

# Trust in sharing sensitive data

*Abdulrahman Azab*

*Senior Advisor, NeIC & University of Oslo*



# NeIC

digital infrastructure for Nordic research excellence

200+  
experts

Since 2012

45 partners

7 M€

Nordic readiness for big opportunities



# Personal sensitive data



## Art. 9:

- Processing of personal data revealing **racial** or **ethnic** origin, **political opinions**, **religious** or **philosophical** beliefs, or **trade union membership**, and the processing of **genetic data**, **biometric data** for the purpose of uniquely identifying a natural person, data concerning **health** or data concerning a natural person's **sex life** or **sexual orientation** shall be prohibited.

# Data Controllers/Processors: Sensitive data services

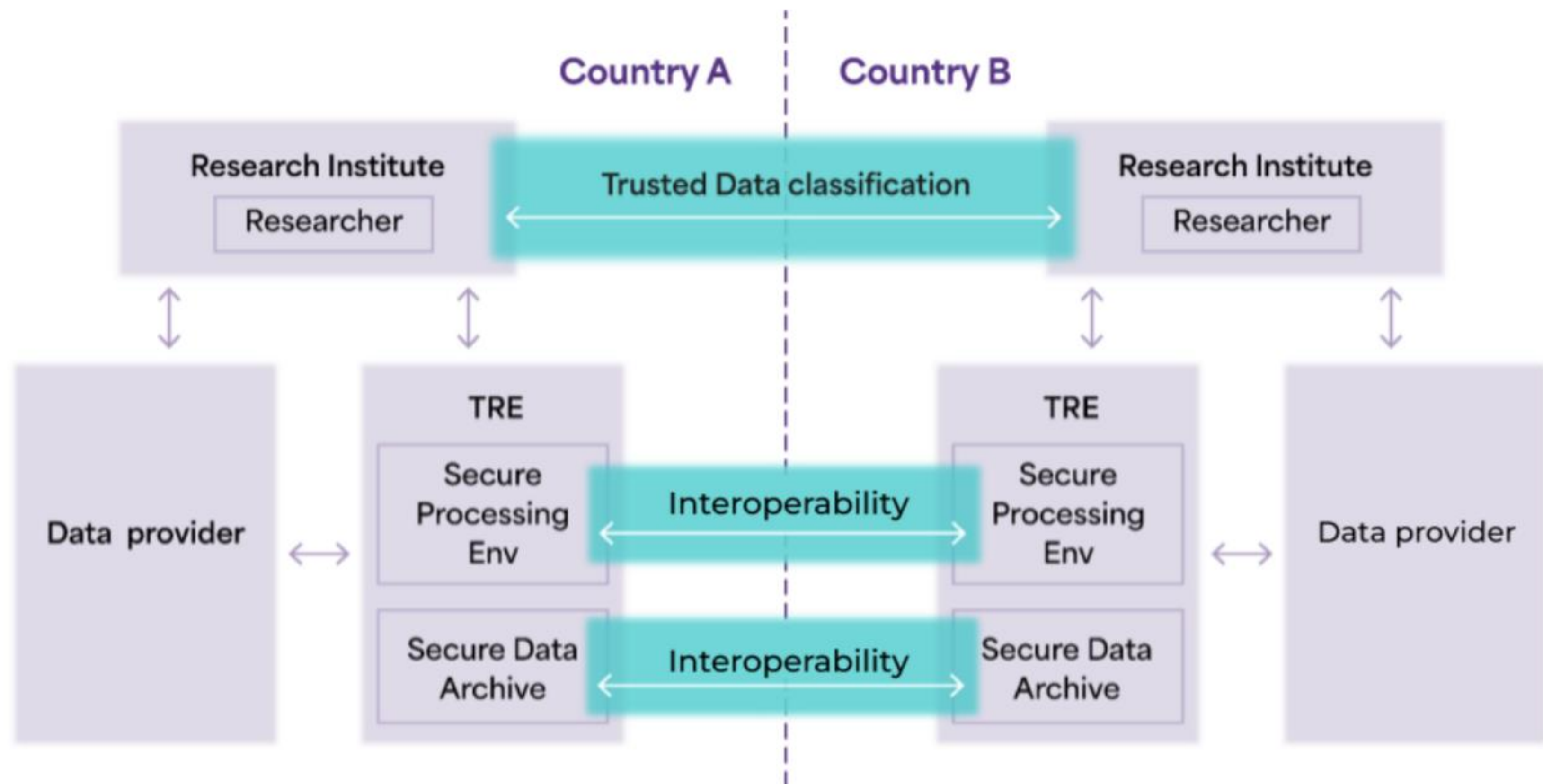
- **Trusted Research Environments (TRE)**
  - o Secure Data Archives (SDA)
  - o Secure Processing Environments (SPE)





# Trust

To allow for transnational federated analysis of sensitive data across multiple TREs, there must be interoperability between SPEs and SDAs.







UNIVERSITETET  
I OSLO



TSD

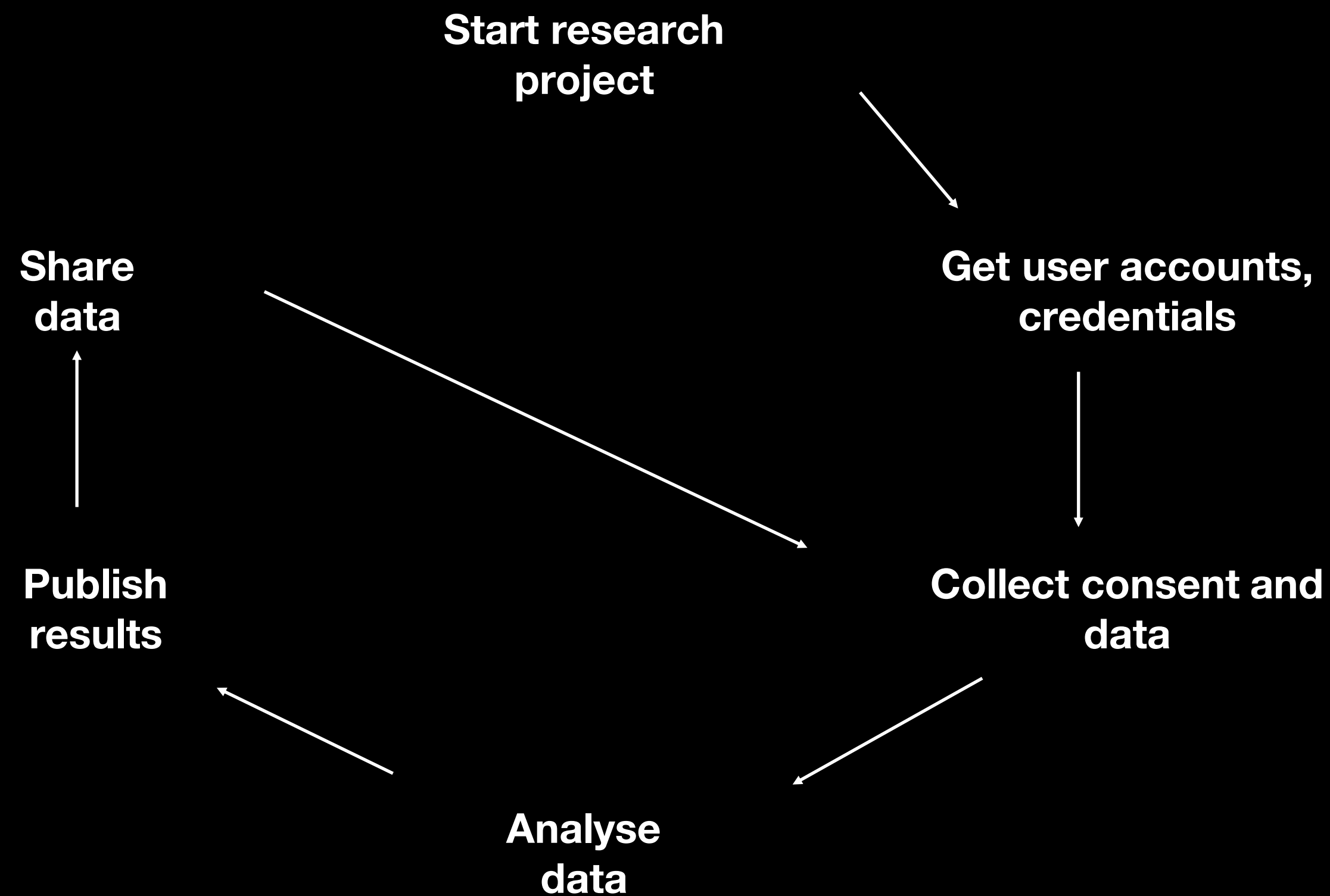
Tjenester for sensitive data  
Services for Sensitive Data

