



PETs Inspiration Workshop for CID, YOUth

Agenda

	<i>Time</i>	<i>Speaker(s)</i>	<i>Remarks</i>
1. Welcome and introduction to CoE-DSC	15 min	Ruben	
2. Understanding CID YOUth	25 min	Yekaterina	
3. Introduction to Privacy Enhancing Technologies (PETs)	15 min	Yekaterina	
4. Practical examples of PETs in use	15 min	Yekaterina	
	<i>Break: 5 min</i>		
5. Relevance for CID, YOUth project and next steps	45 min	Ruben	Open discussion
	<i>Total: 120 min</i>		

Welcome

First, let's have a round of introductions



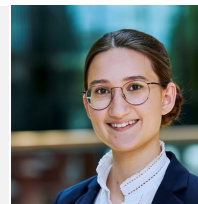
Facilitated by



Pepijn Groen
pepijn@datasharingcoalition.eu



Ruben van den Goorbergh
ruben@datasharingcoalition.eu



Yekaterina Travkina
yekaterina@datasharingcoalition.eu

What we do today – goals of the workshop

1



Introduce CoE-DSC: CoE-DSC supports the development of data spaces and data sharing infrastructure as well as stimulates growth Dutch data sharing community and initiatives

2



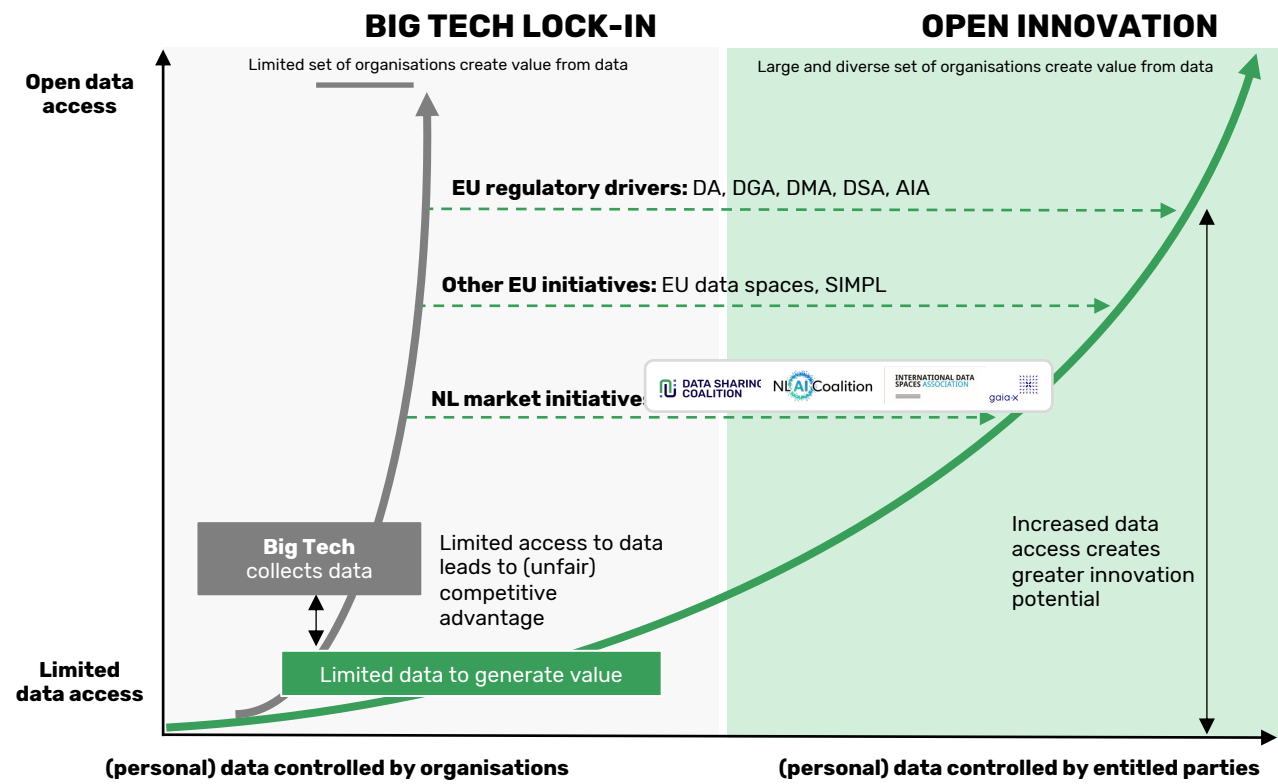
Show value of PETs in practice: Privacy-enhancing technologies (PETs) can help initiatives overcome privacy, commercial and reputational barriers by minimising the data used in analytics, while providing useful insights as seen in CoE-DSC use cases (e.g. elderly care and cancer research)

