EOSC | ENTRUST

European Network of Trusted Research Environments

Deliverable D13.4

Year one version of EOSC-ENTRUST Blueprint & Interoperability Framework

Project Title (Grant agreement number)	EOSC-ENTRUST Grant Agreement 101131056		
Project Acronym (EC call)	EOSC-ENTRUST		
WP No & Title	WP13: Trusted research environment blueprint - Establishing a common terminology and approach		
WP Leaders	Miikka Kallberg (CSC), Pål Sætrom (NTNU)		
Deliverable Lead Beneficiary	2. CSC & 12. UiB		
Contractual delivery date	30/11/2024	**Actual delivery date**	02/12/2024
Delayed	Yes		
Partner(s) contributing to this deliverable	CSC, UiB, SURF		
Authors	Pål Sætrom (NTNU) https://orcid.org/0000-0001-8142-7441 Heikki Lehväslaiho (CSC) https://orcid.org/0000-0002-6263-1356 Christine Stansberg (UiB) https://orcid.org/0000-0002-3665-7843 Haneef Awan (UiO) https://orcid.org/0009-0001-0403-115X Ahmad Hesam (SURF)		
Contributors	Haakon Fannemel Breivik (UiB) Eirik Halseth (UiB) Ingeborg Winge (UiB) https://orcid.org/0000-0001-6139-5999 Miikka Kallberg (CSC) https://orcid.org/0000-0002-1457-3999 Stefanie Kirschenmann (CSC) https://orcid.org/0000-0002-2773-4798 Monica Abrudan (ELIXIR Hub)		

Funded by the European Union

Acknowledgements (not grant participants)	https://orcid.org/0000-0003-0228-4353
Reviewers	Rob Baxter (HDR UK/DARE UK) https://orcid.org/0000-0002-3693-8725 Jan-Willem Boiten (Lygature) https://orcid.org/0000-0003-0327-638X Abdulrahman Azab (SIGMA2) Anne van der Kant (HealthRI)

Table of contents

1. Executive Summary

EOSC-ENTRUST aims to create a European network of Trusted Research Environments (TREs) for sensitive data and drive European interoperability between TREs by development of a common blueprint for federated data access and analysis – the EOSC-ENTRUST Blueprint & Interoperability Framework (ENTRUST Blueprint, for short). The final ENTRUST Blueprint will consist of template legal agreements, architecture specifications, operating procedures, interface definitions, and a glossary of terminologies. This document is the first version of the ENTRUST Blueprint and presents a draft architecture specification and glossary.

All existing specifications and implementations of sensitive data processing environments are based on stand-alone design with no or very limited data interoperability. The inclusion of interoperability can lead to profound changes in the basic assumptions of these environment requirements. Therefore, we have drafted a general purpose vision of this interoperability based on preliminary versions of the DARE UK Federated Architecture Blueprint (DARE UK Blueprint, for short). The DARE UK Blueprint is based on the Five Safes framework and dataspace design thinking to allow research on sensitive data within a network of participating TREs. We have carefully mapped four existing TRE architectures and the first version of ENTRUST Driver use case requirements to the high level DARE UK Blueprint. Our analyses suggest that our architecture proposal should fit existing European TRE provider architectures and support all Driver requirements while maintaining the Five Safes principles for sensitive data research.

Going forward, we will collaborate with the Drivers to evaluate the current architecture proposal and with additional TRE providers to collect broader data on how provider TREs map to the ENTRUST Blueprint capabilities. We will open our proposed architecture for broader comments and suggestions for improvement through a dedicated version-controlled project.

2. Introduction

The EOSC-ENTRUST project aims to create a European network of Trusted Research Environments (TREs) for sensitive data and drive European interoperability between TREs by development of a common blueprint for federated data access and analysis – the EOSC-ENTRUST Blueprint & Interoperability Framework (ENTRUST Blueprint, for short). This document is EOSC-ENTRUST Deliverable D13.4 (D13.4 deliverable, for short) and is the first version of the ENTRUST Blueprint.

The following sections describe the D13.4 deliverable's contributions towards the project objectives, the methods used, the accomplished work, and results before discussing, concluding, and outlining next steps and the deliverable's impact.

3. Contribution towards project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives/key results:

	Key Result No and description	**Contributed**
Objective 1 Create a European network of Trusted Research Environments, linked to EOSC and EuroHPC, to enable transnational collaborative research on sensitive or restricted data.	1. A catalogue of suitable national or institutional TREs as part of the EOSC offering (WP10/11/12, WP13/14/15)	No
	2. A 'starter pack' of exemplar projects to demonstrate how networks of TREs can address European research priorities (WP4/5/6, WP7/8/9)	No
	3. European researchers are aware of capabilities through communication and outreach events (WP4/5/6) and materials delivered to support national TRE training programmes (WP4/5/6, WP13/14/15)	No
	4. Enable federated use via standards and technology for trusted researcher identity, data use and data access linked to developing European framework for trusted electronic identification of individuals (WP16/17/18)	No
	5. Enable researchers and software developers to deploy across multiple TREs via secure FAIR digital objects and workflows (WP16/17/18)	No

	6. EuroHPC capacity that meets the need for secure exascale and GPU (e.g., AI) computing can be identified and connected using the EOSC-ENTRUST framework (WP10/11/12, WP13/14/15).	Yes
Objective 2 Trusted Research Environment providers implement, validate, and promote their capabilities through a European framework using common standards and shared legal, operational and technical language.	1. An established European network of national and institutional TRE Providers (WP10/11/12)	No
	2. A service blueprint that allows technical interoperability between TRE based on the EOSC Interoperability framework (WP13/14/15)	Yes
	3. National and institutional TREs consistently set out their capabilities with common representation for validated legal, operational, semantics and technical aspects (WP10/11/12).	No
	4. Define the security baseline and auditing procedures for TREs to support the Five Safes[1] principles and capture requirements in guidelines for FAIR sensitive data in EOSC (WP13/14/15)	No
	5. Drive TRE composability via policy and process interoperability and set out an EOSC compliant governance model for a TRE services network (WP10/11/12, WP13/14/15).	No
Objective 3 National funders and governments understand the network of TRE capabilities serving their needs, and how TREs support their national priorities and their contributions to selected	1. A machine-readable catalogue of TRE capabilities allowing detailed, comparative analysis of technical capabilities and identification of gaps (WP10/11/12, WP13/14/15).	No
	2. Policy briefs on the capabilities of the European TRE Provider Forum (WP4/5/6, WP10/11/12) and Use Cases of their application in research domains of high societal impact (WP7/8/9).	No
	3. Connection between the EOSC-ENTRUST Provider Forum and the European Data Spaces (WP1/2/3, WP4/5/6).	No

[1] Ritchie, F. (2017, September). The "Five Safes": A framework for planning, designing and evaluating data access solutions. Paper presented at Data for Policy 2017, London, UK

transnational programmes		
Objective 4 The European Network of Trusted Research Environments (ENTRUST) is embedded in the European Open Science Cloud and the European Data Spaces and fosters an ecosystem of public, private and joint-venture providers of TRE services.	1. National and organisational providers are incorporated into EOSC via national members and the European network forms part of EOSC long-term strategy (WP1/2/3, WP4/5/6).	No
	2. The emerging European Data Spaces build their capabilities on the network of existing and developing TRE providers (WP4/5/6).	No
	3. Technological developments required by one Data Space activity can be directed to a forum of TRE specialists, reducing the need for duplication and coordinating investment in foundational technologies (WP10/11/12).	No
	4. A driver project to demonstrate opportunities for public-private partnerships (WP7/8/9)	No

4. Methods

4.1 Deliverable scope

The D13.4 deliverable is the first version of the EOSC-ENTRUST Blueprint & Interoperability Framework (ENTRUST Blueprint). Updated versions of the ENTRUST Blueprint will be released in 2025 and 2026 as outlined in the Draft Roadmap for EOSC-ENTRUST Blueprint.[2]

The complete EOSC-ENTRUST Blueprint and Interoperability Framework (ENTRUST Blueprint) will consist of template legal agreements, architecture specifications, operating procedures, interface definitions, and a glossary of terminologies. This first version of the ENTRUST Blueprint presents a draft architecture specification and glossary; template legal agreements, operating procedures, and interface definitions will be included in future versions of the ENTRUST Blueprint. We focus on the architecture and its necessary

[2] https://doi.org/10.5281/zenodo.12703951

components, as these will form the basis for interface definitions and operating procedures. Moreover, the architecture components may be used for structuring information about TRE solutions in the EOSC-ENTRUST TRE Provider Catalogue[3], the first version of which is published along with the Training package for the Year one ENTRUST Blueprint[4] and the ENTRUST Blueprint itself (this document).

4.1 Methodology

The Architecture work package (WP) (WP13) has held regular online project meetings every two weeks, supplemented with project meetings focused on the ENTRUST Blueprint. Both sets of meetings have been announced in the ENTRUST project calendar and have been open to all project members. Project meeting discussions and the initial use case (Drivers; WP7) requirements and TRE provider forum (Providers; WP10) capability mapping presented in the ENTRUST Blueprint Roadmap[5] identified the The DARE UK Federated Architecture Blueprint[6,7] ("DARE UK Blueprint") as a potential basis for the ENTRUST Blueprint. The DARE UK Blueprint offers an implementation and governance independent, abstract view of interoperability that should be applicable to beyond the specifics of the United Kingdom.

To test the usefulness of the DARE UK Blueprint ideas, capabilities of invited Providers and the initial combined set of Driver requirements[8] were mapped to the DARE UK Blueprint to test the compatibility and to identify potential gaps. Initial results from this mapping work were presented and discussed at the First EOSC-ENTRUST Evaluation & Adoption Workshop, held as a hybrid meeting in Helsinki on September 24-25, 2024. We have used an enterprise architecture methodology following the ArchiMate[9] 3.2 specification for architecture modelling and gap analyses.

We refer the readers to Section 6.5 for abbreviations, terms, and definitions relevant for the project.

[3] Miikka Kallberg, Rob Baxter, Stefanie Kirschenmann, Heikki Lehväslaiho, Deliverable D13.3 Machine-readable First Edition of the EOSC-ENTRUST TRE Provider Catalogue [in review].
[4] Christine Stansberg, Haneef Awan, Deliverable D13.2 Training package for EOSC-ENTRUST Year one Blueprint & Interoperability Framework [in review].
[5] https://doi.org/10.5281/zenodo.12703951
[6] https://dareuk.org.uk/wp-content/uploads/2023/07/DARE-UK-Federated-Architecture-Blueprint-Initial-Draft.pdf
[7] DARE UK. (2024). DARE UK Federated Architecture Blueprint (2.2). Zenodo. https://doi.org/10.5281/zenodo.14192786
[8] Anne van der Kant, Jan-Willem Boiten, Milestone report M7.1 Initial Driver Requirements for TREs from the four Drivers [unpublished].
[9] https://www.archimatetool.com/

5. Description of work accomplished

5.1 The DARE UK Federated Architecture Blueprint

The DARE UK Federated Architecture Blueprint[10,11] ("DARE UK Blueprint") describes an architecture for supporting secure research on sensitive data residing with different sensitive data providers. This includes services for linking sensitive data, services for secure data analyses including the necessary components for running federated analyses, and services for running the secure network itself. The architecture is designed based on the Five Safes framework to allow research on sensitive data within a network of participating TREs. Being technology-agnostic, the DARE UK Blueprint consists of (i) an infrastructure layer, which describes the federation participants and how data and information flows between these, (ii) a data layer, which uses the FAIR principles (Findable, Accessible, Interoperable, Reusable) to frame minimal requirements regarding metadata about the federation and about the data within the federation, and (iii) an organisational layer, which outlines requirements for a federation authority and discusses pros and cons of centralised vs distributed organisational models for a federation authority. The following sections describe the DARE UK Blueprint data space model, which is how the DARE UK Blueprint models sensitive data providers participating in the federation, and the essential components of the DARE UK Blueprint infrastructure layer.

5.1.1 The DARE UK Blueprint Data Space Model

Without loss of generality, the DARE UK Blueprint starts from the situation where individual-level sensitive data from a population are divided into administratively distinct regions, each containing several disjoint datasets about the individuals in the region and the aim is to find ways to enable population-scale data linkage and research in the public interest (Figure 1). Although simplified, this data space model aligns with International Data Spaces Association terminology[12] and captures the current and expected future state of the art in the UK, where different types of individual-level sensitive data, such as health, economic, educational, or environmental data, reside at different regional data providers. Note that this model directly applies to transnational data, such as those residing within the European Union. Based on this data space model, analyses of individual-level sensitive data fall into four data usage patterns (Figure 5.1.1) requiring (Q0) data from one dataset within a single region, (Q1) the same type of data from multiple regions, (Q2) different data types from individuals within the same region, or (Q3) multiple data types from multiple regions. Assuming that most of the future analyses of sensitive research data will be done within a trusted research environment (TRE), the DARE UK Blueprint infrastructure layer describes

[10] https://dareuk.org.uk/wp-content/uploads/2023/07/DARE-UK-Federated-Architecture-Blueprint-Initial-Draft.pdf
[11] DARE UK. (2024). DARE UK Federated Architecture Blueprint (2.2). Zenodo. https://doi.org/10.5281/zenodo.14192786
[12] https://github.com/International-Data-Spaces-Association/IDS-G/tree/main/Glossary

minimal capabilities necessary for supporting these four analysis patterns. Specifically, the DARE UK Blueprint describes capabilities needed to link individual-level datasets between different data providers (Q2, Q3) and run federated analyses across geographically distinct data providers (Q1, Q3).

