.

- Catalogue – A catalogue component is vital for the platform, as a trusted source of workflow metadata for all participants that may be interested or required to join a PET workflow. Additionally, this may extend to non-workflow metadata such as metadata regarding datasets being created by participants, PET execution capabilities and algorithms metadata, and template information for data sharing and usage agreements that may be mandated for both datasets and algorithms. This component addresses the catalogue capabilities required by the Data Entitled Party, the Algorithm Entitled Party, the Platform Provider and the Initiator.

- Workflow orchestrator – As there are two-levels of orchestration – at the higher workflow level, and the lower algorithm level – this component manages the higher-level state machine of a workflow (depending on the algorithm), publishes workflow metadata to a catalogue, and enables registrations and withdrawals for a published workflow.

- Identity and Trust mechanisms – While such PET workflows may be organized and coordinated across trusted and/or untrusted participants, secure communication and data sharing requirements must still be met, and these may be handled through standardized protocols for Identity and Trust. This is a generic technical aspect that may affect all roles.

3.2.3 PET-level components

At the PET algorithm-level, the required functionalities can differ widely across the possible supported algorithms. At the broadest level, these may be categorized as:

- PET-level orchestrator/coordinator – Depending on the requirements of the workflow, some PET algorithms may need a coordinator in order to facilitate execution and output across the participants. Such coordinators may also be responsible for the dissemination of workflow results to the beneficiaries. This component addresses the business role of the Orchestrator.

- PET-Participant Client/Executor – This component may combine the capabilities required by the Data Provider, the Algorithm Provider and the Operator of the execution platform. Instructions are received from the orchestrator/initiator and correspondingly are executed by each participant.

3.2.4 Compatibility with data spaces

The process followed to define the components begins with the core roles from the AI business role model for data spaces. Data spaces also define certain intermediary and governance roles for facilitating communication across the ecosystem of participants, and

these largely overlap with the governance-level requirements for a PET Platform, except for the Workflow Orchestrator. With the addition of this component, a Data Space may provide a modular and secure infrastructure for PET workflows.

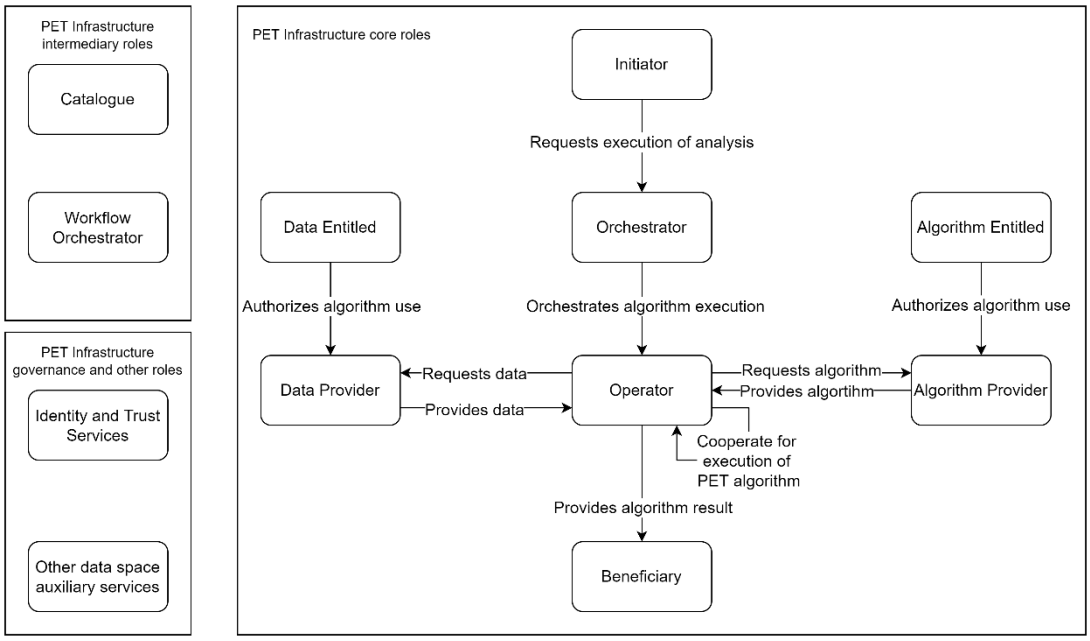

Figure 3: PET Infrastructure role model with data space roles

In the context of a data space, intermediary roles largely undertake the functionalities of the control plane, and the core roles are a combination of both control plane (through the initiator and orchestrator roles), and data plane (through all other roles involving a transfer of data, algorithm or result).