3



Define opportunities & next steps: PETs can be useful for the CID YOUth by enabling data sharing for research purposes while ensuring privacy sensitive data is not disclosed

Large-scale data sharing under control of entitled party stimulates innovation and value creation



Source: IMEC, Universiteit Gent, INNOPAY and TNO analysis

Dutch landscape of data spaces still in its infancy, but increasing funding and interest accelerates development

Dutch initiatives categorised per sector



Other Dutch initiatives



Key figures

Non-exhaustive and indicative

- About 50 NL data space initiatives are in development following Common Data Spaces as identified by the European Commission and more
- Dutch authorities and market parties released more than > EU 450 mln for the development of data spaces
- Dutch organisations contribute more than EU 96 mln € in-kind investments per year
- Currently about 500 Dutch organisations are connected to a 'live' data spaces (mainly through [HDN](#) and [SCSN](#))
- Only about 0,3% of Dutch organisations are currently involved in the development of Data Spaces
- In regards to Open Science development, the Netherlands follows the EU vision and aligns with [EOSC](#)
- National programme for Open Science [NPOS2030](#) involves 78 contributing institutions steered by Dutch universities, federation of university medical centers (NFU) and the Dutch Research Council (NWO)

CoE-DSC supports development of data spaces, infrastructure development and support Dutch data sharing community

CoE-DSC programme tracks

- 1 Data Spaces**



Support data sharing initiatives in the various stages of the development of a (cross-) sectoral use cases
- 2 Harmonisation**



Research and development to arrive at generic data sharing infrastructure and tools (e.g. PETs)
- 3 Community**



Expanding community with new participants, developing new partnerships and share achieved results

What value do we provide for initiatives

- ✓ **Ensuring interoperability** so that data can also be used across organisations, as well as between sectors
- ✓ **Maximum reuse of existing knowledge and solutions**, building on each other rather than reinventing the wheel each time
- ✓ **Insight in EU developments** and providing a channel to EU initiatives
- ✓ Making **scarce data-sharing expertise easily findable** and unlockable to market and initiatives
- ✓ **One central hub** for data sharing challenges

Founding partners CoE-DSC:



Pen-holders:



Collaboration partners:



CoE-DSC represents a large number of organisations that share data, consume data or facilitate data sharing

Over 500 participating organisations... Non-exhaustive ...represent different groups



- Industry associations** that represent their members
- Data sharing initiatives and software providers** that represent their end-users
- Standards institutions** that represent users of standards
- Companies** that create value with data themselves

Agenda

	<i>Time</i>	<i>Speaker(s)</i>	<i>Remarks</i>
1. Welcome and introduction to CoE-DSC	15 min	Ruben	
2. Understanding CID YOUTH	25 min	Yekaterina	
3. Introduction to Privacy Enhancing Technologies (PETs)	15 min	Yekaterina	
4. Practical examples of PETs in use	15 min	Yekaterina	
	<i>Break: 5 min</i>		
5. Relevance for CID, YOUTH project and next steps	45 min	Ruben	Open discussion
	<i>Total: 120 min</i>		

YOUTH Cohort studies collect various longitudinal data to learn about behavioural and psychological development of individuals

Summary description of CID YOUTH



Study Design:



- Ultrasounds
- EEG, MRI
- Eye-tracking
- Biological samples
- PCI movies
- Computer tasks
- Questionnaires