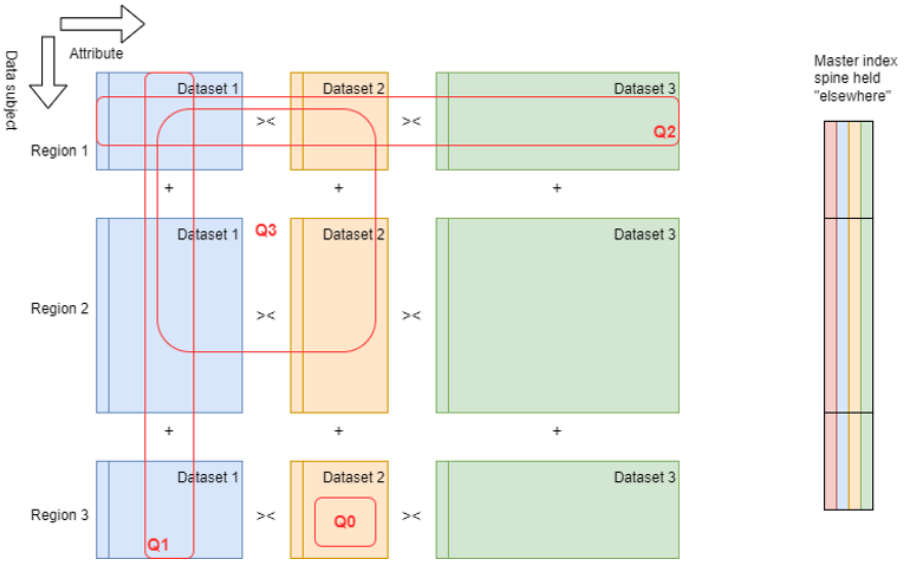

Figure 5.1.1 DARE UK Conceptual data space. Different types of individual-level sensitive data (columns; distinct colours represent distinct datasets) are split between different geographically distinct regions (rows; one row represents one individual). Analyses can use (Q0) a single dataset, (Q1) the same data from multiple regions, (Q2) different data types on the same individuals, or (Q3) multiple data types from multiple regions. Each dataset can have unique individual-level identifiers, with a separate master index providing the map (linkage spine) between dataset-specific identifiers and unique individuals. [Figure from DARE UK Federated Architecture Blueprint report from 2023[13].]

5.1.2 The DARE UK Blueprint Infrastructure Layer

The DARE UK Blueprint 2.0 infrastructure layer describes the services forming the federation and the individuals using the federation, referred to as Participants and Actors, respectively. Focusing only on Participants, the infrastructure consists of one set of Federation Services, TREs, and one or more of Software Services, Index Services, Discovery Services, and Job Submission Services that are connected through common Security Server interface (Figure 5.1.2). The DARE UK Blueprint makes no assumptions nor sets any requirements regarding specific TRE implementations; instead, it assumes that participating TREs follow a commonly agreed minimal standard, such as that defined in SATRE[14] or the ISO27001 standard for information security management systems[15]. Importantly, the architecture divides TRE into three functionally defined capabilities: Research Analytics Zones (RAZs), which give researchers access to sensitive data, Secure

[13] DARE-UK-Federated-Architecture-Blueprint-Initial-Draft.pdf
[14] https://satre-specification.readthedocs.io/en/stable/
[15] https://www.iso.org/standard/27001

Data Zones (SDZs), which support data management functions such as dataset linkage, and Query Management Zones (QMZs), which provide components needed for TREs to support services such as dataset discovery or federated analyses. A TRE may contain one or more of these Zones.

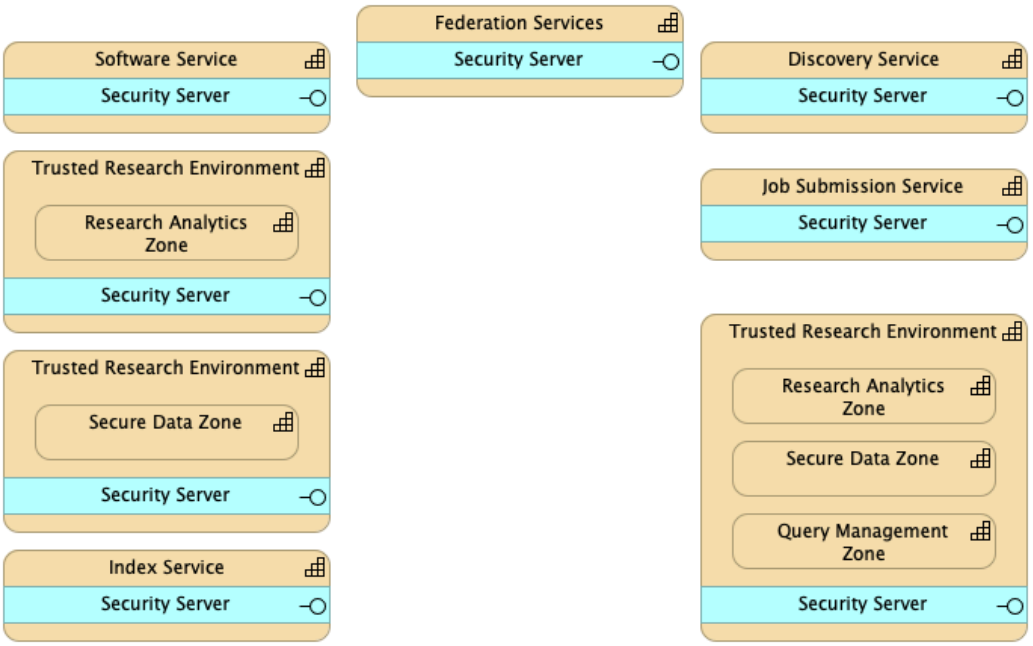

Figure 5.1.2 Participants in the DARE UK Blueprint infrastructure layer are connected through a common Security Server application interface. Participants in the DARE UK Blueprint are modelled as strategic capabilities and include one set of Federation Services, Trusted Research Environments (TREs), and one or more of Software Services, Index Services, Discovery Services, and Job Submission Services. TREs have at least one of three capabilities: Research Analytics Zones (RAZs), Secure Data Zones (SDZs), and Query Management Zones (QMZs). We illustrate three possible TRE configurations: a RAZ-only TRE, a SDZ-only TRE, and a TRE with all three zones.

Although the infrastructure layer also describes data and information flow between participant and actor roles, we will focus on Participants and their minimal required functionality as defined in the DARE UK Blueprint and describe these in the context of the actors and their roles in the infrastructure (Figure 5.1.3).

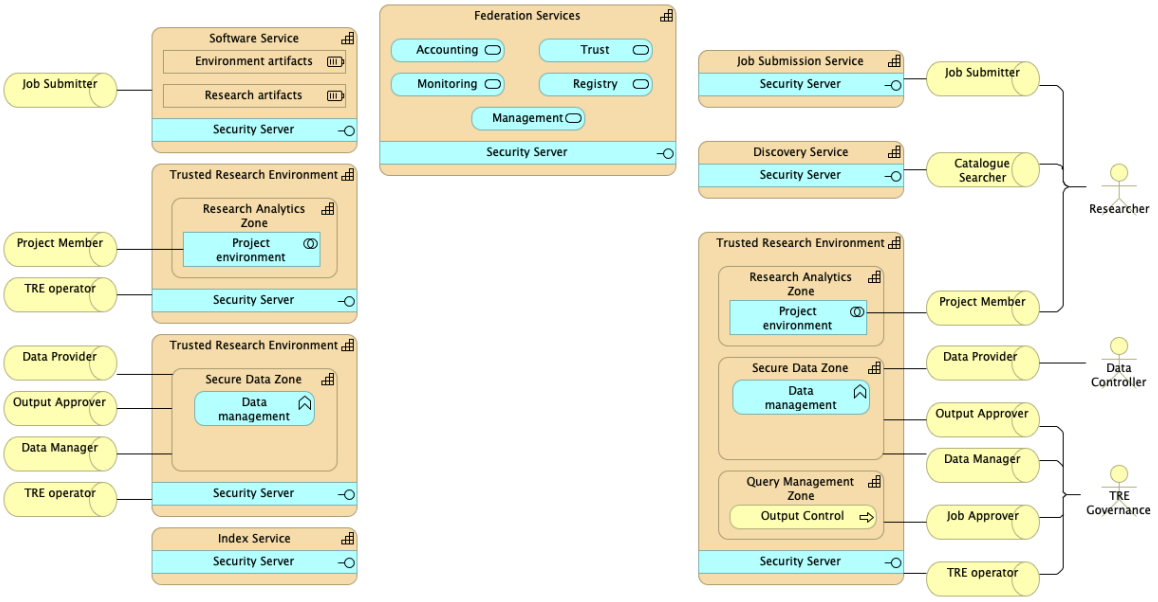

Figure 5.1.3 Minimal required functionality, actors, and roles in the DARE UK Blueprint infrastructure layer.

The DARE UK Blueprint identifies Researcher, Data Controller, and TRE Governance as actors in the federation. A Researcher can only access the federation as a Catalogue Searcher exploring externally available metadata through the Discovery Service, as an external Job Submitter to the Job Submission Service and Software Service, or as a Project Member with access to a Project Environment within a RAZ. Note that in the DARE UK Blueprint, a Researcher can only have access to sensitive data as a Project Member. The Project concept represents an approved research activity and contains information about individual Researchers that are part of the project and the data the project is authorised to use. Only authorised research data are available through the Project Environment.

Data Controllers are legally responsible guardians of sensitive datasets. In the DARE UK Blueprint, Data Controllers act as Data Providers releasing their datasets for authorised research projects through Data management in TRE SDZs.

TRE Governance is the team of people running a TRE. Whereas all TREs require a TRE Operator for running day-to-day technical services, TRE Governance takes on additional roles depending on the TRE's specific capabilities. Specifically, SDZs require a Data Manager who is responsible for the SDZ's data management functions and an Output Approver who is responsible for checking research outputs to be released from the TRE to the outside world. QMZ requires a Job Approver who is responsible for reviewing and accepting or rejecting jobs submitted from a Researcher acting as Job Submitter. Results from accepted jobs go through an Output Control process potentially involving Output Approver.

The remaining federation Participants - Index Services, Software Services, and Federation Services - represent distinct essential capabilities in the federation. Index Services provide linkage spines (see Figure 5.1.1) allowing SDZ Data management to link distinct datasets at the individual level and provide joint datasets for authorised projects. Software Services provide two types of resources: Environment artefacts, which represent Project Environment configurations or states, and Research artefacts, which is a proxy for external research artefacts, such as workflows, containers, or scripts, available in external repositories. Federation Services provide Accounting, Management, Monitoring, Registry, and Trust services that together provide the coordinating functions defining the federation. For example, the Registry service records the Participants that are part of the federation, the Datasets provided by Data Providers, the approved Projects, and users of the federation.

5.2 Mapping ENTRUST TRE Providers to the DARE UK Federated Architecture Blueprint

5.2.1 Mapping NORTRE to the DARE UK Blueprint 2.0

NORTRE[16] (Norwegian TREs) is a collaboration between the three main institutional research infrastructures for sensitive data in Norway, TSD[17] (services for sensitive data) at University of Oslo (UiO), HUNT Cloud[18] at the Norwegian University of Science and Technology (NTNU) and SAFE[19] (secure access to research data and e-infrastructure) at University of Bergen (UiB). The three partners share knowledge and expertise so scientists and data controllers from Norway and around the world can collect, analyse, store, share and collaborate on sensitive data in an optimised and trustworthy manner. HUNT Cloud is certified according to ISO27001, while SAFE and TSD are currently progressing towards this.

The Norwegian organisation around sensitive data differs significantly from the UK, in that the data holders do not have their own TREs associated with them. In the Norwegian model, the research institutions are mostly Data Controllers themselves, while the NORTRE partners serve as Data Processors. There was an attempt at building an all-in-one national health analysis platform (HAP) for Norwegian registry- and population based data, which incorporated all zones belonging to the DARE UK TRE. This was however paused in 2021 and eventually abandoned in 2022, due to violation of GDPR / Schrems II. At this point, the NORTRE partners were asked to provide an alternative solution based on their existing academic services, in which they would serve as an EU Secure Processing Environment (SPE) where researchers could have their data sent after a successful application to the

[16] https://nortre.no/
[17] https://www.uio.no/english/services/it/research/sensitive-data/index.html
[18] https://about.hdc.ntnu.no/
[19] https://www.uib.no/en/foremployees/131011/safe

national health data access body Helsedataservice[20]. As a result, NORTRE currently supports the Norwegian health authorities in developing EHDS-ready SPEs and effective and secure solutions for transporting data from the national health data access body to these[21] [22]. In addition to registry- and population based data, NORTRE partners also serve many users who have approval from the regional ethical boards to collect data themselves, which they then import into the TREs for analysis and management. In the first phase of EOSC-ENTRUST, SAFE and TSD have attempted to map their architecture to DARE-UK. HUNT Cloud will follow with a mapping in the next phase of the project.