# TSD in a nutshell

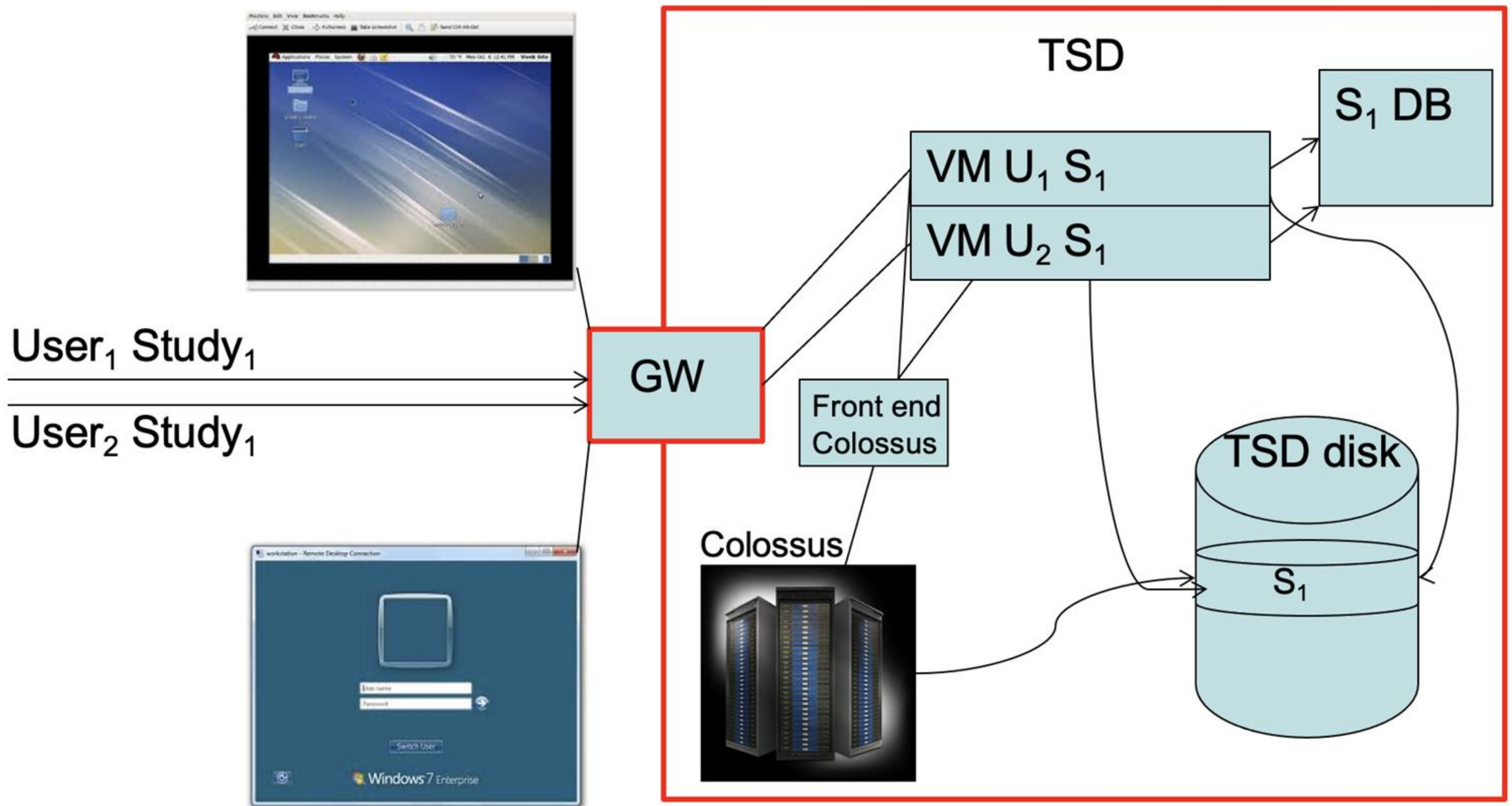
- In production since 2014
- Secure on-premises cloud hosted at University of Oslo
- PaaS, remote login with 2FA, own IdP
- APIs for app development
- VMs, HPC, containers, databases, backup
- Web services for the entire research life-cycle
- 8PiB sensitive data
- 8000 users, 1800 active projects, 80 institutions
- 1TB data moved in and out daily



# Project lifecycle



# Using TSD



# Services

**TSD Publication Portal**  
Download Files

**Publishing  
Sharing**

**Archiving**



Profile

Change your password / get new QR Code.

Log in

Apply for access to a project

Currently only available for users with a Norwegian electronic ID.

Log in (ID-Porten)

Project administration

- Review pending applications
- List all people in project
- Group management
- External import links

Log in

**Self  
service**

**Data capture  
Consent  
App development**



Welcome to the TSD Data Portal

To use this service you should already hold a valid TSD user account.

Import Files Export Files Record From Media Device

**Analysis  
Compute  
Software  
Storage  
Backup  
Database  
s**



# TSD API overview

TSD API: Secure, event-driven, multi-tenant HTTP API.

Typical TRE deployments include firewalls for network protection.

Relevance to GDPR: Article 32 stresses the need for security in data processing, which TREs embody through firewall protection.

# TSD API: Designed for GDPR Compliance

Tailored to be GDPR-compliant from the ground up.  
Focuses on transparency, security, and a user-centric approach.  
Ensures robust data protection in line with GDPR standards.

# API Design and Data Flow

Users → Clients (web services) → Application servers via proxies.