- YOUTH Cohort studies are part of Consortium on Individual Development (CID) work and are carried out to research neurocognitive development of children
- The research aims to investigate the behavioural, psychological, and social influence on development of individuals
- In YOUTH studies various longitudinal data is collected from children (from fetal stage to infancy and adolescent age) including ultrasounds, MRIs, results from behavioural & cognitive experiments
- YOUTH data adheres to [FAIR](#) (Findable, Accessible, Interoperable, Reusable) principles, and the project is a part of the Open Science movement with the goal to enable free dissemination of knowledge
- The Yoda environment is used to share data between researchers. Researchers need to register into Yoda and make data requests for their studies

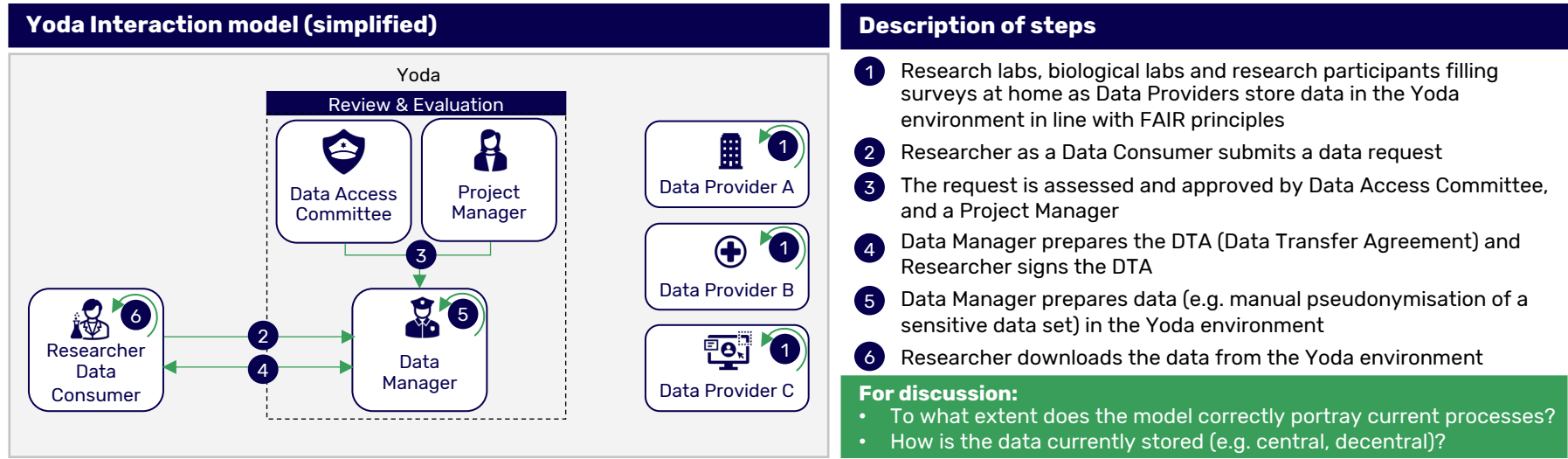
Parties involved in YOUTH studies:

- 2832 babies and 1338 children and their families
- Researchers from Utrecht University & other Dutch universities
- Practitioners & researchers at Utrecht Medical Centre (UMC) and Kinder Kennis Centrum (KKC)

For discussion:

- What are other involved parties besides the ones listed? E.g. are hospitals, schools, or any other social institutions involved?

Researchers request YOUth data stored in Yoda, requests are thoroughly assessed and if approved, data is prepared & sent



- Roles & responsibilities:**
- | | | |
|---|--|--|
| <p>Data Provider(s) in this case are:</p> <ul style="list-style-type: none"> • Research labs (UU, UMC, KKC) • Biological labs (KKC) • Research participants at home | <p>Data Consumer(s) - a Researcher who is registered in Yoda and needs data for their studies</p> | <p>Data Access Committee - a group of members responsible for review of the data requests</p> |
| <p>Data Manager - a party responsible for preparing data and DTAs (Data Transfer Agreements)</p> | <p>Project Manager - a party responsible for review of data requests and study registrations</p> | |

Legend: → Interactions

Current trust mechanisms to ensure privacy and reputation require significant time & resources, and leave risk for breaches

Trust aspects for data sharing



1 Privacy:

YOUth studies involve sensitive data of children, that falls under GDPR, and thus requires care when sharing between parties.