SAFE (Secure Access to Research data and E-infrastructure) at UiB

SAFE was established in 2015 as a strategic investment for UiB to support and enhance research on sensitive data and is based on the Norwegian Code of conduct for information security in the health and care sector (Normen[23]). It is a specialised infrastructure developed by the IT-division at UiB for secure storing and processing of sensitive data for employees, students and external parties. SAFE serves about 2000 end-users in about 500 projects and hosts more than 1PB of data, from humanities, social sciences, law, economics, psychology, art and health. The solution consists of virtualized servers hosted at the university's on-premise data centre where users get a dedicated project-specific environment to work in.

Architecture and capabilities

The SAFE infrastructure (Figure 5.2.1) is based on virtual Windows- and Linux servers that integrate with a wide range of scientific software according to the user's needs, including secure video recordings, Whisper[24]-based automatic speech transcription and REDCap[25] for data collection. The technological configuration consists of VMware virtual machines with associated NetApp storage and backup. SAFE's infrastructure is segregated into two zones. The terminal server zone (VMWare hosted), and the file server zone (NetApp), roughly correspond to the DARE UK RAZ and SDZ zones, respectively. The terminal servers act as access points to the file server zone, where the sensitive information is stored. Both the virtual machines and the file servers have a high level of redundancy. The infrastructure is also secured within two distinct firewalls that separate SAFE from the outside world and separate the terminal server zone and file server zone, respectively.

Projects in SAFE are managed by a System Owner, who is usually, but not always, the project leader/PI. This user acts as the Data Controller, while SAFE acts as the Data

[20] https://helsedata.no/no/helsedataservice/
[21] https://helsedata.no/no/helsedataservice/nyhetsarkiv/ny-gap-report-fra-spuhin/ny-gap-report-fra-spuhin/
[22] https://www.regjeringen.no/no/aktuelt/meir-helsedata-til-forsking-innovasjon-og-naringsutvikling/id3056396/
[23] https://www.ehelse.no/normen/normen-for-informasjonssikkerhet-og-personvern-i-helse-og-omsorgssektoren
[24] https://openai.com/index/whisper/
[25] https://redcap.vumc.org/

Processor. Access to a SAFE server is granted upon approval of the project by a Regional Ethical Board.

Users log in to SAFE from their own client, which must adhere to the respective requirements for usage and connection to SAFE. They connect using their UiB account, which follows strict access management in Active Directory with specialised procedures for access to SAFE and a MakeMeAdmin[26] system for added security. External users get access through UiB guest accounts that need to be vouched for by the Data Controller and confirmed. All users connect to a SAFE VPN profile that is access-managed by the SAFE team, connect with Microsoft Entra MFA[27], and connect to the server via Remote Desktop, which is also access managed by the SAFE team. Within a SAFE server, the PI/ System Owner fully controls access management, which can be granulated on folder- and file level. This is done by updating an access document only available to them within SAFE and notifying the SAFE team, who will execute the change after an assessment of the details.

Files can be exported and imported through the use of the "sluice" file share system, which supports automatic virus scanning and encryption of files. Export is strictly access-managed, and by default only the System Owner can export and request export-privileges for other users. The files are automatically AES-256 encrypted upon export, after which the key is available from the user project area. The exported files are further logged and copied in a folder accessible only to the System Owner and the SAFE team for non-repudiation. Import is only available to the users with access to a SAFE server. Almost all of SAFE's operations and daily tasks are automated through PowerShell and Python scripts, especially related to access management in order to avoid as much human error as possible. An API for file transfer is also being developed for increased integrity and availability. Detailed guidelines and training material on how to use SAFE is provided through the UiB SAFE internal web pages.

[26] https://github.com/pseymour/MakeMeAdmin
[27] https://learn.microsoft.com/en-us/entra/identity/authentication/concept-mfa-howitworks

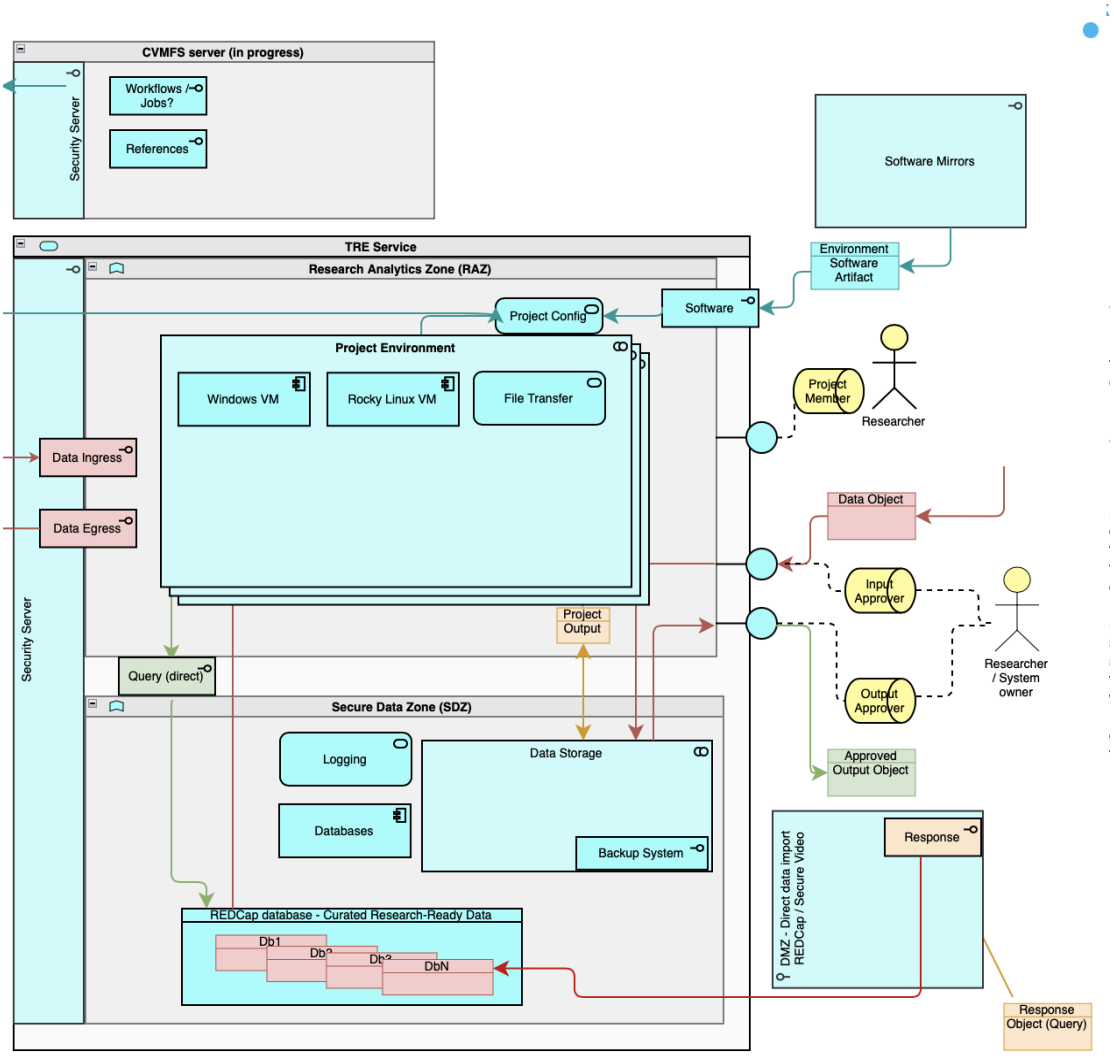

Figure 5.2.1 Mapping of the SAFE TRE architecture to the DARE UK Blueprint architecture graph, using the DARE UK layout. SAFE holds zones corresponding to the DARE RAZ and SDZ. Data input and output is controlled by the PI / System Owner.

Matches in the architecture

Some components within the DARE blueprint seem to describe corresponding components within SAFE reasonably well, such as the DARE RAZ and the SAFE terminal server zone. SAFE has a File server zone, which partly matches the DARE SDZ, but without the TRE governance actors described in the DARE definition. Access control is managed by the PI and executed by the SAFE personnel. A full list of matches is provided in Table 5.2.2.

Gaps in the architecture

As described in the NORTRE section, Norwegian data holders do not currently host their own TREs, but rather grant access to data by requiring that they are stored and analysed within SPEs. The requirements of SPEs are currently being developed by the national health

authorities. SAFE therefore does not hold the DARE components related to Information Governance and Query Management. Further, data input and output is controlled by the PI / System Owner on behalf of the University of Bergen as Data Controller, and not TRE Governance.

TSD (services for sensitive data) at UiO

TSD was established in 2012 and serves about 10,000 end users in 2000 projects from a wide range of scientific fields. TSD's services include secure processing environments that integrate with custom data collections tools for web and smartphones (Nettskjema), research participant consents services and high-performance computing (HPC). The technological configuration consists of VMware, a HPC system (3,200 CPU), an Enterprise Storage System and tape backup. Parts of the HPC capacity and storage is provided through the TSD-Sigma2 collaboration. TSD delivers Windows and Linux-based virtual machines to the end-users with a wide range of tools and software. Further details on the risk analysis for these services are provided here.

Architecture and capabilities

The Services for Sensitive Data (TSD; Figure 5.2.2) is a specialised eInfrastructure developed at the University of Oslo (UiO) that provides a secure, scalable Platform-as-a-Service (PaaS) for handling sensitive data in research. Designed to ensure compliance with GDPR, TSD offers virtual workspaces, remote login, and API integration to facilitate the secure collection, analysis, and publication of sensitive data, particularly in health and clinical research.

Through data processing agreements, TSD processes sensitive data on behalf of its customers from public, health, education, and commercial sectors, including some EU entities, while ensuring data controllers retain responsibility for obtaining the necessary legal approvals. Managing user accounts within the industry, especially for EU users, presents specific challenges due to regulatory requirements. For instance, creating accounts for EU users requires the Principal Investigator (PI) to provide detailed information about each user. The platform is deployed on-premises at UiO and supports services such as data collection, dynamic consent, high-performance computing (HPC), and data publication.

The infrastructure provides users with secure remote access to virtual machines (VMs) within a dedicated, project-specific virtual workspace. TSD implements several layers of security, including firewalls, segmented VLANs, and two-factor authentication for all logins. Project separation is maintained through a combination of VLANs, micro-segmentation, and firewalling, ensuring strict security and compliance for all hosted data.

TSD offers various tools and services for working with sensitive data, such as PostgreSQL and MSSQL database hosting, an integrated survey tool (Nettskjema) for secure data collection, and APIs for data transfer and system integration. HPC is supported through the Colossus cluster, which is optimised for data processing with Slurm[28] job scheduling, InfiniBand network and GPFS for high-speed file access. Access to the Colossus cluster is restricted and requires a Principal Investigator (PI) to formally apply for resources while applying for a project in TSD, ensuring that only authorised users can submit jobs. When submitting jobs, a user must log in to a designated host in the specific project (referred to as System VM in Project Environment), and users are guided to avoid including sensitive information in the filename of the job scripts. The `/proc` filesystem is mounted with `hidepid=2`, ensuring that non-root users can only view their own processes, and interactive jobs are not permitted so to prevent unauthorised access. The job submission system, managed by Slurm, assigns job IDs and maintains a controlled queue, further enhancing privacy. Each job receives its own temporary directories `/tmp`, `/var/tmp`, and `/dev/shm`, managed by a Slurm plugin, and it can also request dedicated scratch area on local disk, all of which are isolated from other jobs by permissions and/or private file system namespaces. Communication between compute nodes is restricted with no SSH access allowed within or between nodes, further ensuring that jobs cannot access each others' data. Job information is confined to each project, and system-wide cluster load is summarised anonymously through `qsumm`[29]. Finally, GPFS file systems prevent root access by remapping root's UID[30] to `65534/nobody`, adding another layer of security. . Together, these measures establish a secure environment for handling jobs for specific projects on the Colossus cluster.

TSD's self-service portal enables users to manage project access, credentials, and consent digitally, enhancing usability while maintaining security. The TSD Identity and Access Management (IAM) API provides a robust solution for managing users, groups, projects, institutions, and capabilities within the TSD ecosystem. Built using FastAPI[31] and based on the PG-IAM framework[32], this API allows seamless management of core entities such as persons, users, groups, institutions, and grants. Integration with external systems and apps is facilitated via secure APIs, enabling machine-to-machine communication using OAuth 2.0.

Physical security at TSD is ensured through controlled server-room access, surveillance, and disaster recovery plans. The platform also features robust monitoring, antivirus measures, and a backup system.

[28] https://www.uio.no/english/services/it/research/sensitive-data/help/hpc/job-scripts.html
[29] https://www.uio.no/english/services/it/research/sensitive-data/help/hpc/queue-system.html
[30] https://www.ibm.com/docs/sv/aix/7.1?topic=passwords-root-account
[31] https://fastapi.tiangolo.com/
[32] https://github.com/unioslo/pg-iam

TSD's compliance and security measures have been thoroughly evaluated, including unscheduled external penetration testing, to ensure robust defence against unauthorised access and data leakage. The most recent penetration test focused on two main objectives: first, to gain unauthorised access to the secure TSD environment from the Internet, and second, gain access to other projects' data within the TSD environment as an authorised user. Both of these objectives were not achieved. The system is also subject to regular audits, and every significant change undergoes risk evaluation.