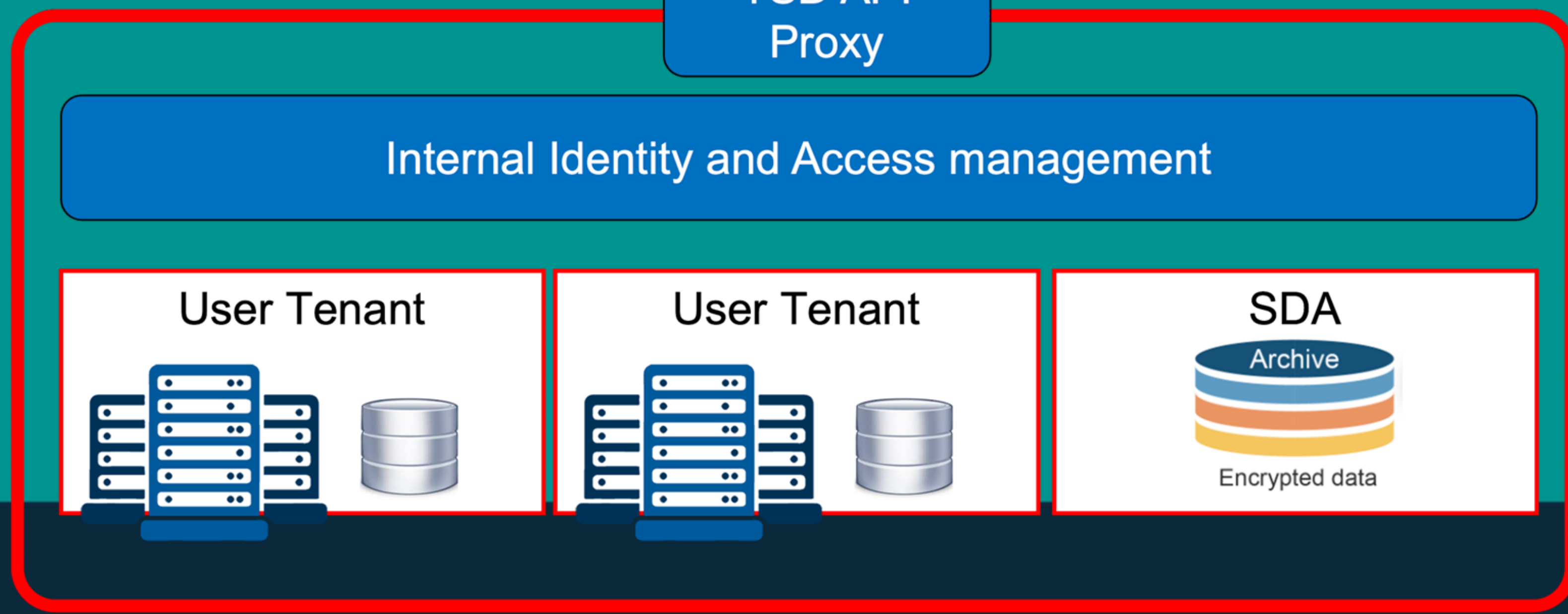
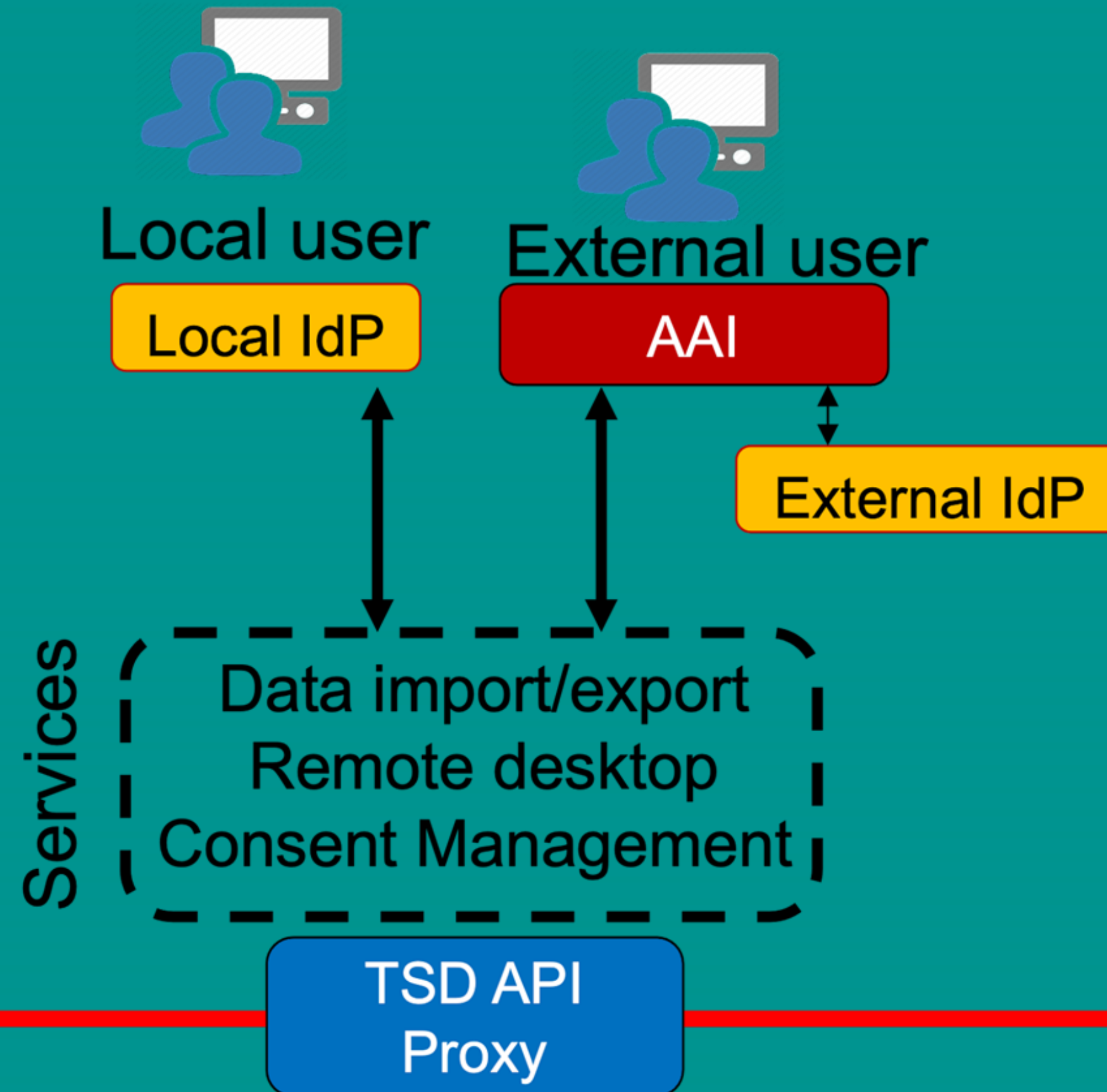
Two proxies: External (network bridge) and Internal (management & tenant subnets).

Standardized access control rules.

Efficient bug management.

Mandatory access control enforcement.





# Authentication Process

Use of OpenID Connect provider for user redirection and authentication.  
ID token issuance for authenticated users, followed by token exchange.

## GDPR Reference:

Article 5(1)(a) emphasizes lawful, fair, and transparent processing.

Article 5(1)(f) - Personal data processing must ensure data security, including protection against unauthorized processing.

# Authorization Mechanism

Authorization server checks access token, host name, HTTP method, and URI.

Routing through external and internal proxies.

ID token issuance for authenticated users, followed by token exchange.

Central authorization server evaluates access control grants.

App servers implement additional controls if necessary.

## GDPR Reference:

Article 25 - Data Protection by Design and Default. The authorization process ensures only those with proper rights can access data.

Article 32(1)(d) emphasizes regular testing of technical measures. The dual-check mechanism aligns with this requirement.



# Message Brokers and Consumer Applications

- Use of exchanges and queues for message processing.
- Replicating exchanges externally based on need.

## GDPR Relevance:

This structure upholds the principle of integrity and confidentiality as stated in Article 5(1)(f).

# Flexible Service Development Benefits

Central authorization combined with event-driven integrations.

Customers can create own clients and subscribe to tenant-specific queues.

GDPR:

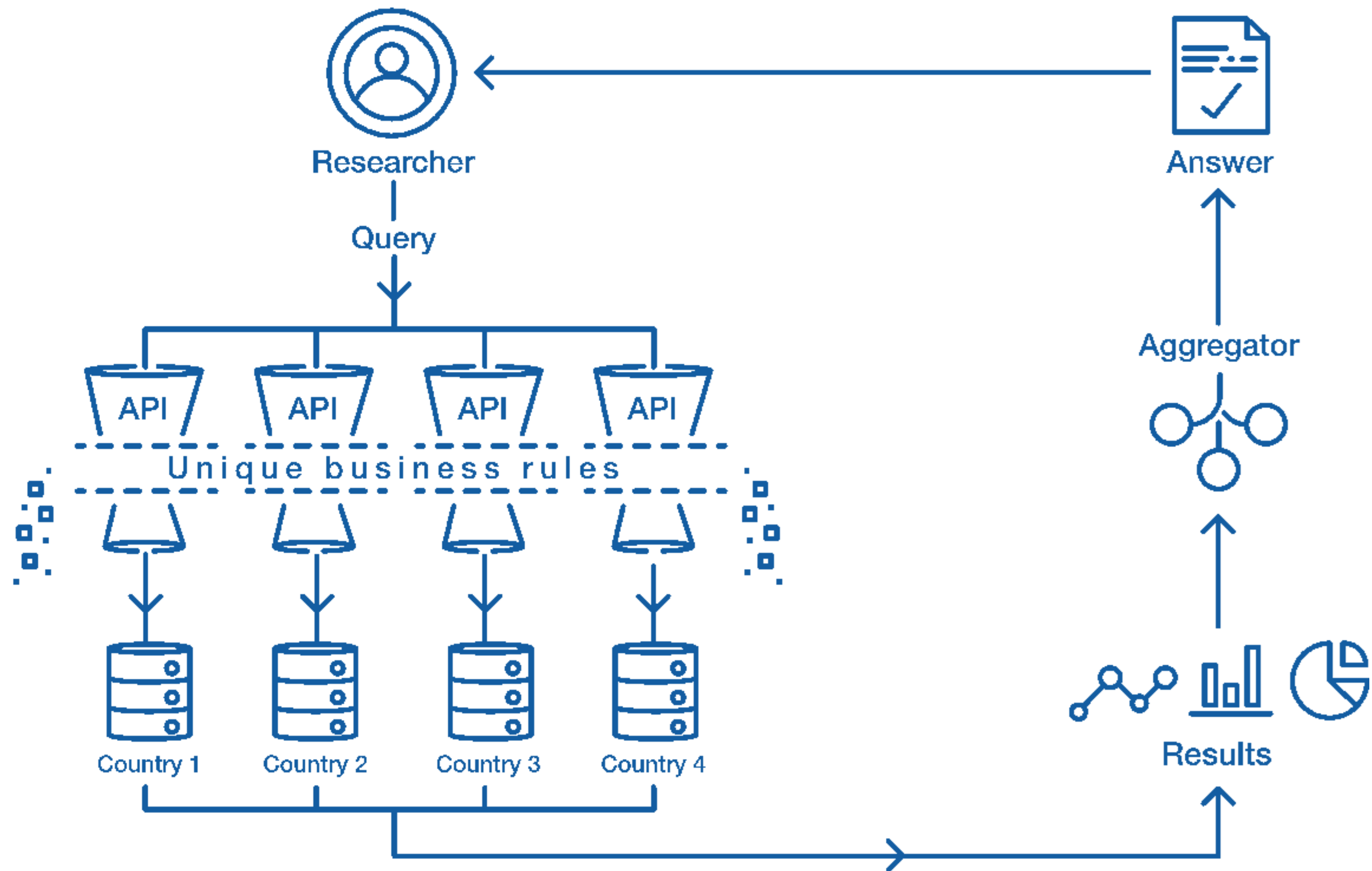
Article 24 - Responsibility of the controller to ensure and demonstrate compliance. The flexibility here allows TREs to adapt to changing data protection needs.