2 Reputation and integrity:

Ethical standards for research should be followed when sharing collected data, which requires care when handling data requests. (See Dutch [Code of Conduct for Research integrity](#)).

Selected trust mechanisms

- **Manual assessment of requests:** data requests from researchers are manually reviewed before acceptance
- **Manual dataset pseudonymisation:** data is manually pseudonymised for each request
- **Data transfer:** data is downloaded by researchers from the Yoda, as queries cannot be made to the Yoda itself
- **Consent:** procedures on family participation consent and possibility for withdrawal (see [here](#))
- **Identity assurance:** only registered and authorised researchers are allowed to download and use data
- **Legal agreements/contracting:** Prior to using data researchers are required to sign DTA (Data Transfer Agreements)

For discussion:

- To what extent are selected trust mechanisms correctly explained and exhaustive?
- To what extent are challenges correctly explained and exhaustive?
- What is the future concern for processes that are now handled well (e.g. any foreseen bottlenecks for current processes when the initiative scales)?

Challenges with trust mechanisms




- Manual processing of requests takes time and resources, making sharing data a costly process which is hard to scale
- Since researchers download the data, it increases privacy risks due to:
 - (1) Each new copy of a dataset increases risk of a data breach
 - (2) Downloading new data over time increases chances of potential re-identification (e.g. stacked MRI scans may reveal the face of a child)

Agenda

	<i>Time</i>	<i>Speaker(s)</i>	<i>Remarks</i>
1. Welcome and introduction to CoE-DSC	15 min	Ruben	
2. Understanding CID YOUth	25 min	Yekaterina	
3. Introduction to Privacy Enhancing Technologies (PETs)	15 min	Yekaterina	
4. Practical examples of PETs in use	15 min	Yekaterina	
	<i>Break: 5 min</i>		
5. Relevance for CID, YOUth project and next steps	45 min	Ruben	Open discussion
	<i>Total: 120 min</i>		

PETs have capabilities to lower privacy, commercial and reputational barriers for data collaboration participants

3 identified barriers hinder the development of data sharing collaborations

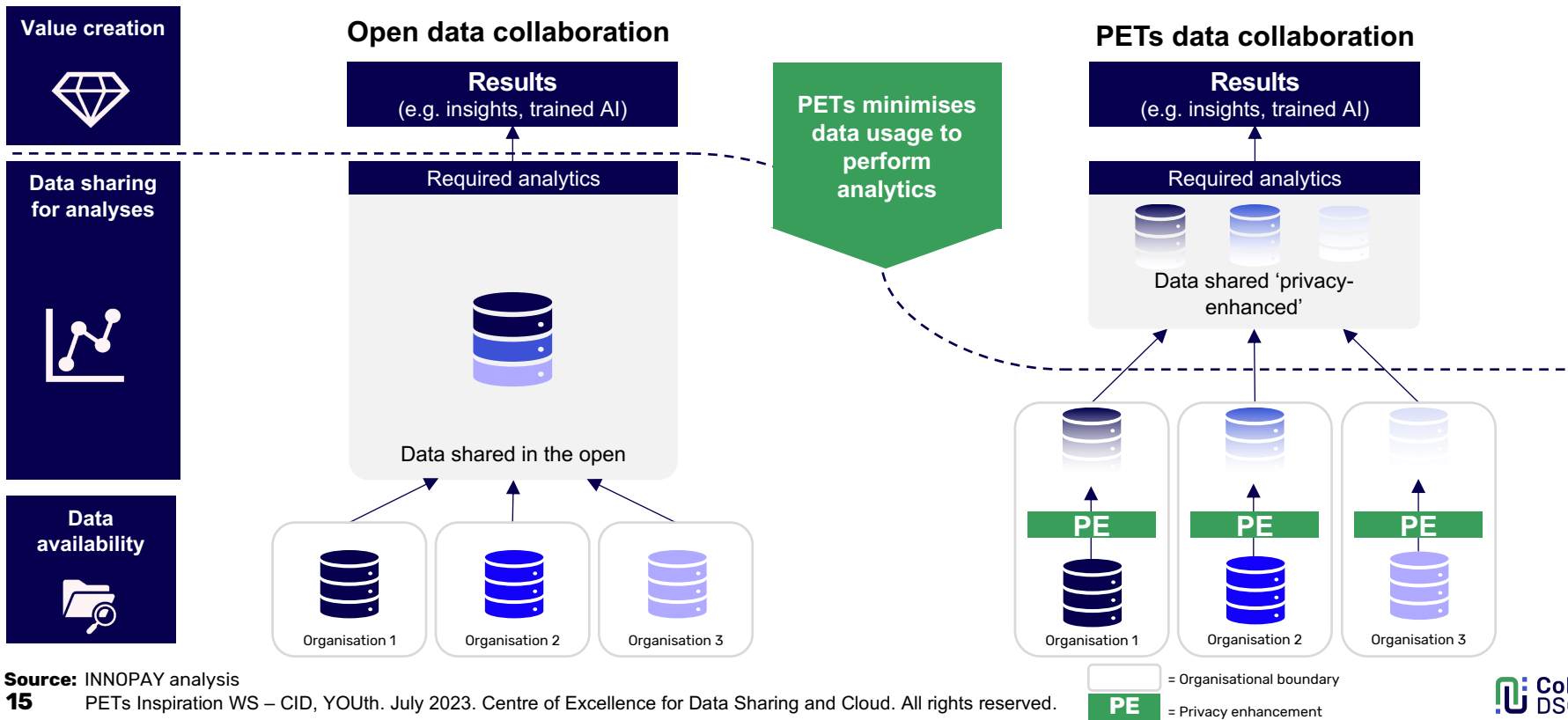
	1. Privacy Barrier	2. Commercial Barrier	3. Reputational Barrier
			