3.2.5 Workflow interactions

A generic non-normative example of how workflows may be defined, published, joined and executed in a PET infrastructure are described in the sequence diagram below.

- An initiator of a workflow must first define the metadata for a workflow, including the steps, communications, purpose, and a preferred format of datasets.
- The initiator publishes the information to a catalogue, along with an open call for participation, or a targeted request to a few specific data/algorithm providers.
- Potential participants for the workflow may discover the workflow from the catalogue service, and if interested, they may register a compatible dataset. For useable results, provenance of dataset creation and data quality checks may be described and requested in the workflow definition.
- Alternatively, potential participants might publish metadata about available datasets to a local/ecosystem catalogue, from where an initiator of a workflow may discover participants of interest.

- Once the initiator identifies that sufficient participants have registered, the workflow may be closed for participation. The workflow is then passed on to the workflow orchestrator component (in many cases, this might also be the initiator, and possibly also a data/algorithm provider in the process. The roles are separated for clarity.)
- The workflow orchestrator coordinates the distribution of operations and collection of results across multiple steps as defined for the specific PET algorithm to be executed.
- Once the algorithm is executed, results are revealed to the beneficiary/ies.

In addition, a standardized workflow state machine is necessary (which is outside the scope of this document). Such a state machine is to be updated by the Initiator and the workflow Orchestrator.