# References

- <https://gdpr-info.eu/>
- <https://github.com/unioslo/tsd-api-docs/blob/master/architecture/design.md>



## Federating data using APIs



Federated Data Systems: Balancing Innovation and Trust in the Use of Sensitive Data



SANE portal

Galaxy portal

TSD API

TSD API

TSD API

TSD API

Proxy

Interoperability

Proxy

Interoperability

Proxy

SPE

SDA

TRE1

SPE

SDA

TRE2

SPE

SDA

TREn

ST Repo



Secure Tools/workflows sharing

# Discussion and Q&A



# Event-Driven Integrations with RabbitMQ

Integration of app servers with internal RabbitMQ message broker.

Message dissemination: Metadata without service data, ensuring privacy.

GDPR Perspective: Event-driven processes with meta-data handling reflect the principle of data minimization as highlighted in Article 5(1)(c).





# **SANE – Secure ANalysis Environment in SURF Research Cloud**

Martin Brandt

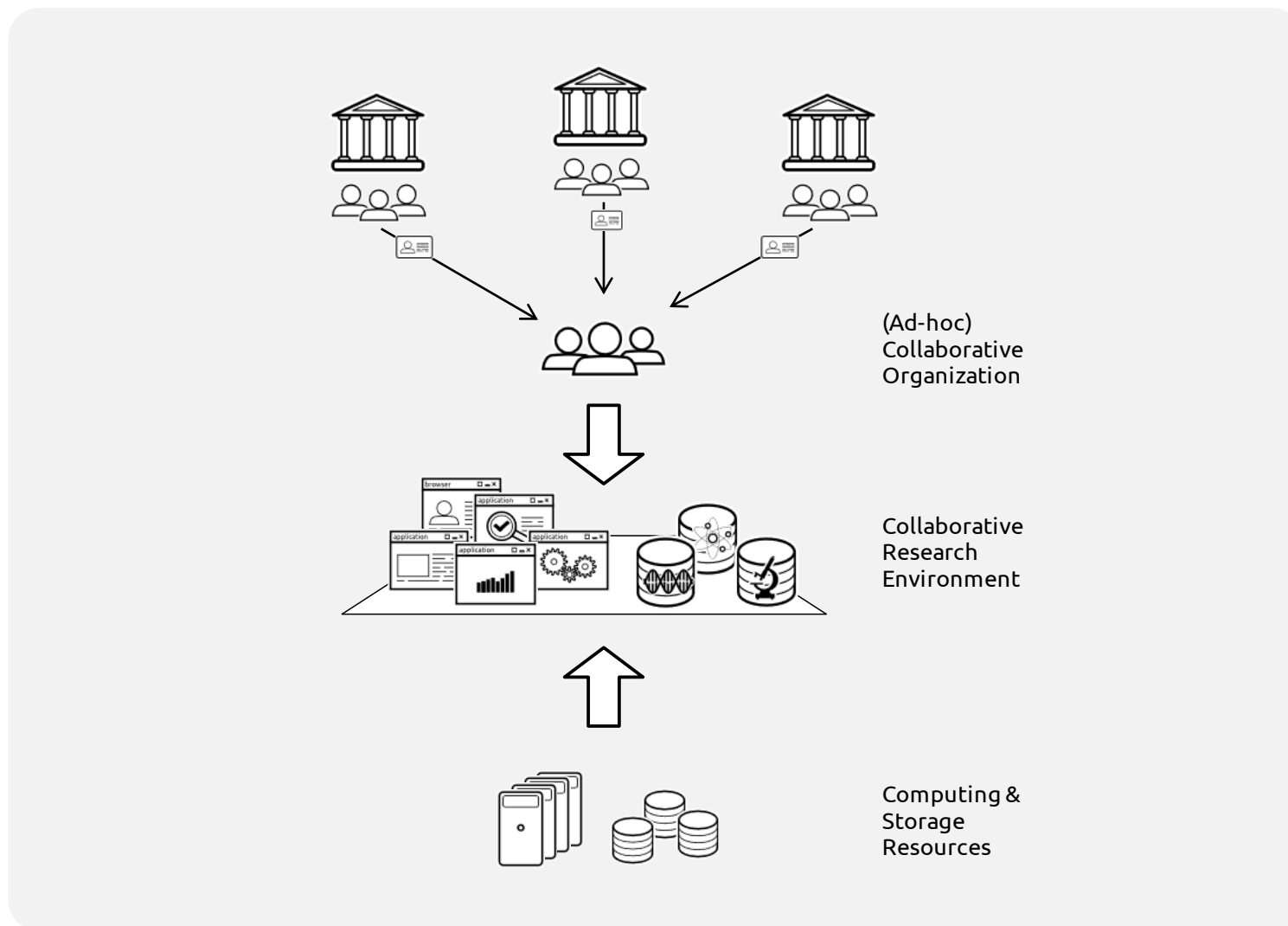
20 September 2023



# | SURF Research Cloud

SURF Research Cloud is a portal for building virtual research workspaces efficiently.