Description	Data that involves Personal Identifiable Information (PII) is difficult to share under GDPR, and thus compliance is required from organisations	Organisations are hesitant to share data because it constitutes commercial value and is considered a key asset	Organisations are hesitant to share data due to reputational risks which include damages resulting from the misuse of data
Example context	Typical for initiatives where sensitive data is involved. E.g., medical records of the patient, travel patterns of an individual etc.	Typical for initiatives where competitors are involved. E.g., transportation providers, healthcare providers, insurance providers etc.	Typical for initiatives where data is re-used for other purposes than for which it was originally collected (E.g., monitoring sector performance, sustainability metrics, etc.)

AI and PETs are new drivers to lower those barriers

- **Privacy Enhancing Technologies (PETs)** consist of various mechanisms that **allow for obtaining insights from analyses**, without revealing the source data
- As a result, participants in the data collaboration are ensured to have a **full control over their data** that should remain private

PETs ensure that sensitive data remains private during analytics while providing useful insights and results

Privacy Enhancing Technologies (PETs) consist of various mechanisms that allow **to compute insights** without revealing the source data.



Source: INNOPAY analysis

15 PETs Inspiration WS – CID, YOUth. July 2023. Centre of Excellence for Data Sharing and Cloud. All rights reserved.

PETs are part of a broader set of agreements to organise trust for data sharing

Triple A model visualisation

Applicability



Use case 1



Use case 2



Use case n

Accessibility

Accessibility of data is based on 9 building blocks

Data standards

Cost model

Metadata

Operational agreements

Exchange protocol

Security

Legal agreements

Governance

Identification,
Authentication,
Authorisation (IAA)

Availability



Data source 1



Data source 2



Data source n

- The use of PETs gives substance to and influences part of the building blocks to make data sharing possible
- PETs alone are never sufficient to organise data sharing, because for example they do not provide an answer to governance issues
- This is why PETs are always part of a broader set of mechanisms to organise trust for data sharing

Differential Privacy, Synthetic Data, MPC and Federated Learning are emerging PETs, each providing specific value to researchers