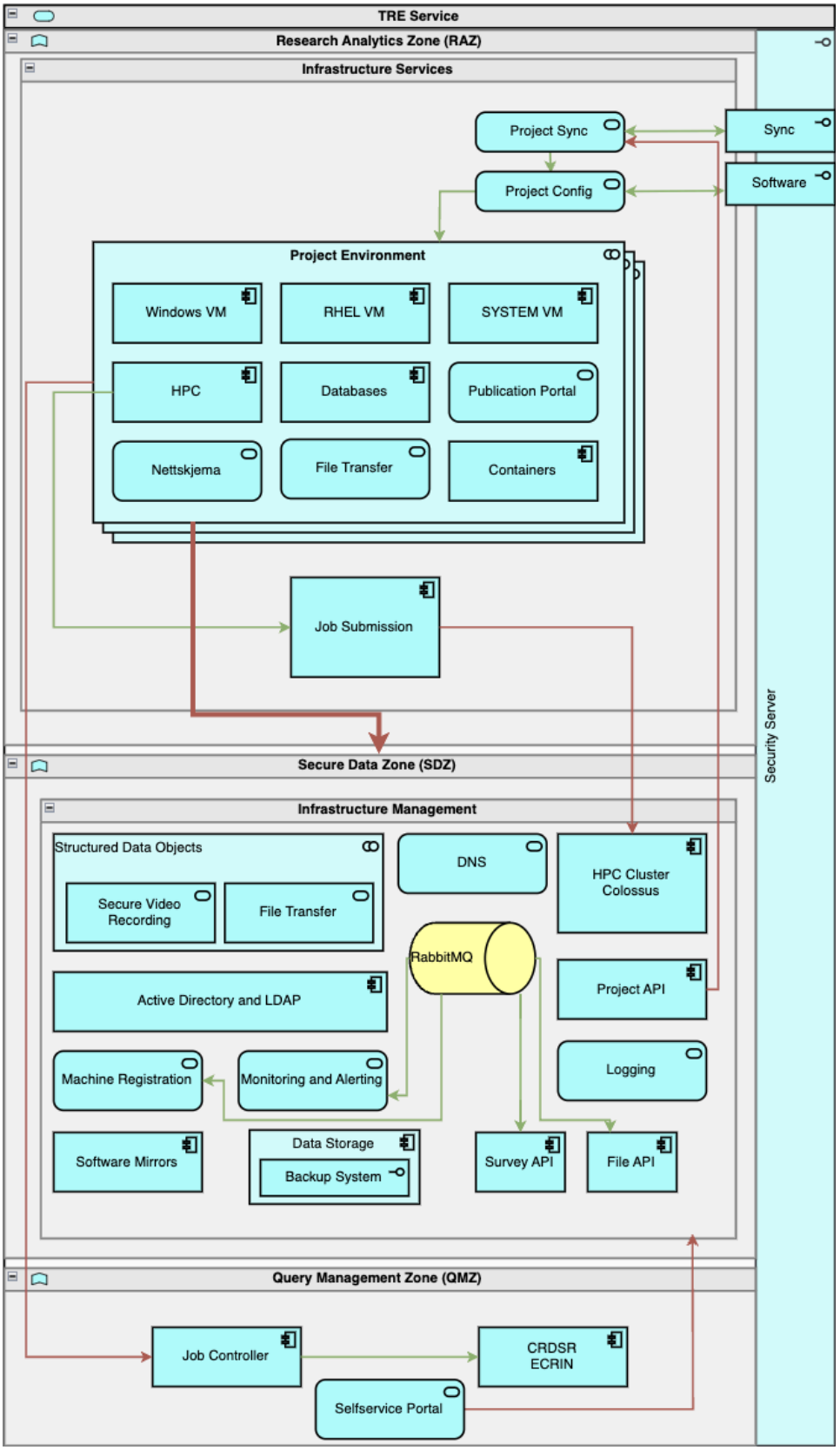

Figure 5.2.2 Mapping of the TSD TRE architecture to the DARE UK Blueprint.

Table 5.2.2 outlines the current capabilities and existing gaps within the TSD infrastructure.

Matches in the architecture

Data pooling, security server and job submission services, all operational and contributing to secure data transfer, query management, and remote analysis.

Gaps in the architecture

Federated analytics, indexing service, and discovery service are highlighted as gaps. These gaps represent areas for enhancement to achieve a comprehensive, interconnected TRE infrastructure.

In summary, while TSD is well-equipped for data access, governance, and analysis within a controlled environment, extending features like federated analytics and dataset discovery will further enhance its usability and support for collaborative research.

Towards a NORTRE federation

The NORTRE partners TSD and SAFE have demonstrated a federation between their services by allowing transport of data through the TSD API, as illustrated in Figure 5.2.3 below. This is further described in EOSC-ENTRUST Milestone 15[33].

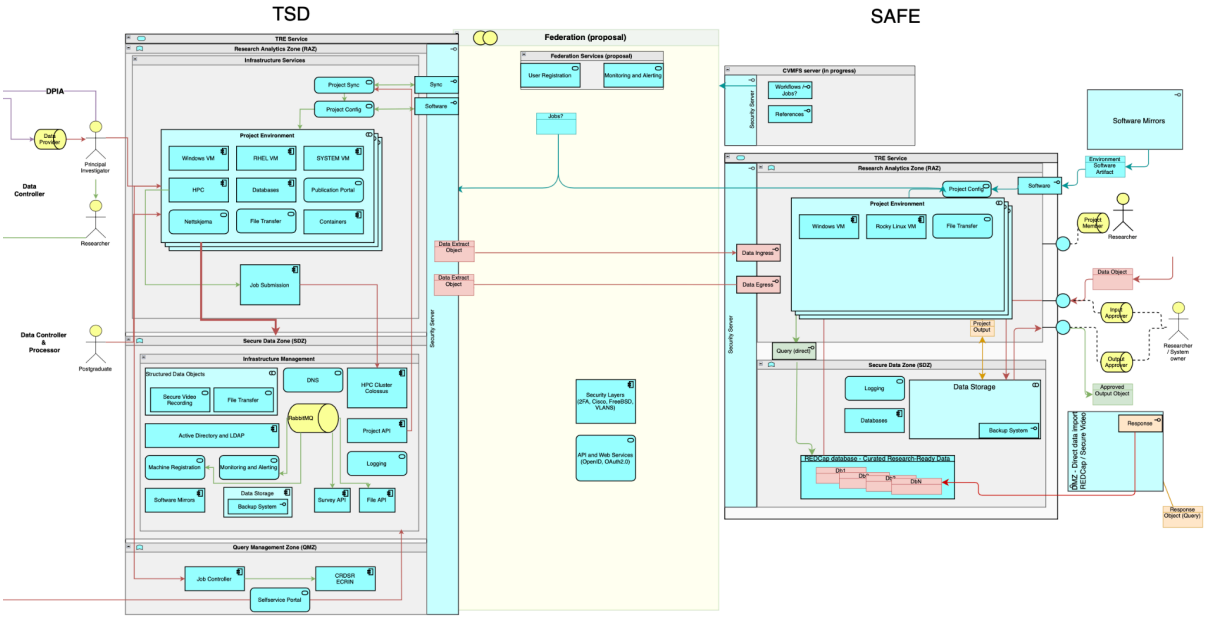

Figure 5.2.3 Current level of federation between the NORTRE partners TSD and SAFE, using the architecture layout provided in the DARE-UK Blueprint: Data transport through the TSD API.

[33] MS15 - Key selected interoperability challenge demonstrations implemented.pptx

5.2.2 Mapping SURF SANE environment to DARE UK UK Blueprint 2.0

SANE architecture and capabilities

SURF Secure ANalysis Environment (SANE) is a virtual, fully shielded computing environment containing pre-approved analysis software 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.

SANE is implemented with existing ISO27001-certified SURF services: SURF Research Cloud and SURF Research Access Management (SRAM).

The target system architecture in Figure 5.2.4 puts SANE (referenced as "Trusted Research Environment") in the context of a set of services and actors.

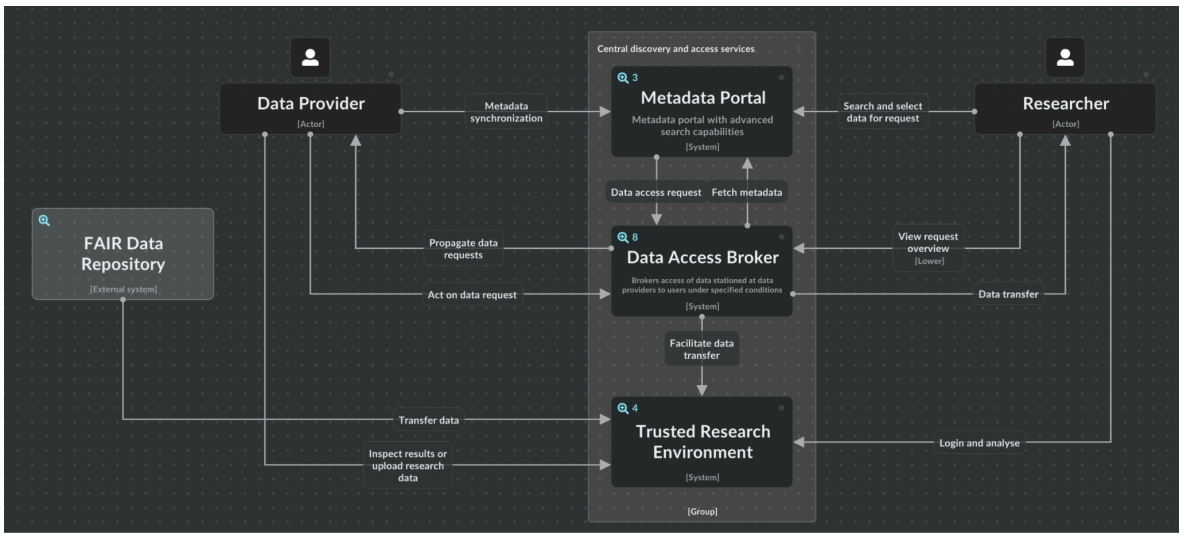

Figure 5.2.4 SANE in context of an overarching architecture depicting the more typical Dutch landscape. SANE is generically referenced here as "Trusted Research Environment" (IcePanel link: https://s.icepanel.io/1NHIm1tCPclyjm/nEES)

Mapping of architecture components

A mapping of the SANE architecture components to the DARE UK architecture components is summarised in Table 5.2.1. A visual mapping can be shown as in Figure 5.2.5.

Table 5.2.1 Mapping of SANE architecture components to the DARE UK architecture components.

DARE UK TRE Component	Corresponding SANE TRE Component	In production?	Description

Research Analytics Zone (RAZ)	Tinker SANE	yes	Tinker SANE provides the means for a Project Member to gain direct access to the data their project is approved to use, in an environment suitable for the analyses their research requires. It is a virtual desktop environment (Windows, Linux) that is restricted from internet access, and contains the software and tools that were requested (besides the default set of tools).
Secure Data Zone (SDZ)	Data Provider Portal	yes	This is the stepping stone environment for a data provider (i.e. data controller), or appointed Governance actors, to upload data in and out of the secured project space, either for data ingest or for disclosure control of output data. All other roles within a project are prohibited from accessing this component.
Query Management Zone (QMZ)	N/A		
Federation service	N/A		The Data Access Broker (see Figure 5.2.4) could potentially provide a standard way for a TRE to federate with a SANE-based TRE by automating data- or algorithm transfer (depending on what kind of federated analysis is needed).
Discovery service	Metadata Portal	yes	A service to discover datasets from various Data Providers by means of their metadata. A Metadata Portal is typically domain-specific and hosted by the corresponding communities. They typically offer advanced searching features, such as powered by knowledge graphs. Ideally a request for a dataset can be made from here, which is to be propagated to the corresponding Data Provider.
Job submission service	N/A		Blind SANE in some ways resembles the Job Submission Service, but lacks a scheduling mechanism (e.g. like Slurm). Blind SANE allows the Researcher to headlessly submit a workload to a compute environment that hosts the sensitive dataset in question. However, it does not allow you to interactively inspect the (running) job, nor allows you to export any logs that could potentially contain data that was not meant to be seen (for which Blind SANE was

			created).
Index service	N/A		Is not a specific service for SURF. Generating index keys is the responsibility of the Data Provider
Software service	Catalogue items in SURF Research Cloud	yes	Software for a SANE environment must be requested by the Researcher from the Data Provider. Once approved, software is made available through an existing set of Catalog Items or custom made via Ansible or Docker. During the cloud deployment of the SANE environment, the software is installed either from a public repository or a private repository if chosen by the Data Provider. After the deployment and before handing over access to the Researcher, all internet access is closed off to ensure the security of the environment.
Security Server	SRAM / Openstack	yes	SURF Research Access Management (SRAM) is used for federated identity management and for defining project groups. Openstack security group rules are used to control the egress / ingress of each SANE component (VPC, ports on VMs). Any federation between other TREs (not exclusively SANE) or connections to Data Provider repositories must be applied to the Openstack security group rules
Data Controller	Data Provider	yes	The Data Provider role is taken by a person who, or an institute which, has the authority to grant access to data in SANE. In some cases the Data Provider can choose to automate (part of) the approval process
Information Governance (IG)	Data Provider	yes	There is no distinction made between a Data Provider and a delegator role (such as an IG). A Data Provider is free to assign someone the data provider capabilities within a project or automate certain tasks
N/A	Blind SANE	yes	In a Blind SANE environment the Researcher(s) within the project cannot see or interactively interact with the data but instead must blindly execute their analysis on the data provided by the Data Provider. This is done by submitting a Python script (public URL) or a

			Docker image (or Dockerfile) during the configuration wizard of Blind SANE, which is then executed in a non-interactive VM.
N/A	Secure Data Storage	yes	In SANE the data that is made available by the Data Provider is explicitly NOT directly put into the SANE (Blind or Tinker) environment but in a shared storage system that is connected to the SANE and Data Provider Portal via a private network (VPC in Openstack). Therefore the shared storage is effectively shared between the Research and Data Provider role. This allows the Data Provider to easily add more data. It also allows the Data Provider to get easy access to any output (i.e. results) that the Researcher generated and wishes to export outside the TRE.
N/A	Data Access Broker (DAB)	no	A broker service that is able to automate the transfer of data from a Data Provider (who might not have a full blown TRE). The DAB is also able to forward data access requests made from the Metadata Portal to the corresponding Data Provider, who might have an automated response based on the provided metadata or data licence that are part of the data access request.