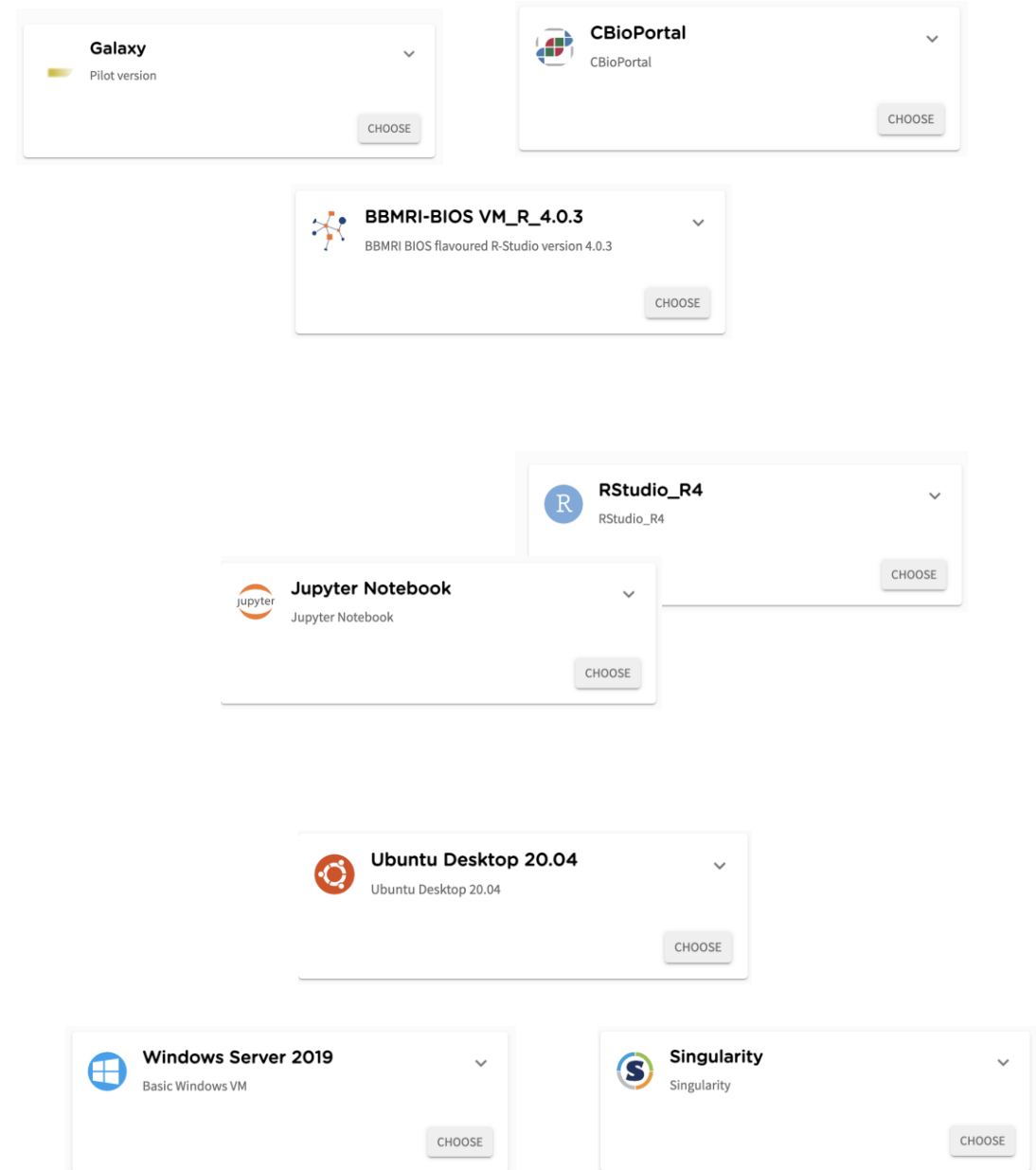
- Use preconfigured workspaces and datasets
- Or add your own. ...
- Connect to your data
- Start working, together



# | SURF Research Cloud

## Easily create reproducible research environments

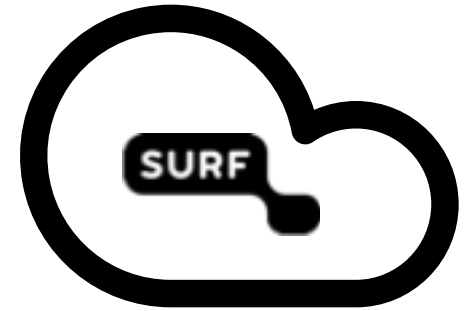
- Create workspaces in an intuitive user interface
- Short and long lived cloud environments
- Directly use linked contracts and budgets at SURF
- (Including NWO funded SURF Infrastructure grants!)
- Use catalog items as recipes for configuration
- The platform sets up cloud infrastructure



# | SURF Research Cloud

Use the best resources for your research

- Use cloud resources at SURF infrastructure, or commercial cloud with SURFcumulus
- Special compute resources like: GPUs, High memory VMs etc.
- Connect to datasets with ResearchDrive or iRods
- Use tools and applications offered by SURF and created by research communities and supporters
- Share your own tools and data in self-maintained [catalog items](#)



# | SURF Research Cloud

## Work together securely

- SURF Research Cloud is integrated with [SURF Research Access Management](#).

Authentication & Authorisation Infrastructure-as-a-Service which is based on the [The European AARC-project](#) architecture.

- Groups can be created based on users federated institute identity called Collaborative Organisations (CO)
- COs are used for access to catalog items, and all members of a CO will be users in workspaces started for that CO



# | Sensitive Data

## **Sensitive data remain unused**

“Although non-academic parties have an increasing number of interesting datasets available, there is currently no infrastructure available allowing researchers to analyse sensitive data in a way that data providers remain in control. As a result, most potential data providers are reluctant to share their datasets and so they remain unused (such as governments, heritage institutions or commercial parties like the Chamber of Commerce or Funda). Yet scientific breakthroughs would be possible if these datasets were available.”

<https://www.surf.nl/en/news/sane-secure-data-environment-for-social-sciences-and-humanities>

# | Sensitive Data

Sensitive data types:

- Privacy sensitive data
- Ethical use of data
- Copyright

Example use case:

“A language researcher wants to do computational linguistic research on a set of 1600 eBooks”

Or

“ A researcher wants to combine CBS microdata on neighbourhood income with medical data”

# | Approaches

- User access to data in a controlled environment (Tinker)
- No user access to data, only algorithm access (Blind)

Types of SURF Research Cloud use:

- SURF internal projects: SANE (Secure ANalysis Environment)
- SURF Research Cloud as resource provider: Fair Data Cube

# | SANE

“SANE is a virtual, fully shielded computing environment containing pre-approved analysis software (such as R and Jupyter notebooks) and access to the sensitive data. It allows the data provider to maintain complete control while still allowing the researcher to study the data in a convenient manner.”

<https://odissei-data.nl/en/2022/02/secure-analysis-environment-sane-secure-data-for-social-sciences-and-humanities/>

# | SANE

Project consortium:

SURF and CLARIAH, ODISSEI

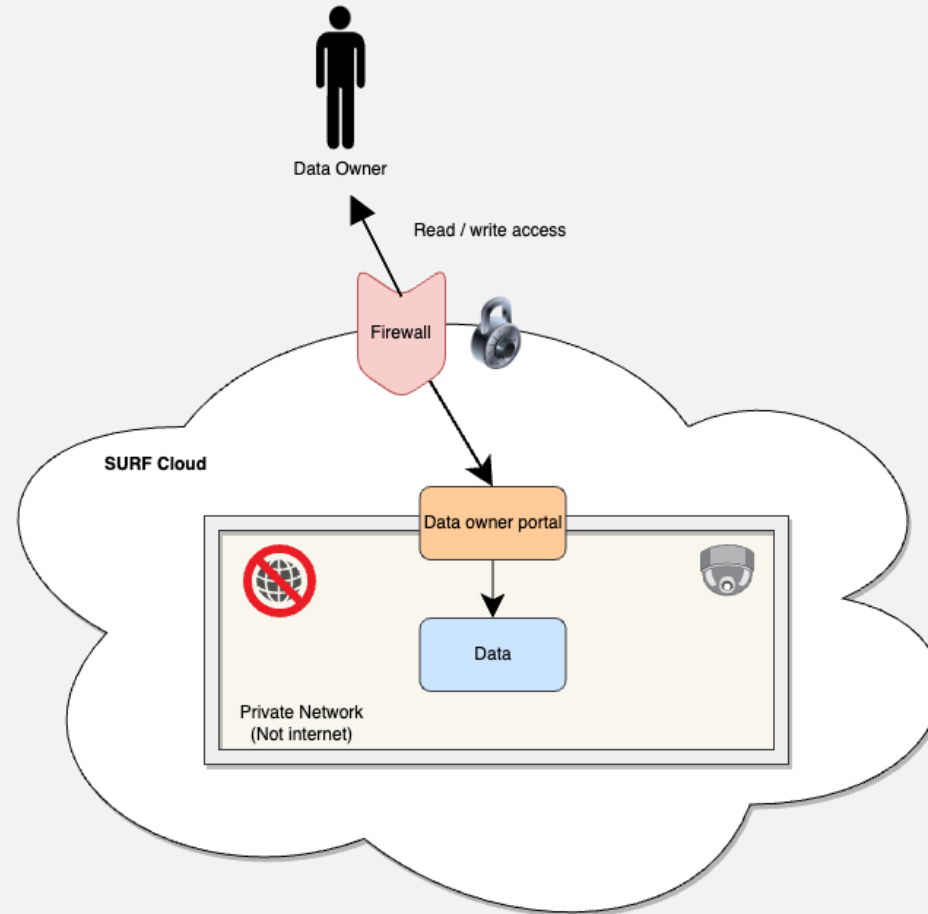
Currently a pilot with the prototype is active with:

- University of Utrecht
- Koninklijke Bibliotheek
- Instituut Beeld en Geluid
- Kamer van Koophandel

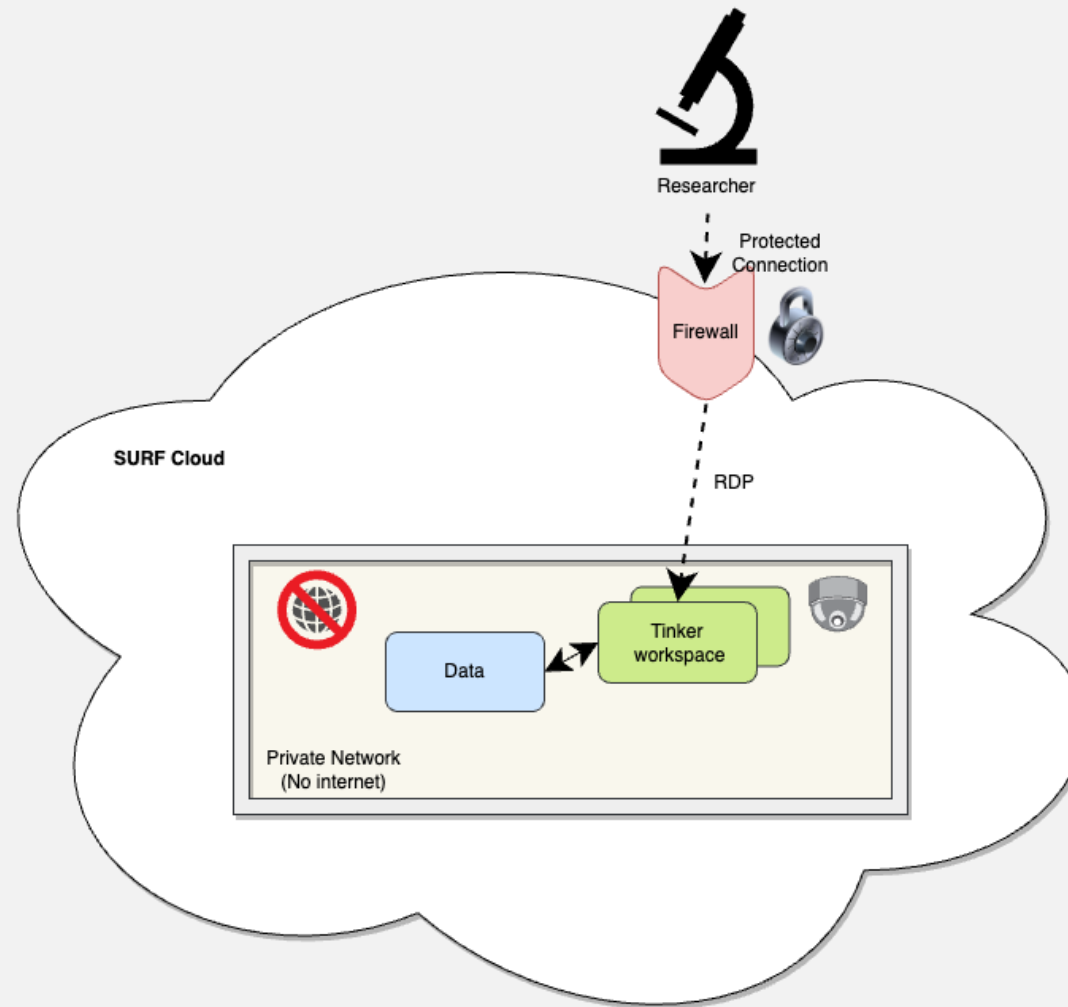




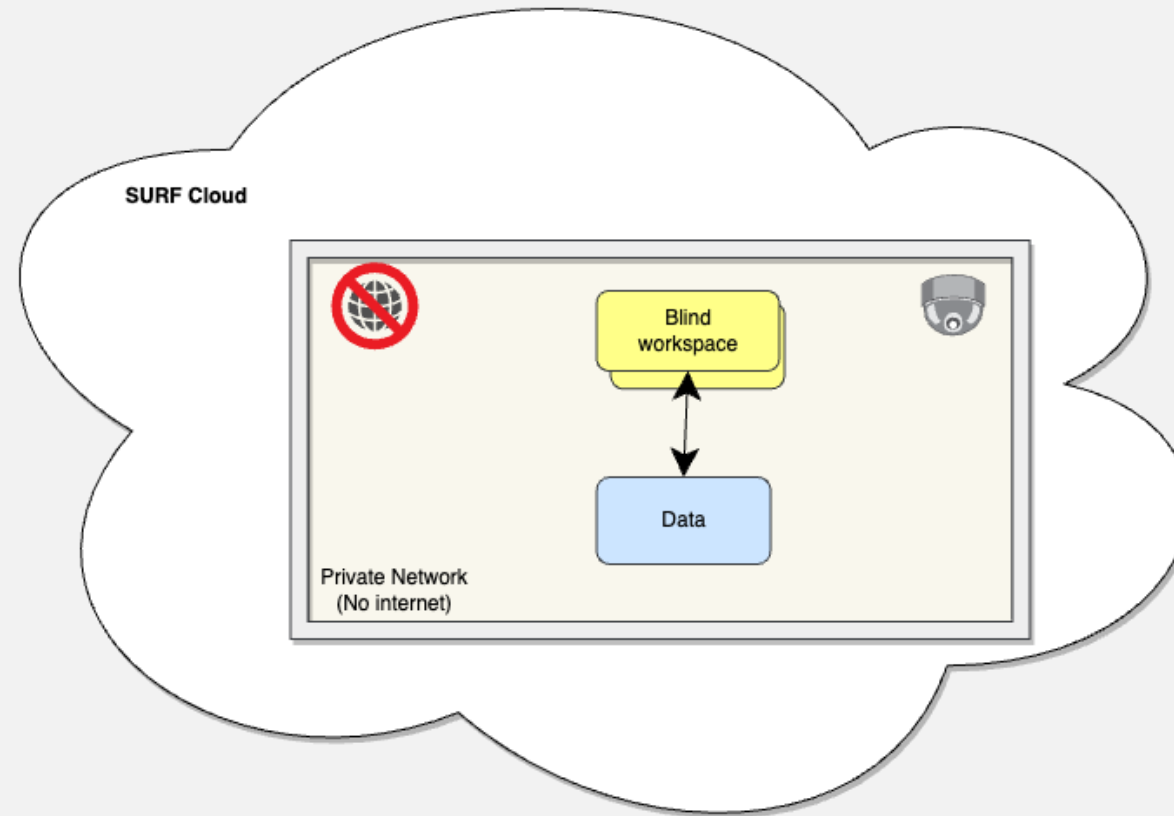
# | SANE - Data provider



# | SANE – Tinker



# | SANE – Blind



# | SANE – Data provider

Steps to create a new project for a dataset:

- Create or request Collaborative Organisation (CO)

Access is based on membership and sub group membership in collaborative organisation

- Create project elements in SURF Research cloud
- Data server
- Internal network
- Clone of template catalog items: Blind and / or Tinker
- Transfer project data to data server using data provider portal workspace
- Invite researchers into CO

# | SANE – Researcher

To use the SANE data a researcher can use the following steps:

- Request access to data
- Accept invite for Collaborative Organisation
- Request SURF Research Cloud budget
- Independent of SANE project
- Can be a small NOW grant, provided by SURF for example
- Start workspaces using catalog items provided by the project
- (Optional) Request extra software / data to be added to the project by the data provider
- Request results from data provider



## Create your workspace

 Restart workspace creation

✓ Collaborative organisation — 2 Catalog item — 3 Dataset(s) — 4 Cloud provider — 5 Options — 6 Name

### Choose the catalog item you want to use

These are the catalog items you can launch. Which catalog items are available to you depends on your collaborative organisation memberships. (see 'Profile')



#### SANE linux data owner portal

Only access for data owner



CHOOSE



#### SANE tinkler workspace

SANE tinkler workspace



CHOOSE



#### SANE Blind template

SANE Blind template



CHOOSE


BACK

CONTINUE

## Select wallet and cloud provider

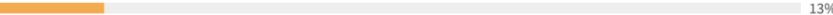
Please select your wallet and the cloud provider you prefer.