	Privacy assessment tech	Privacy enhancing tech		
	Differential Privacy	Synthetic Data	Multi-Party Computation	Federated Learning
Description	DP provides a measure on how much personal data is exposed by a specific data analysis algorithm. It shows mathematical parameters on privacy	Synthetic data tools transform datasets into new datasets with similar statistical properties, while removing privacy sensitive information from the original data set	MPC enables organisations to perform computations on data via encrypted decentral analysis so that no party learns anything beyond its own input and the output of the computations	Federated learning combines locally trained AI models into one improved model. The original data sets are not combined and are not shared, but stay at each data provider
Value	DP allows researchers to have a mathematical guarantee that an algorithm used does not reveal sensitive data	Synthetic data allows researchers to run particular analyses without using sensitive source data itself	MPC allows researchers to gain insights from datasets stored on different servers without revealing the data	FL allows researchers to send algorithms to the data stored in a server to gain insights without revealing underlying data
Visual				

Indicative

Legend: Raw data Data with minimised PII AI model Synthetic data generation Encryption Value Measure of privacy parameter

Source: INNOPAY analysis

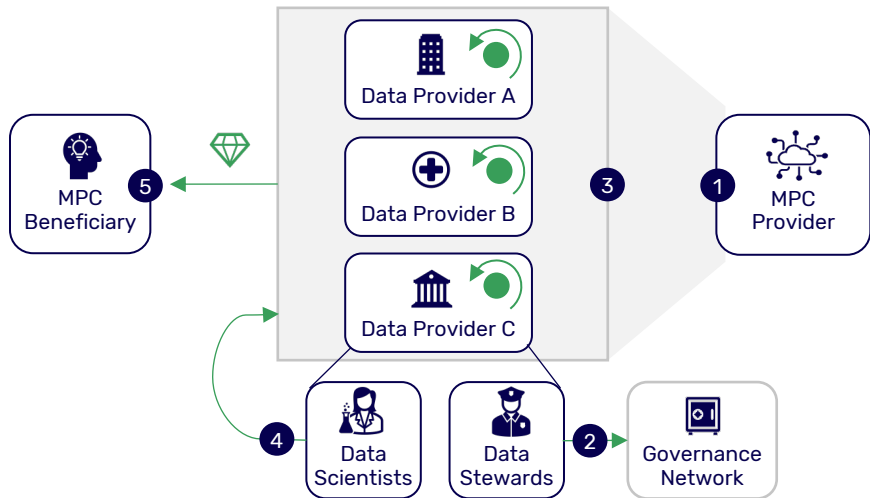
17 PETs Inspiration WS – CID, YOUth. July 2023. Centre of Excellence for Data Sharing and Cloud. All rights reserved.

Agenda

	<i>Time</i>	<i>Speaker(s)</i>	<i>Remarks</i>
1. Welcome and introduction to CoE-DSC	15 min	Ruben	
2. Understanding CID YOUth	25 min	Yekaterina	
3. Introduction to Privacy Enhancing Technologies (PETs)	15 min	Yekaterina	
4. Practical examples of PETs in use	15 min	Yekaterina	
	<i>Break: 5 min</i>		
5. Relevance for CID, YOUth project and next steps	45 min	Ruben	Open discussion
	<i>Total: 120 min</i>		

PETs for Elderly Care Monitoring (MPC ensures data privacy, while governance framework provides control over requests)

Description of the interaction model



Key results / learnings from the use case

- The Dutch elderly care sector can benefit from data collaboration to generate statistical insights and measure impact on policies (WMO, WLZ, ZVW)
- Most of data used is privacy-sensitive and therefore trust is difficult to achieve
- Multi-Party Computation (MPC) is selected as technology to organise trust for relatively low costs while safeguarding data privacy
- Linksight (MPC provider), DSW, Delft Municipality and Pieter van Foreest collaborate to generate statistics in Delft region, and plan to scale up
- Scaling up creates tensions between participants that want to have fast operations for gaining insights from data, and have strict control over the data. Such dynamics cannot be resolved by MPC alone and requires a governance framework
- CoE-DSC supported Linksight in developing governance framework with:
 - Baseline mechanisms per all types of requests (e.g., digital identity procedures, contracting, accessing insights)
 - Additional mechanisms depending on whether participants in the compute group want to (A) exercise direct control, (B) delegate control to a trusted party to maintain pace, or (C) have a tailor-made compromise for control and pace

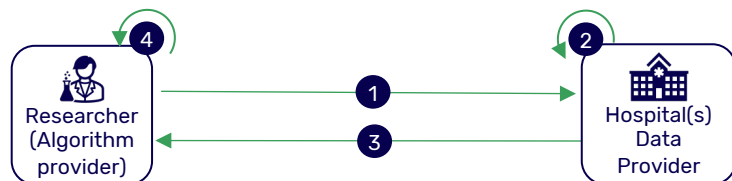
Participants



Source: CoE-DSC analysis, read the detailed report [here](#).