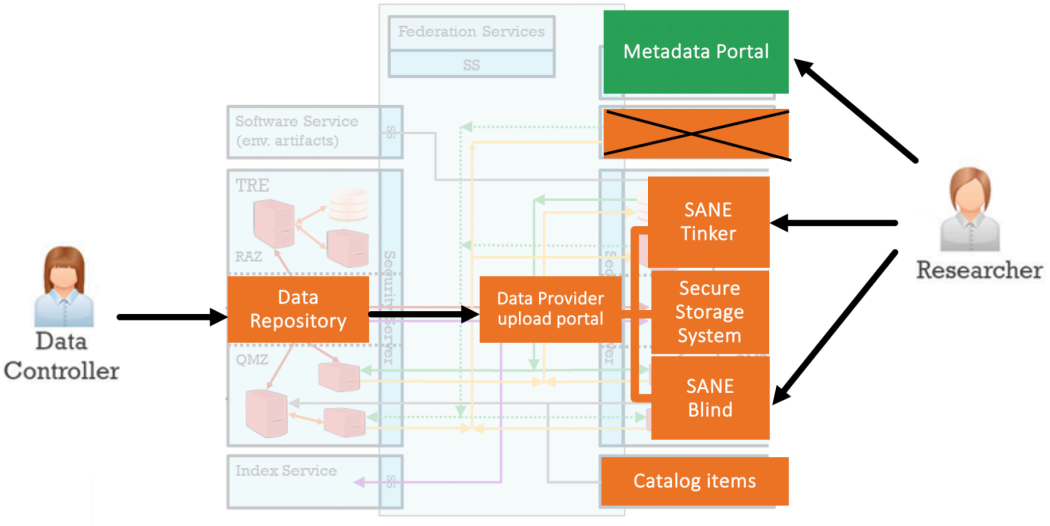

Figure 5.2.5 Visual Mapping of SANE architecture components to the DARE UK

architecture components.

Matches in the architecture

See table in the Mapping section and the corresponding description

Gaps in the architecture

The following gaps were identified to be present in the DARE-UK architecture but not in the SURF architecture:

- **QMZ**: Within the SURF architecture, the subselection of a dataset (by means of querying or otherwise) is left to the responsibility of the Data Provider. If the Data Provider supports a standard way of performing a subselection, then the Data Access Broker is the component within the SURF architecture that is able to request the said subselection and transfer the subselection to the corresponding SANE environment.
- **Job Submission Service**: Currently the SURF architecture does not have a way to submit jobs other than 'submitting' an analysis script or Docker container to be run on the provided data via Blind SANE. We anticipate that in order to serve scientific workflows SANE must also support a more 'ad-hoc' deployment of secure environments that have the same security features in place as Blind SANE but are short-lived and are scheduled by some orchestration service (e.g. Kubernetes or Nomad). Alternatively, a Slurm-based cluster or supercomputing facility could be used as a back-end for deploying the Blind SANE workload.
- **Index Service**: similar to the missing of a QMZ, the SURF architecture assumes that data indexing is done by the Data Provider and therefore falls outside the scope of a TRE provider. There are TRE providers in the Netherlands (e.g. CBS) that offer an index service, but SURF does not see this as a mandatory part of a generic TRE provider.

The following gaps were identified to be present in the SURF architecture but not in the DARE-UK architecture:

- **Blind SANE**: Some Data Providers are keen on not having Researchers actually see the data (e.g. publishers that provide their ebooks for research purposes). For these cases we provide Blind SANE so that analysis is still possible, albeit in a Blind Environment. A Blind SANE goes hand-in-hand with a 'Sandbox SANE' environment, which allows a subset or a synthetic dataset to be visibly used in a Tinker SANE-like environment in order to prototype the Blind SANE analysis script or container image.
- **Secure Data Storage (SDS):** Currently, with SANE, data is not directly placed inside the Tinker or Blind SANE environments but is rather placed in something called the Secure Data Storage (SGS). In the current SURF implementation the SDS is an

Openstack volume that is mounted to a “Secure Data Server” which is a non-interactive VM that hosts a Samba server. The Secure Data Server is only discoverable in the private network. The SDS enables additional data and software packages to be easily added via the Data Provider Portal.

5.2.3 Mapping CSC SD Services to the DARE UK Blueprint 2.0

The design of end-user services at CSC – IT Center for Science Ltd is limited by its status as the Finnish national research service provider where contracts with ministries fully pay the costs. CSC cannot charge end users for its services. As a result, CSC can maintain only scalable automated services where its role is limited to the data processor. This has direct implications for the service architecture.
The CSC SD Services[34] are currently deployed in three use cases relevant to this analysis with different requirements: academic use, secondary use of health data under national authority of Findata, and secondary use of health data under a single register holder.

SD Services have been built as a general-purpose sensitive data management environment to support both the active research phase and the re-use of previous research outcomes. It was planned as a second-generation environment with a modular design based on experiences gained from first-generation, manually maintained environments. The goal was to maximise the security by design to embed security features seamlessly into the user interface and to make it scalable by distributing manual processes to data controllers, end-users, and supervisors.

[34] https://research.csc.fi/sensitive-data-services-for-research

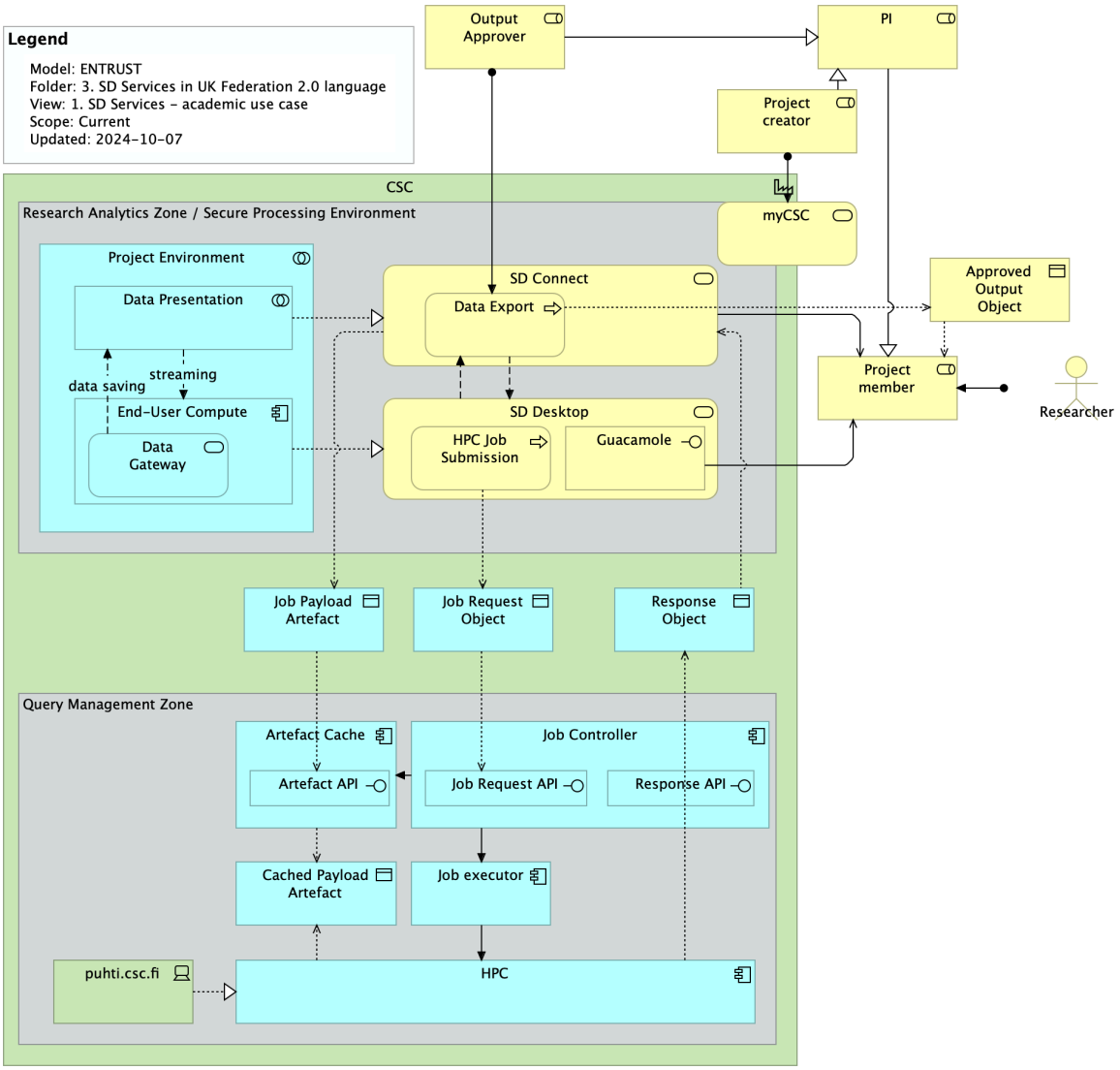

Figure 5.2.6 Academic use of SD Services.

The primary use of research data (Figure 5.2.6) assumes that researchers have full control of the sensitive data they want to process. In the case of personal data, this usually means that they work on consented social, health, or genetic data. It is assumed that the environment should promote communication and collaboration among the users over restrictions and that all users in the project are equal except when exporting results out of the system. Export is the most critical processing step and needs confirmation from an appointed project member, "Output Approver". This role is initially give to PI who can pass it on to another project member.

The sensitive data processing environment for project members, "Project Environment", is an instance of an isolated environment of the TRE "Secure Analytics Zone" that corresponds to the EU Secure Processing Environment (SPE) concept. It consists of two parts: SD

Desktop[35] which is a network-isolated virtual computer, "End-User Compute", and SD Connect[36] which is a secure data storage that does automatic encryption of all data passing through it. SD Connect is the only way to get data into SD Desktop for processing and it does this by exposing decrypted contents to SD Desktop using the streaming capabilities in the secure and public crypt4gh algorithm. When needed, users may copy parts of this content to internal drive or volume. The Data Export tool of SD Desktop allows users to copy files back to SD Connect for more permanent storage, transfer to other services, or export out of the environment.

Project creation, service selection, and user approval are done with a dedicated web-based, company-wide tool myCSC without any intervention from service maintainers. Users login with a CSC identity but authentication is based on a federated system where identity is provided by the users' organisation. Access to secure services needs a separate multi-factor authentication (MFA) step.

The alpha version of an "Indirect Query" implementation of “Job Submission Service” to send secure HPC requests to one of the CSC supercomputers in "Query Management Zone" is in place. Users compose a job request file that combines compute and storage description with data and software references and launch it. The job enters a dedicated workload manager queue for sensitive data. When the job enters the execution, it reserves a compute node, transfers all data into it, isolates it completely during the execution, transfers encrypted results to the user's SD Connect, and cleans up the node and caches.

Academic users can also apply to re-use published access-controlled datasets using tools that are described later in the secondary use single registry holder use case.