### 1. Choose one of the available wallets

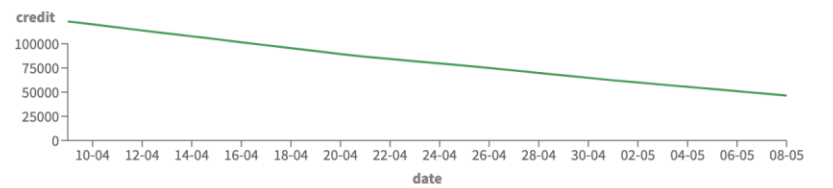

**Account for internal project for RSC opex & devops**
^

SELECTED ✓

**Budgets**  
 1 connected budget(s)

**Budget 1**  
 13%

Contract:	SRC OPEX during 2023	Description:	Workspace & storage usage
Valid from:	01-01-2023	Valid until:	31-12-2023
Total:	349000 (credit)	Used:	302550
Remaining:	46450		



### 2. Choose the cloud provider that best fits your needs

#### SURF HPC Cloud

SURF HPC Cloud


Ubuntu 20.04

1 core - 8 GB RAM

2 core - 16 GB RAM

4 core - 32 GB RAM

SANE blind - translators

 Due to limited capacity no 4 GPU workspaces are available, at the moment.

SELECTED ✓

### 3. Choose the flavour(s) of the cloud provider

Choose operating system:

Ubuntu 20.04

SELECTED ✓

Choose size:

1 core - 8 GB RAM



CHOOSE


2 core - 16 GB RAM



CHOOSE

# | SANE - Blind

## 1. Choose the expiration date of the machine

 13-05-2023 (20:18)

## 2. Workspace name, domain name and description

Name

Docker literary translators

Hostname

dockerliterary

Description

Description

## 3. Workspace parameters

Docker image name: (e.g. python:latest)

Docker image name

Dockerfile repository URL: (e.g. https://github.com/myorg/myrepo)

Dockerfile repository URL

## | SANE – Data provider

When a researcher requests results the data provider will start a data provider portal to analyze the researcher results and release them to the researcher if they contain no sensitive data.

# | SANE – conclusion

- Data provider in control
- Has control over user access and software used
- Use Federated Identity and Access Management for authorization
- Using the collaborative organisations from SURF Research Access Management
- Researcher is invited, but can use his own budget
- SURF Research Cloud is used as platform, building blocks used here can be used for other SURF Research Cloud catalog items and projects:
  - Restricted windows / Linux desktop access
  - Role based access to workspaces
  - Limited catalog item selection for Collaborative organisations

## | SANE – future

- Audit of the system
- Encrypted disks
- VPN access
- Controlled repositories for blind scripts and docker images
- Automate steps to setup project
- Use safe links to APIs instead of internal cloud data server



| **Questions?**

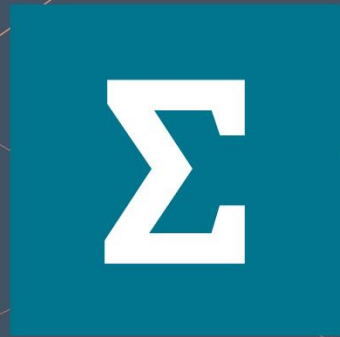


**Thank you for  
your attention!**

✉ Martin.brandt@surf.nl

<https://www.surf.nl/en/surf-research-cloud-collaboration-portal-for-research>

**SURF**



sigma2

National provider of e-infrastructure

# Archiving and Sharing Sensitive Data

EOSC Symposium, Madrid 20-22 Sept. 2023

# Research in the Digital Era



*Researchers need e-infrastructure for securely storing, analysing and reusing large volume of data, to enable **collaborative research** and be able to contribute to the **worldwide knowledge**. For the advancement of our society and business*



# Sensitive Data and the Collaborative Challenge

## What data? What sensitivity?

Not only personal sensitive data ...

... Also ...

Personal Data (video, audio, raw data, IoT/sensors etc)

.... And Also ...

Confidential data (business confidentiality)

## What Challenge?

Storing and Computing ... (Security, Isolation)

... Also ...

Sharing (API, Multitenancy, Access control)

... And Also ...

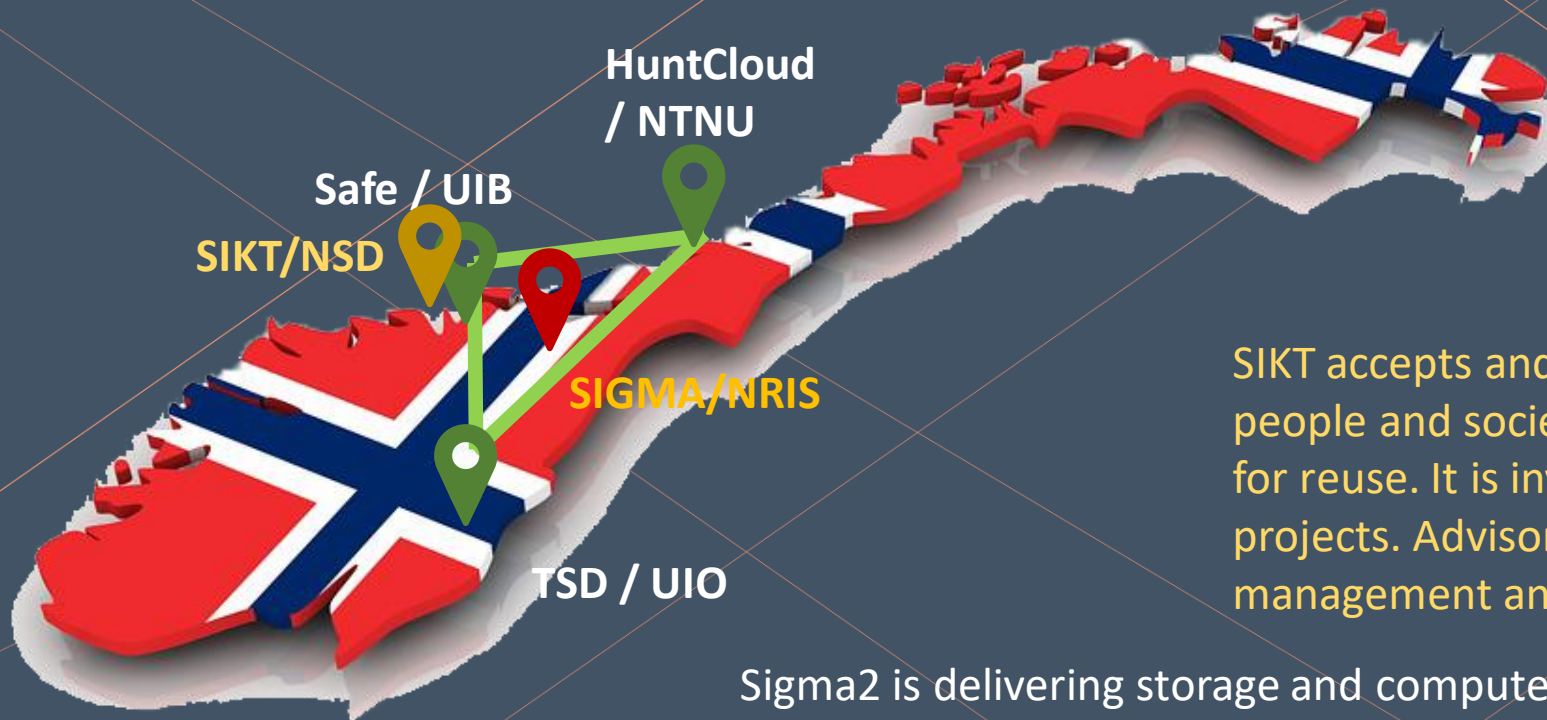
**Data archiving and reusing** (Identity vetting, data-owner / data processor agreement, consent management, etc.)

# Why do we need Research Data archives for sensitive (!) data?

We have already BioBanks, Healths Registries, Social Sciences Databases, Domain Specific Repositories (es. ELIXIR EGA)...

So Why? For what?