PETs for health cancer research (federated machine learning case under Personal Health Train for identifying cancer)

Description of the interaction model



Iterative loop of steps 1 - 4

Note: in this case, both the researcher and hospital are beneficiaries of the algorithm, since hospitals re-use AI model to identify cancer of their patients and researcher gains insights for their studies

Participants



Key results / learnings from the use case

- Collaboration of health providers and researchers in the Dutch health sector is vital for improving cancer recognition and treatment enhancement
- However, it is challenging to share data between health institutions and researchers, as it often involves sensitive patient data that which falls under GDPR
- In this case Federated Learning (FL) was deployed as a service to enable improvement of disease recognition, while ensuring sensitive data remains private. With FL researchers train AI locally on hospitals data to recognise lung cancer tumours in medical X-ray images
- Compared to traditional set up where data goes to the algorithm of the researcher, here algorithm travels to data (A2D), ensuring that sensitive data remains undisclosed (read more [here](#))
- The set-up ensures that data doesn't leave the security premises of the data provider, and also allows to securely match patients across data sets without unique identifiers (read more [here](#))
- The improved algorithm is re-used by both researcher and hospital, making it an iterative loop.
- To enable efficient data use for FL, participants agree on FAIR principles (Interoperability, Reusability Findability and Accessibility of data)

Source: NL AIC Analysis based on input from participants, [NL AIC Health data infrastructure report](#); For more on Personal Health Train (PHT) read [here](#)

20 PETs Inspiration WS – CID, YOUth. July 2023. Centre of Excellence for Data Sharing and Cloud. All rights reserved.

Agenda

	<i>Time</i>	<i>Speaker(s)</i>	<i>Remarks</i>
1. Welcome and introduction to CoE-DSC	15 min	TBD	
2. Understanding CID YOUth	25 min	TBD	
3. Introduction to Privacy Enhancing Technologies (PETs)	15 min	TBD	
4. Practical examples of PETs in use	15 min	TBD	
	<i>Break: 5 min</i>		
5. Relevance for CID, YOUth project and next steps	45 min		Open discussion
	<i>Total: 120 min</i>		

What we do today – goals of the workshop

1



Introduce CoE-DSC: CoE-DSC supports the development of data spaces and data sharing infrastructure as well as stimulates growth Dutch data sharing community and initiatives

2



Show value of PETs in practice: Privacy-enhancing technologies (PETs) can help initiatives overcome privacy, commercial and reputational barriers by minimising the data used in analytics, while providing useful insights as seen in CoE-DSC use cases (e.g. elderly care and cancer research)