[35] https://research.csc.fi/-/sd-desktop
[36] https://research.csc.fi/-/sd-connect

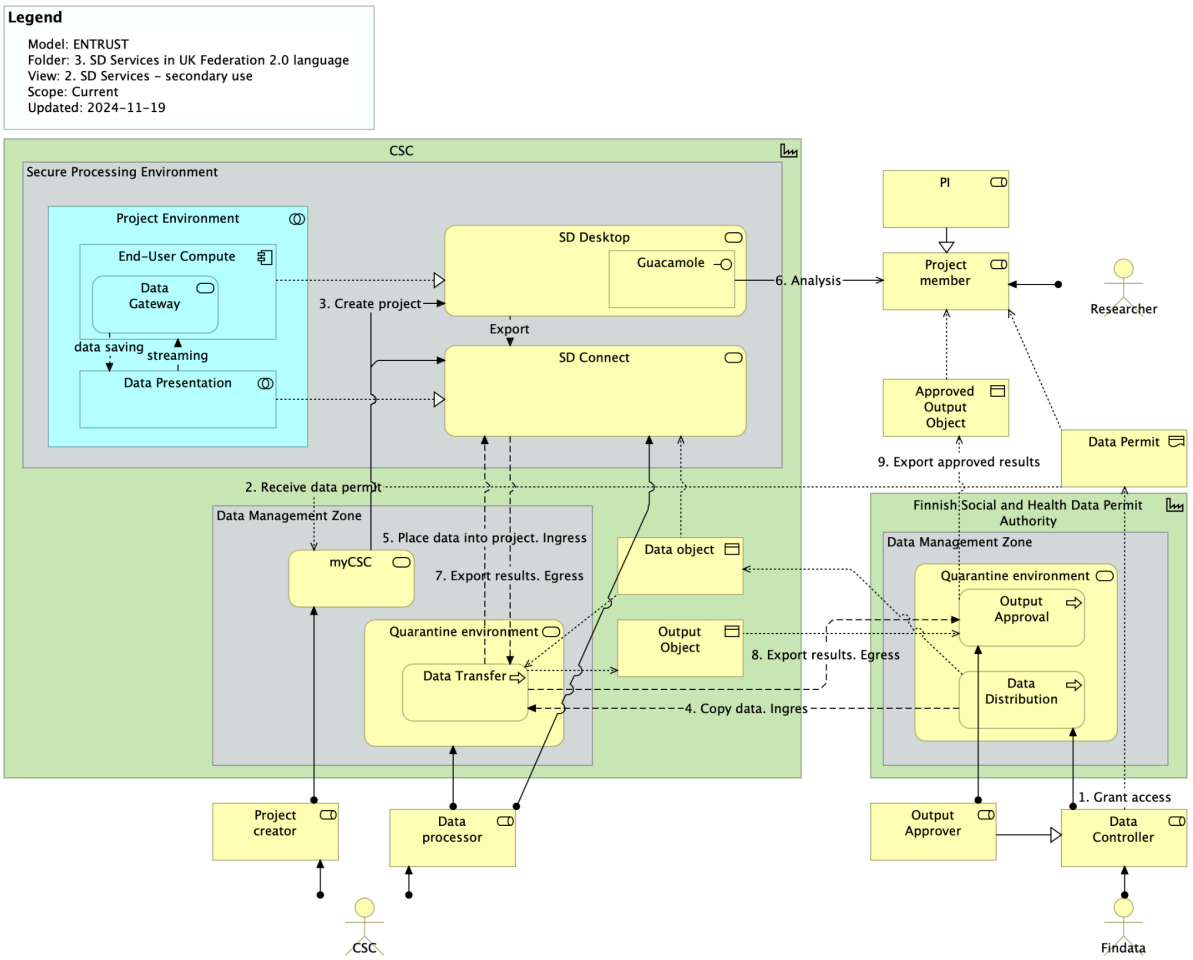

Figure 5.2.7 SD Services support for the Finnish Secondary Use Act under the national authority.

The 2019 Finnish Secondary Use of Health Data Act[37] and the subsequent regulation on secure data processing environments[38] added complications to the academic use workflow (Figure 5.2.7). The act established the Finnish Social and Health Data Permit Authority, Findata, to coordinate and manage the use of social and health registry data[39]. Users need an explicit permit for the data that may only be processed within a secure environment. To fulfil these requirements, CSC is obliged to manually establish the project for the permit holders and transfer the dataset from Findata, effectively establishing the TRE "Data Management Zone". After accessing their environment, users have no other means to import data than through the system clipboard.

Manual intervention is also needed for returning research results to Findata for sensitivity inspection. Technically, this means that users have no access to their SD Connect and they

[37] https://findata.fi/en/services-and-instructions/legislation/
[38] https://findata.fi/en/services-and-instructions/regulations/
[39] https://research.csc.fi/example-case-7-sensitive-data-reuse

are not able to import and export data themselves. The SD Connect can only be used by a service provider administrator.

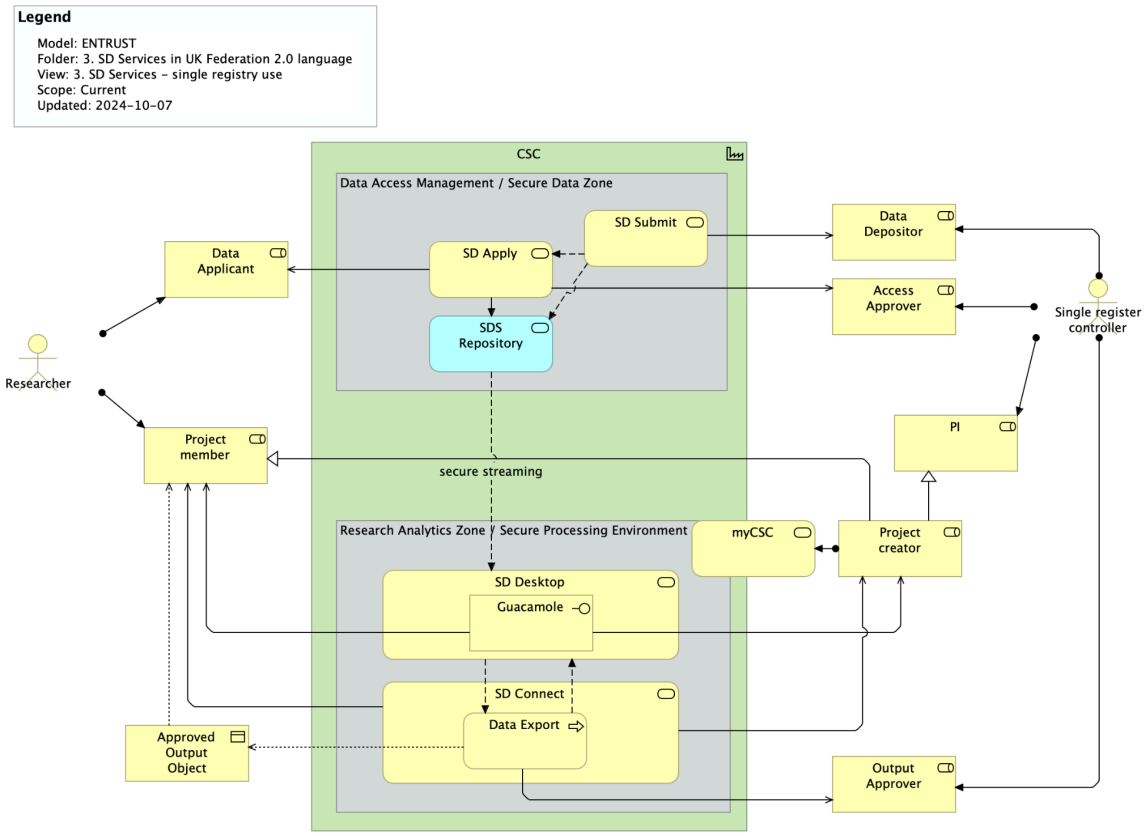

Figure 5.2.8 SD Services support for Finnish Secondary Use Act for data controlled by one data controller.

The Secondary Use Act did not transfer all registry data controller responsibilities to the central authority. When data access is requested for **datasets available from one health data controller**, it is allowed to manage the process without Findata. The most common application of this at CSC is when a medical student needs to work on clinical thesis data under the supervision of a senior adviser. The advisor as the data controller uses the automated secondary use data management service SD Submit to deposit the data and make it available to the student only through the access control service SD Apply which enables the strong identification of the user to automate the authorisation of the data processing within SD Desktop. Since the availability of the dataset is not made public, it is not necessary to have detailed, formal metadata included in the submission. The supervisor can deliver those as part of the dataset and give additional verbal instructions. The supervisor is the project owner and has the power to inspect the security of exported results (Figure 5.2.8).

This extends to the DARE UK Blueprint concept "Secure Data Zone" to data access management that is not currently covered in its architecture. The detailed description of these CSC use cases try to emphasise the the importance of integrated data access management to the overall functionality, security and scalability of sensitive data management. Alternatively, data access management could be seen as a separate federation service that ensures secure and timely delivery of sensitive data based on active permits.

In addition these SD services, CSC maintains the public metadata discovery service Etsin[40] which is part of CSC FAIR data services that matches Federation service “Discovery Service”.

Matches in the architecture

SD Services match well to the overall structure of the DARE UK Blueprint architecture. With the abundance of detailed requirements, the Finnish legal regulation fails include secure and practical rules for transferring sensitive data into secure processing environments. This defining of rules to classify objects based on the sensitivity levels and linking them to secure transfer protocols is also the core of the interoperability requirements.

Gaps in the architecture

The importance of secure data transfer also highlights the blind spot of the DARE UK Blueprint. In taking the TRE as a starting point of its exploration of interoperability requirements it pools several functional service classes under one term. A highly secure approach for interoperability needs to treat each of three functional zones of TRE to have clearly defined, secure data transfer interfaces between them. In the EOSC interoperability plans this will need to be taken into account.

The Data Access Management capability is not part of the DARE UK Blueprint. The SD Apply service demonstrates its importance to both individual TREs and any future federations of sensitive data processing. The underlying, standardised technology unifies secure storage of data[41], data access management[42], and automatic streaming of data to the secure processing environment for the permit holder[43]. The benefits of this approach in increased

[40] https://etsin.fairdata.fi/
[41] Crypt4GH: a secure method for sharing human genetic data https://www.ga4gh.org/news_item/crypt4gh-a-secure-method-for-sharing-human-genetic-data/
[42] REMS, Resource Entitlement Management System is a tool for managing access rights to resources, such as research datasets, https://github.com/CSCfi/rems
[43] GA4GH Passport https://www.ga4gh.org/product/ga4gh-passports/

security and scalability when associated with high level of assurance of user identification needs to taken into account an future interoperability plans.

5.2.4 A summary of EOSC-ENTRUST TRE Providers mapped to DARE UK Blueprint

In summary, most features and components presented in the DARE-UK architecture blueprint can be found also among these EOSC-ENTRUST TRE Providers that were mapped to it, with the exception of the Federation service, see Table 5.2.2. A fundamental difference between the DARE UK Blueprint and the TREs SAFE, TSD, and CSC is that PIs instead of TRE Governance act as Output Approvers for Project environments. Conceptually, PIs are specialisations of researchers and project members, but in the DARE UK Blueprint, Output Approvers always belong to TRE Governance. As a federation among TREs is not yet well developed anywhere in Europe, it is not surprising that components related to this are least developed among the TREs, although the TSD API facilitating data transfer between TSD and SAFE (demonstrated in Milestone 15) is a beginning.

In a wider context, the next generation European Genome-phenome Archive[44] (EGA) is a federation. The Federated EGA[45] (FEGA) is formed by national nodes. Two of the Providers in this evaluation are also FEGA nodes: FEGA Norway[46] is built on TSD technology and FEGA Finland[47] is based on SD Services technology. They both share the same functionality of providing national storage of submitted sensitive data that is handled through secure, standard APIs. Datasets in FEGA nodes are available to users through data applications approved by the dataset specific Data Access Committees and approved datasets are streamed to secure processing environments of the federation.

Table 5.2.2. DARE-UK TRE components mapped to selected EOSC-ENTRUST TRE Providers.

DARE-UK TRE component	Description	SAFE (UiB)	TSD (UiO)	SANE (SURF)	CSC SD
Research Analytics Zone (RAZ)	Project members can gain access to data approved for analytics.	YES, Terminal server zone	YES	Tinker SANE	YES, SD Desktop
Secure Data Zone (SDZ)	The Secure Data Zone (SDZ) is a business zone dedicated to data management and information governance. In this environment, data managers have full access to the SDZ,	YES, File server zone similar, but PIs manage access themselves	YES	Secure Storage System	YES; SD Connect

[44] The European Genome-phenome Archive https://ega-archive.org/
[45] FEGA https://ega-archive.org/about/projects-and-funders/federated-ega/
[46] FEGA Norway https://ega.elixir.no/, https://tryggve.tsd.usit.uio.no/
[47] FEGA Finland https://research.csc.fi/-/fega/

	while researchers do not.				
Query Management Zone (QMZ)	It sits alongside an SDZ and provides different methods of access to approved research-ready datasets stored within the SDZ.	Project specific, certain users have built databases featuring this	YES[48]	Blind SANE	YES, SD HPC
Indexing Service	A trusted service to link datasets across different TREs for creating a unified research dataset	NO	NO	Is not a specific service, generating index keys is the responsibility of the data provider	NO
Discovery Service	A service to find and assess datasets across federated TREs.	NO	NO	Metadata Portal	YES, Fairdata.fi Etsin
Job Submission Service	Service for remote execution of analysis jobs, including submission and security validation.	NO	YES	Step in the Blind component	YES, As part of SD HPC
Software Service	Provide access to sources of software from outside the Federation.	YES	YES	Catalogue items in SURF Research Cloud	
Federation service	The Federation Services group comprises services for registries (of services, users, projects, etc.), trust (security certificate management and signing), management (of standard shared software), monitoring and accounting.	NO	NO	N/A	NO
Security Server(SS)	Secure gateway for all inter-TRE traffic to maintain CIA[49]	Indirectly, through the TSD API	YES[50]	SRAM is used for federated identity management. Openstack security group rules are used to control the egress / ingress between TREs	Federated identity governance, OICD JWT identities, SD Connect, GA4GH standard passports, encryption and streaming of data

[48] P.K. Dahl, Interoperability challenge demonstrations, https://doi.org/10.5281/zenodo.13860340
[49] CIA triad: Confidentiality, Integrity and Availability
[50] Supported by OIDC JWT tokens as electronic identities, GA4GH passports are yet to be fully implemented.