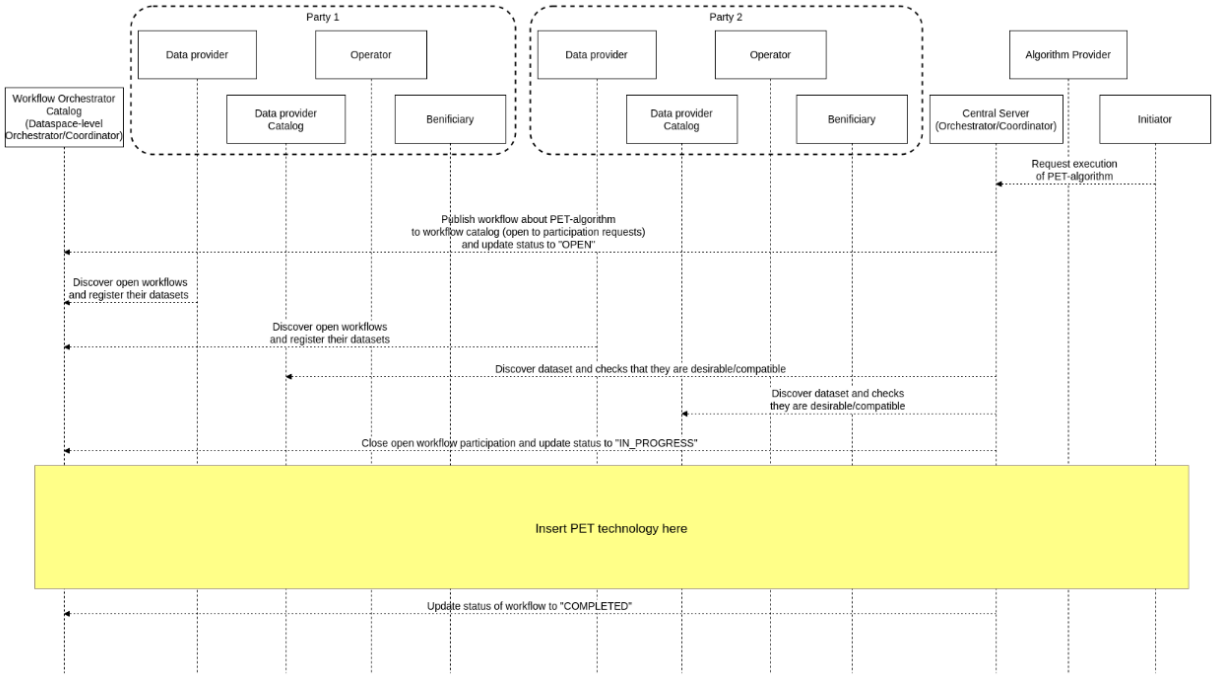

Figure 4: Sequence diagram of a workflow which can be used to run PET technologies in a data space

4 Conclusion & Research Questions

This report discusses the integration of Privacy-Enhancing Technologies (PETs) in Data Spaces. The AI role model from the NL AIC served as a foundation to introduce a PET-specific role model. These adaptations constituted a redefinition of role relationships and the introduction of a new role, the PET Initiator. Also some tasks and names of roles have changed.

We gave a view on the possible roles that different parties involved can take in a future PET infrastructure. In the future, these roles could in part be taken up by traditional Trusted Third Parties (TTPs). The insights in orchestration show that PETs are not necessarily a threat to current TTPs, but would ask of them to fulfill a different role. Note that the goal of a TTP is to perform an analysis. Traditionally this has meant requiring access to all data, but using PETs they can achieve this goal without getting insight into the data.

From the role model, capabilities have been defined that are necessary for an eventual implementation of a PET Platform embedded in a data space. A data space provides a solid basis for basic required PET functionality. A PET dataspace will take some work to set up initially, but in the long run it will make future collaborations easier.

An initial example of sequence diagrams have been proposed, showing how a PET interaction is to be initiated and distributed across the network of participants. Workflow orchestration should typically follow an “orchestration” phase where all the participants’ and datasets’ metadata are collected, before executing the actual algorithm. This has been elaborated using sequence diagrams as illustration.

This report is a first start in exploring the orchestration of PETs and Data Spaces in a real-world scenario. There is still much research needed on this topic. Some research questions that arise are:

- TTPs have normally been able to do model development on ‘plaintext’ data. Can the model development be done in a secure way as well, or do we need a ‘hybrid’ solution?
- How do the exact requirements of the parties involved map to the capabilities of the data space?
- A governance structure will be needed because too many analyses or queries can lead to a deducibility of data. What should this governance structure look like?

5 References

[1] International Data Spaces Association. (2019). REFERENCE ARCHITECTURE MODEL. https://internationaldataspaces.org/wp-content/uploads/IDS-RAM-3.0-2019.pdf
[2] International Data Spaces Association. (2023). REFERENCE GUIDE FOR INTRA AI DATA SPACE INTEROPERABILITY. https://nlaic.com/wp-content/uploads/2023/04/NL-AIC-intra-AI-Data-Space-Interoperability-v3.2.pdf

[3] Data Spaces Support Centre. (2023, September). Data Spaces Blueprint | Version 0.5. https://dssc.eu/space/BPE/179175433/Data+Spaces+Blueprint+%7C+Version+0.5+%7C+September+2023
[4] Lo, S. K., Lu, Q., Zhu, L., Paik, H.-Y., Xu, X., & Wang, C. (2022). Architectural patterns for the design of federated learning systems. *Journal of Systems and Software, 191*, 111357. https://doi.org/10.1016/j.jss.2022.111357.
[5] https://www.sciencedirect.com/science/article/pii/S0164121222000899
[6] OECD (2023), "Emerging privacy-enhancing technologies: Current regulatory and policy approaches", OECD Digital Economy Papers, No. 351, OECD Publishing, Paris, https://doi.org/10.1787/bf121be4-en.
[7] Leveraging the benefits of combining data spaces and privacy enhancing technologies -– Joint BDVA and CoE DSV White Paper - Leveraging the benefits_FIN_2.04 (coe-dsc.nl)
[8] https://digital-strategy.ec.europa.eu/en/policies/simpl

ICT, Strategy & Policy

www.tno.nl

TNO innovation for life