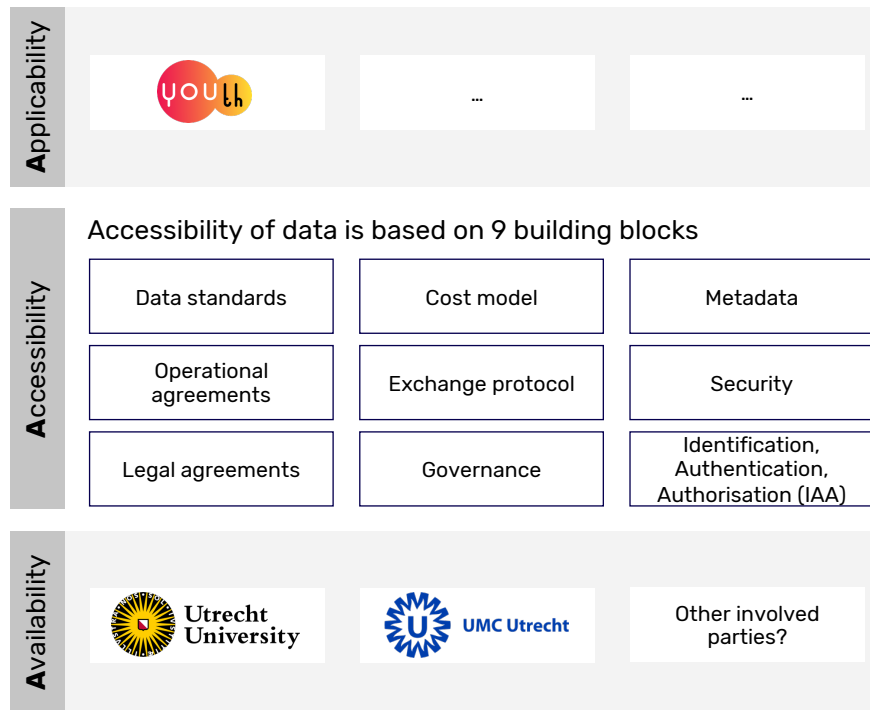
3



Define opportunities & next steps: PETs can be useful for the CID YOUth by enabling data sharing for research purposes while ensuring privacy sensitive data is not disclosed

Discussion aims to address: potential for PETs, requirements for the set-up, and broader aspects to enable YOUth data sharing

Triple A model visualisation



Preliminary points for discussion

Addressing how PETs can serve as trust mechanisms:

- Potential of using synthetic data, instead of original sensitive data
- Potential of Federated Learning where algorithm travels to the data, instead of sending data directly
- Potential of MPC as a tool for researchers to analyse data in a decentral way without revealing it

Addressing requirements for potential PETs set-up, e.g.:

- Ease of implementation, low set up costs, reduced time spent etc. (what are the other requirements?)
- To what extent commercial implementations are viable to set in a public institute?

Addressing a broader aspects to enable 'YOUth' data sharing beyond just PETs solutions:

- Out of 9 building blocks, what aspects are relevant for YOUth and can be further improved? (e.g. Security, IAA, Governance, etc.)

What's next – what can we do together?

 **Session**

 **Date**

 **Plan (Agenda)**

 **Goal**



Achieved

Inspiration workshop: PETs possibilities

6 July 2023

- Providing insight into the activities of CoE-DSC and previous PETs projects
- Possibilities of PETs for the CID YOUTH research

Share insight into the possibilities of PETs and how they contribute to the exchange of data



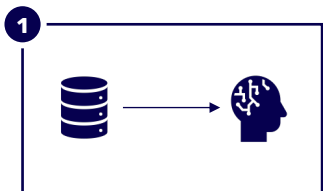
Potential Next Step:

Define relevant use cases for YOUTH to increase efficiency and enhance trust in research data sharing

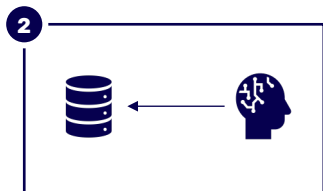
Appendix

There are 4 AI collaboration models each with their own rationale: D2A, A2D, TPP and NP

AI data spaces collaboration models Simplified view



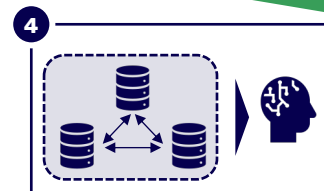
Data to Algorithm (D2A)



Algorithm to Data (A2D)



Third Party Processing



Network Processing

Description			
Rationale (relevance) per model			

From CoE-DSC use cases, the rationale and (dis)advantages were identified for the 4 AI collaboration model archetypes



Advantages and disadvantages per model

<ul style="list-style-type: none"> + Rich data available for AI beneficiary + Required infrastructure is mature - Risk of mis-use of data as data is processed outside control of DP 	<ul style="list-style-type: none"> + Keeps DP in control over data and executed algorithms on the data - Requires algorithm execution capabilities from the DP - Risk of mis-use of algorithm by DP 	<ul style="list-style-type: none"> + Enables data sharing relations with very little trust (only trust in a third party is required) - Not scalable due to involvement of the central third party or potential vendor lock-in 	<ul style="list-style-type: none"> + Keeps DPs in control over data and executed algorithms - High set-up costs due to lack of standardisation of infrastructure - Reduced computation power due to overhead from Network Processing
---	--	---	---