Data Controller	In the TRE domain, Data Controllers take the role of Data Providers, releasing data approved for research to projects via TRE Data Managers.	Approval occurs outside SAFE, once inside SAFE, the users control the data themselves	Yes[51]	Data Provider	Data Controller process for secondary use of health data is external to CSC, for all other use cases SD Submit and SD Apply
Information Governance (IG)	information governance (IG) professionals act as intermediaries between data providers and data consumers, ensuring all necessary ethical, data protection and legal approvals are in place for a research project to proceed.	As above	Yes[52]	Data Provider (is responsible). Data Access Broker can be used to automate certain tasks	YES for FEGA service, but for other use cases it is Data Provider responsibility external to CSC.
Data Pooling	Moving datasets between TREs and linking them within a single TRE	TSD-SAFE API	YES[53]	YES	SD Connect
Governance Metadata	Metadata that capture information about what purposes the Federation is being used for / Standards to define who has access to data and under which conditions.	NO	YES	YES	YES
Authentication & Authorization	Establishing user identity and managing permissions for access to sensitive data.	YES	YES	YES	YES

5.3 Mapping ENTRUST Driver requirements to the DARE UK Federated Architecture Blueprint

The ENTRUST Driver WP (WP7) milestone report[54] provided an initial combined set of TRE requirements from the four ENTRUST use cases (Drivers). The requirements were based on defined user journeys for each Driver and were analysed within the context of the SATRE[55]

[51] In projects where a researcher is affiliated with UiO as the institution responsible for research, UiO assumes the roles of both data controller and data processor.
[52] https://tryggve.tsd.usit.uio.no/docs.html
[53] As demonstrated; the transfer of data between TSD and SAFE.
[54] Anne van der Kant, Jan-Willem Boiten, Milestone report M7.1 Initial Driver Requirements for TREs from the four Drivers [unpublished].
[55] https://satre-specification.readthedocs.io/en/stable/

framework. The work resulted in a set of requirements per Driver, with each requirement either being mapped to specific SATRE capabilities or being identified as missing from the SATRE framework. In the following sections, we have further analysed each Driver's user journey and requirements against the DARE UK Blueprint, focusing on mapping the Driver's data usage patterns against the DARE UK Data Space Model (see Section 5.1.1) and on mapping requirements missing from the SATRE framework or other requirements related to interoperability against the capabilities in the DARE UK Blueprint infrastructure layer (see Section 5.1.2).

5.3.1 Driver 1 – Federated Human Genomics

Driver 1 tackles secure management and cross-border sharing of sensitive genomics data for research, by providing an infrastructure that makes genomics datasets FAIR while ensuring that each dataset's holder (Data Controller) maintains control of how and by whom the dataset is reused. Dataset reuse (see Section 5.1.1) (Q0) and meta-analyses between datasets (Q1) are clearly relevant for Driver 1, as is combining genomics datasets with personal data from other sources (Q2, Q3), such as health data or population-based studies. Note, however, that usage patterns Q2 and Q3 typically require the Data Controller to facilitate linking with other data, as individual-level data are typically stored with dataset-specific pseudonymized IDs. Approving and facilitating dataset reuse with other sources (Q2) therefore falls on each dataset's Data Controller, making research involving usage pattern Q3 extremely challenging.

The Driver 1 user journey steps involve data discovery, access application, data access and analysis, export of results, and compliance and reporting. Of these, data discovery, data access and analysis, and export of results map directly to the DARE UK Blueprint Discovery Service and RAZ capabilities and Output Approver role, respectively. Compliance and reporting may be facilitated through the DARE UK Blueprint Software Service and Federation Services capabilities, whereas dedicated services for handling data access applications are missing in the DARE UK Blueprint.

Within the ENTRUST Driver WP (WP7) milestone report, Driver 1's gap analysis identified the following three key requirements:

- (D1.C1) anonymisation,
- (D1.C2) internet connection from within the TRE, and
- (D1.C3) scalable infrastructure.

Internet connection from within the TRE (D1.C2) and scalable infrastructure (D1.C3) can be mapped to the DARE UK Blueprint's Software Service capability and the DARE UK Blueprint itself, respectively. Anonymisation (D1.C1), in the form of guidelines for techniques to protect individual privacy across all participating institutions, is indirectly addressed in the DARE UK Blueprint. Specifically, whereas the DARE UK Blueprint does not specify specific requirements for protecting individual privacy, it assumes that the participating TREs follow

a commonly agreed minimal standard for information security management systems, thereby ensuring that data privacy is protected within each TRE. The DARE UK Blueprint's Security Server then ensures privacy when exchanging data between TREs in the federation. The result is that the DARE UK Blueprint satisfies the Five Safes principles for research on sensitive data within a network of participating TREs. We do note, however, that Driver 1 may require specific pseudonymisation services to facilitate dataset reuse.

5.3.2 Driver 2 – SSHOC – Social Sciences & Humanities Open Cloud

Driver 2 aims to ease the sharing of sensitive administrative and social science data across countries. Analyses involving geographically distinct datasets (Q1) are consequently Driver 2's main focus, but Driver 2 also mentions several use cases involving combined analyses with other data (Q2, Q3), such as integration with health records or questionnaires, data transfer of digital behaviour datasets, and linking anonymised surveys with sensitive data.

The Driver 2 user journeys include data discovery, project application, user certification, data access and analysis, and handling output requests. Of these, data discovery, data access and analysis, and handling output requests map directly to the DARE UK Blueprint Discovery Service and RAZ capabilities and Output Approver role, respectively. Dedicated services for handling project applications and user certification are out of scope in the DARE UK Blueprint.

Within the ENTRUST Driver WP (WP7) milestone report, Driver 2's gap analysis identified the following six key requirements:

- (D2.C1) data transfer between environments,
- (D2.C2) compliance with various national legal frameworks for data sharing,
- (D2.C3) mechanisms for timely deletion of data post-expiration,
- (D2.C4) safe researcher training and certification,
- (D2.C5) creation and dissemination of general training modules, and
- (D2.C6) data citation.

Secure data transfer (D2.C1) is part of the DARE UK Blueprint and is handled by the Security Server combined with the Index Service and SDZ or QMZ depending on the data usage case. Timely data deletion (D2.C3) is handled by RAZ Project environments with Project-specific durations. Data citation (D2.C6) is partly handled as datasets already available within the federation are findable and citable through the DOI Discovery Service. Datasets produced by a research project – for example, by collecting own data or linking or curating existing datasets – may be made available by ingesting these into SDZs, but an explicit service for publishing such datasets are out of scope in the DARE UK Blueprint. As the DARE UK Blueprint is a technical blueprint for federation, explicit modelling of compliance with multiple frameworks (D2.C2) is out of its scope. Dedicated services for researcher training and certification (D2.C4) and general training (D2.C5) are out of scope of the DARE UK Blueprint.

5.3.3 Driver 3 – Clinical Trials Data

Driver 3 works to identify and overcome challenges in sharing and reusing data within the clinical research community. This includes both dataset reuse (Q0), meta-analyses between different clinical trials datasets (Q1), and combining clinical trial data with health care data from the same individuals (Q2).

The Driver 3 user journeys involve data discovery, project application, user training in information governance (IG), data access and analysis, handling output requests, and publishing with data credit attribution. Of these, data discovery, data access and analysis, and handling output requests map directly to the DARE UK Blueprint Discovery Service and RAZ capabilities and Output Approver role, respectively. Dedicated services for handling project applications and user IG training and certification are missing in the DARE UK Blueprint. Publishing with data credit attribution will likely require persistent Dataset identifiers and services for publishing Datasets within the federation, which are missing from the DARE UK Blueprint.

Within the ENTRUST Driver WP (WP7) milestone report, Driver 3's gap analysis identified the following nine key requirements:

- (D3.C1) data encryption for data in transit,
- (D3.C2) software applications,
- (D3.C3) archiving for reproducibility or future validation,
- (D3.C4) basic training in using the TRE's project space,
- (D3.C5) data kept only within the agreed time period,
- (D3.C6) credit attribution to the original data generator,
- (D3.C7) dataset discovery,
- (D3.C8) metadata,
- (D3.C9) data standardisation/data anonymisation.

Secure data transfer (D3.C1), timely data deletion (D3.C5), and dataset discovery (D3.C7) are part of the DARE UK Blueprint (see D2.C1, D2.C3, and the analysis of Driver 3's user journeys, respectively). Both standard and custom software applications (D3.C2) are handled through the Software Service capability, but machine learning software involving training and exporting models on federation data would likely require specialised Output Control. Archiving for reproducibility (D3.C3) can be done by using Software Service to store Project environments, including project analysis software, and SDZ for archiving project data; however, the latter may require dedicated archiving or publishing services that are out of scope of the DARE UK Blueprint. Metadata specifications (D3.C8) are part of the DARE UK Blueprint data layer; any domain-specific needs could potentially be handled by domain-specific SDZs and Discovery Services. The D3.C9 requirement for data anonymization resembles that of Driver 1 (see D1.C1); data standardisation should likely be handled as part of a domain-specific dataset publishing service (see D3.C3). A dataset publishing service would also facilitate credit attribution to the original data generator

(D3.C6). Dedicated services for basic training in using Project environments (D3.C4) is missing in the DARE UK Blueprint.

5.3.4 Driver 4 – Health & Environmental Science in Public-Private Partnerships

Driver 4 focuses on collaborative data processing between higher education and research institutions and private sector entities. Data analyses combining datasets that are geographically distinct (Q1) or from different domains (Q2) are likely relevant; for example, Driver 4 mentions combining health records with data from health technologies, such as wearable devices.

The Driver 4 user journeys focus on the process of setting up a collaborative project within the specific TRE implementation SD Services (see Section 5.2.3). This process largely involves providing role-based services within Project environments.

Within the ENTRUST Driver WP (WP7) milestone report, Driver 4's gap analysis identified the following three key requirements:

- (D4.C1) service design,
- (D4.C2) ethical guidance, and
- (D4.C3) backup.

Service design (D4.1) may largely be facilitated by providing role-based services within Project environments, but could also require interaction with other DARE UK Blueprint capabilities such as Discovery Service, Index Service, and Software Service. Whereas Software Services (environment artifacts) can store and provide specific environment configurations, backup of project environment contents (D4.3) is typically handled by individual TREs. Ethical guidance (D4.C2) of users will likely involve training and certification services not covered by the DARE UK Blueprint.

5.3.5 Summary of Driver Requirement Mapping

Table 5.3.1 summarises the identified Driver requirements and how these map to components in the DARE UK Blueprint.

Table 5.3.1 Mapping of key Driver requirements to the DARE UK Blueprint. "Requirement" lists Driver requirements. "DARE UK Blueprint component" lists matching Blueprint components; a "–" indicates that components supporting the requirement are missing. Columns "Driver 1" - "Driver 4" show the source of the requirement, indicated by "X".

Requirement	DARE UK Blueprint component	Driver 1	Driver 2	Driver 3	Driver 4
Data discovery	Discovery Service	X	X	X	
Data access and analysis	RAZ	X	X	X	X

Handling output requests	Output Control	X	X	X	
Encrypted data transfer	Security Server	X	X	X	X
Software applications	Software Service			X	X
Timely data deletion	Project environment		X	X	
Archiving for future validation	Software Service, SDZ			X	X
Internet connection from within TRE	Software Service	X			
Scalable infrastructure	Blueprint itself	X			X
Metadata	Data layer			X	X
Service design	Project environment				X
Compliance with varied legal frameworks	–		X		
Compliance and reporting	(Federation Services, RAZ)	X			
Data access requests	–	X	X	X	
Safe User training and certification	–		X	X	X
General user training	–		X	X	
Pseudonymization	–	X		X	X
Data publishing (FAIR)	–	X	X	X	X

5.4 Mapping ENTRUST-aligning European sensitive data projects to the DARE UK Federated Architecture Blueprint

5.4.1 The EU Secure Processing Environment (SPE)

The EU Data Governance Act (COM/2020/767) amended the General Data Protection Regulation (2016/679) with elements addressing personal information and research. According to this act, when a data user accesses data from a data subject it should happen in a Secure Processing Environment (SPE):

> *the physical or virtual environment and organisational means to ensuring compliance with the requirements of Regulation (EU) 2016/679, in particular data subjects' rights, intellectual property rights, and commercial and statistical confidentiality, integrity and accessibility, ensuring compliance with applicable Union and national law, and allowing the entity providing the secure processing environment to determine and supervise all data processing actions, including to display, storage, download, export of the data and calculation of derivative data through computational algorithms*

The relation of SPE to the independently developed concept of Trusted Research Environment (TRE) that has no single definition has been unclear. With the development of the DARE UK Blueprint 2.0 that divides TRE in three functional zones (Figure 5.1.3), it can

be seen that the EU SPE corresponds to the TRE Research Analytics Zone (Figure 5.4.1) as they both provide sensitive data processing to data users.