# What is happening in Norway?

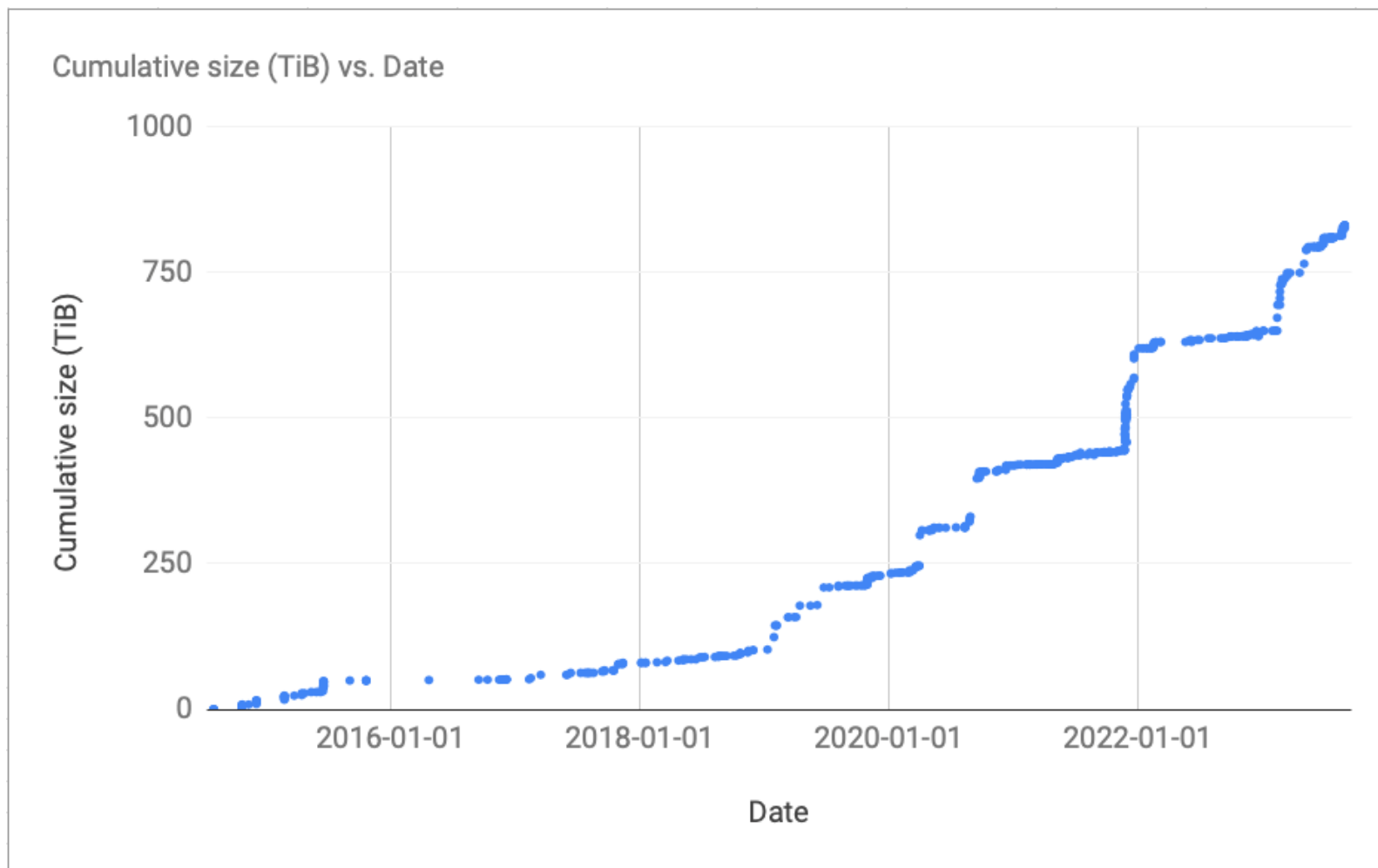


Through the ENTRUST PROJECT (INFRAEOSC-01-06) the three TREs are consolidating the already on-going work to create a unified APIs layer to support interoperability between the sites.

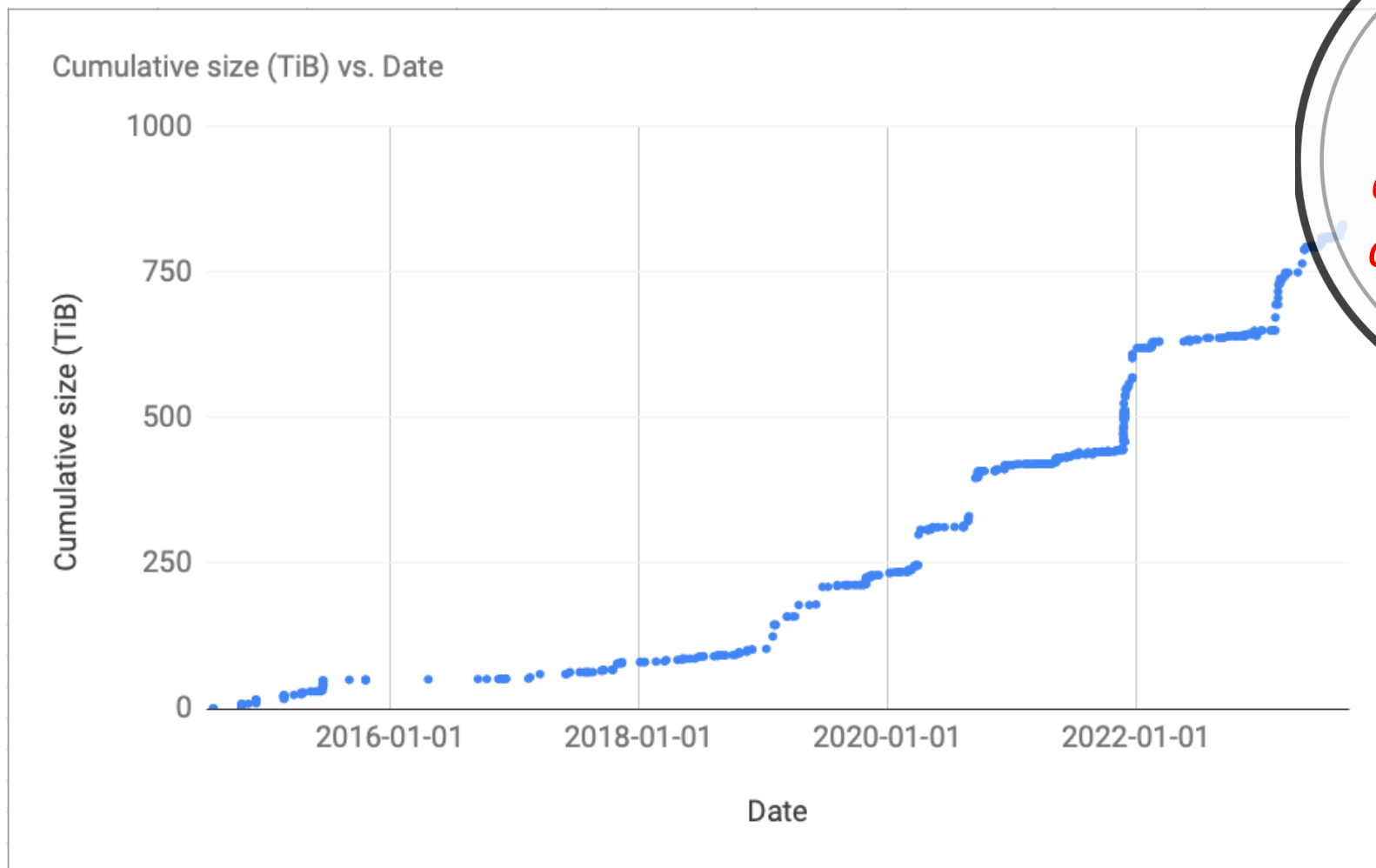
SIKT accepts and archives research data on people and society and makes the data accessible for reuse. It is involved in several European projects. Advisor for personal research data management and consent.

Sigma2 is delivering storage and compute service, including TSD. The largest archive for research data is operated by Sigma2 NRIS. As part of the EUDAT CDI, Sigma2 through the ENTRUST project will focus on sustainability models and procedures for networks of interoperable TREs. But also delivering archive storage and services for sensitive data (of different type)





Sigma2  
Research  
Data  
Archive



*Archive2021  
Project: use  
cases based also  
on sensitive data  
(\*)*

Sigma2  
Research  
Data  
Archive

# Is this possible at all? How much can be automatize?