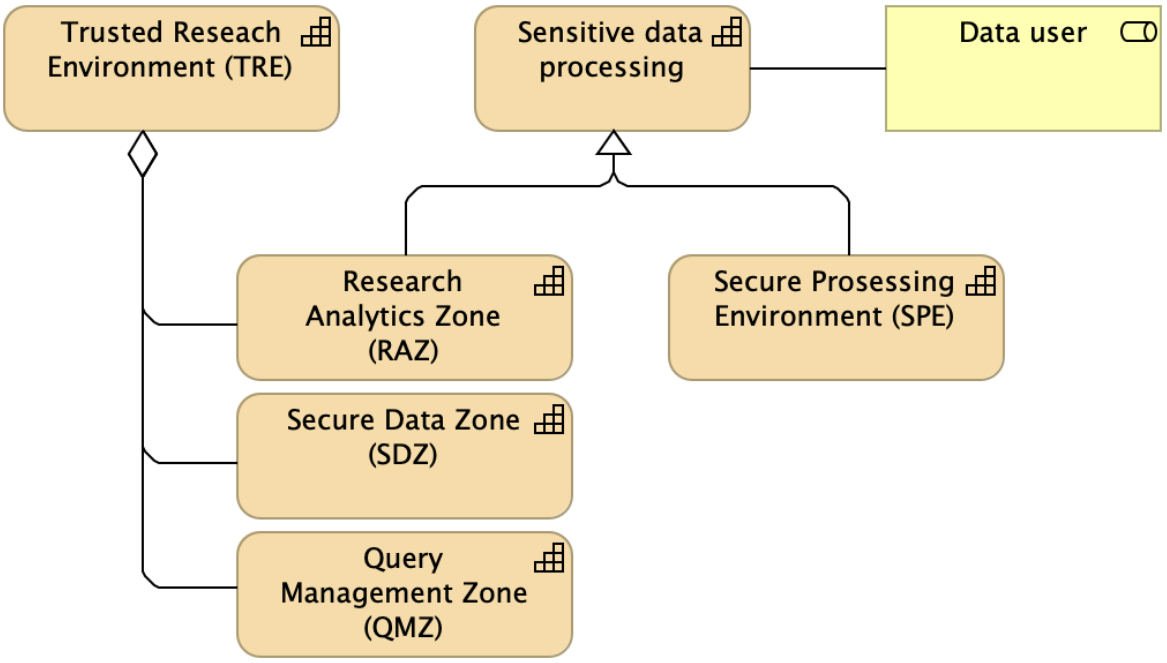

Figure 5.4.1 Secure Processing Environment (SPE) corresponds to the Research Analytics Zone (RAZ) functional component of the UK Trusted Research Environment (TRE).

5.4.2 The European Health Data Space

The almost finalised European Health Data Space (EHDS) Act will outline the requirements of SPE for the secondary use of health data. The final wording should be out before end of November 2024. The guidelines for the SPE implementing acts that will contain more detailed requirements will be released by the end of year 2025.

5.4.3 EOSC projects (SIESTA and TITAN)

Neither SIESTA[56] nor TITAN[57] have released initial versions of their architectures.

6. Results

The following five sections contain the five components of the EOSC-ENTRUST Blueprint and Interoperability Framework (ENTRUST Blueprint): template legal agreements, architecture specifications, operating procedures, interface definitions, and a glossary of

[56] SIESTA https://eosc-siesta.eu/
[57] TITAN https://titanproject.eu/

terminologies. This first version of the ENTRUST Blueprint presents a draft architecture specification and glossary; sections on template legal agreements, operating procedures, and interface definitions are included but intentionally left empty and will be presented in future versions of the ENTRUST Blueprint. We have focused on the architecture and its necessary components, as these will form the basis for interface definitions and operating procedures in future versions of the ENTRUST Blueprint.

6.1 Template legal agreements

To be included in future work in WP14 and WP15

6.2 Architecture specifications

The mapping of selected ENTRUST TRE Provider capabilities (Section 5.2) and Driver requirements (Section 5.3) to the DARE UK Blueprint (Section 5.1) found that it largely aligns with existing Provider capabilities and supports many Driver requirements. Nonetheless, the mapping identified potential critical gaps in the DARE UK Blueprint. The following sections present the ENTRUST Blueprint architecture designed to address these gaps (Figure 6.2.1), and involve Federation Services, Research Analytics Zones, Secure Data Zones, and Index Services.

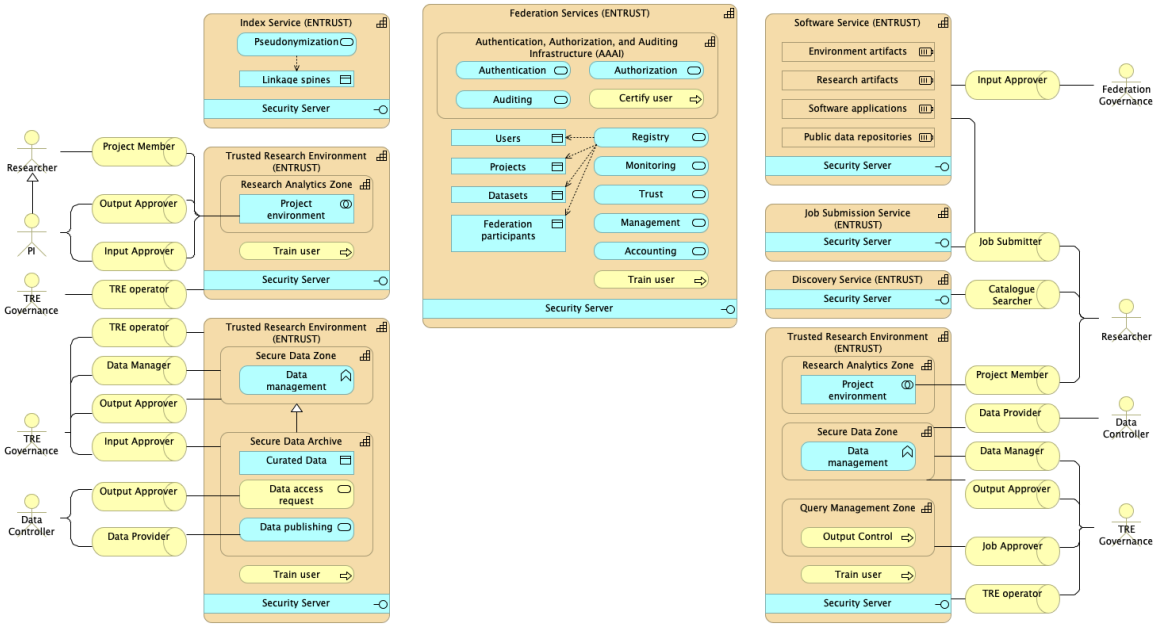

Figure 6.2.1 ENTRUST Infrastructure architecture.

6.2.1 Federation Services

Three of four Drivers identified Safe User training and certification as essential requirements. Whereas established specifications or standards for individual TREs, such as SATRE, do cover user training and certification, a federated network of TREs would require some agreed minimal common training and a common process of certification of its users. Specifically, only certified Researchers should be Project Members and have access to Project environments. We have therefore included a Train user process as part of Federation Services; the accompanying ENTRUST Training package deliverable will define specific training requirements. As certification is closely linked with user authentication and authorization, we have grouped the Certify user process within a dedicated Authentication, Authorization, and Auditing Infrastructure (AAAI) capability in Federation Services (Figure 6.2.2). We note that a process of removing user certification should also be necessary – for example, if a user is found to be violating ethical standards for sensitive data research – but we excluded this process from this first version of the ENTRUST Blueprint. We also note that the Federation Services could be extended with a registry of national and regional legislations and regulations for participating TREs as a potential framework for compliance with legal frameworks and regulations, but leave this as a potential future extension.

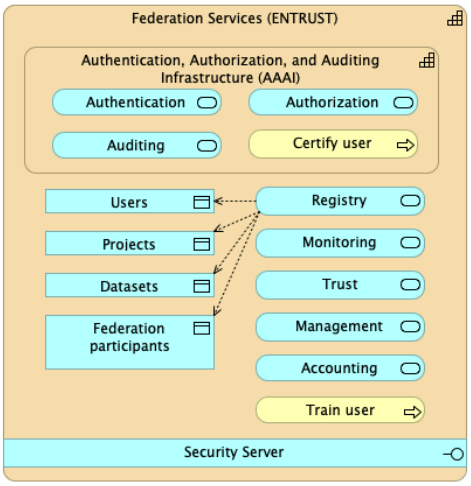

Figure 6.2.2 ENTRUST Federation Services.

6.2.2 Research Analytics Zone

Two of four Drivers identified general user training, including basic training in using the TRE's project space, as an essential requirement. We have therefore included Train user as a required process as part of a TRE (Figure 6.2.3) in addition to the Train user process of the Federation Services. We model TRE training separately from federation training, as the architecture should allow for different TRE participants having different Project environment implementations.

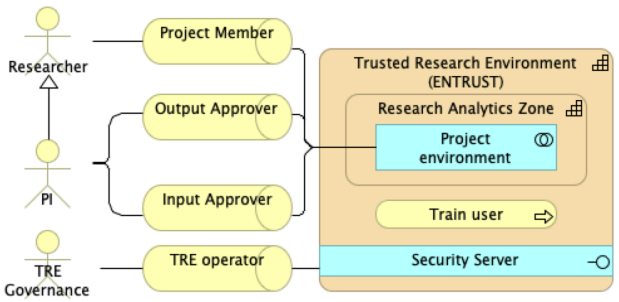

Figure 6.2.3 ENTRUST Research Analytics Zone.

More importantly, three of the Providers, SAFE, TSD, and CSC SD Services, identified a gap in the RAZ architecture. Specifically, their architectures require the PI, which is the Researcher leading a Project, to act as Output Approver for the Project's Project environment. Moreover, PIs have the additional separate role of Input Approver to Project environments. These aspects are included in the ENTRUST RAZ model (Figure 6.2.3).

For TREs that provide RAZ and are not data controllers themselves, there are three main arguments for why the PI and not TRE Governance should be the Output Approver. First, the data controller or its representative grants access to their data for research on specific conditions (i.e. the research project), which must include the named individuals with permission to access the data. The safest approach is to limit this list to the project's Researchers and exclude TRE Governance. Otherwise, TRE Governance must be named in the permit or legally represent the original data controller somehow. Second and third, placing the Output Approver role on TRE Governance when the TRE is not the data controller forms a serious bottleneck that will severely limit infrastructure scalability and raises difficult questions about the legal responsibility of the TRE vs that of the project PI. We do note that the ongoing work on EHDS, secondary use of health data, and the requirements of Secure Processing Environments may clarify some of these aspects. We also note that in our current understanding, RAZ is comparable with the EHDS SPE (see Section 5.4.1).

6.2.3 Secure Data Zone

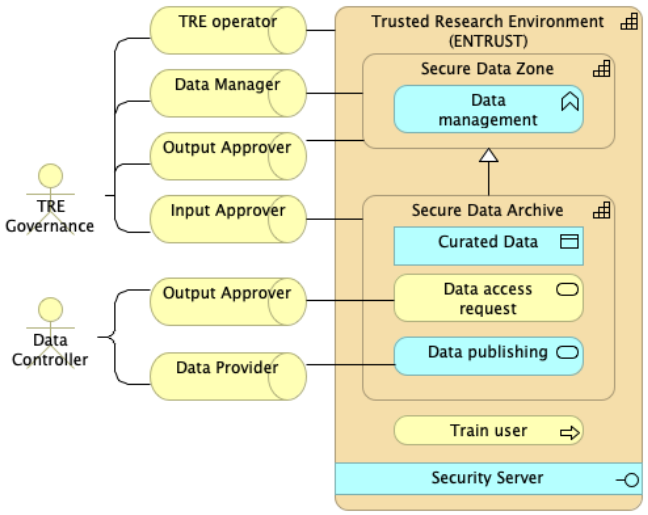

Figure 6.2.4 ENTRUST Secure Data Zone. The Secure Data Archive is a specialised Secure Data Zone storing published Curated Data and handling Data access requests to such published data. Note that the Data Controller would typically have the roles of Data Provider and Output Approver when publishing data and handling data access requests, respectively.

Three of four Drivers identified handling data access requests as essential requirements; two Drivers identified functions related to making data FAIR, including standardising and publishing data, as essential requirements. As both requirements represent specialised optional SDZ functions, we have included these functions as services that are part of a specialised SDZ called a Secure Data Archive (SDA, Figure 6.2.4). Secure Data Archives can be generic or dedicated discipline-specific sensitive data archive implementations, such as the Federated European Genome-Phenome Archive (FEGA). Note that in our modified architecture, SDAs store and provide Curated Data, which for example can be datasets from published studies or dataset freezes from ongoing data curation efforts by appointed Data Custodians. Importantly, Curated Data should have unique IDs and accompanying metadata to enable credit attributions. Note that general SDZs can still store sensitive data as part of ongoing data management or data curation at the SDZ; however, curated data available for research should be published and stored in an SDA.

6.2.4 Index Services

Two of four Drivers identified anonymisation functions as essential requirements. As noted in the mapping, the Drivers may instead require specific pseudonymisation services to enable data reuse involving individual-level merging of datasets (Q2, Q3; see Figure 5.1.1). Specifically, for any given dataset of Curated Data with pseudonymised project-specific IDs, the linkage spine that maps the project-specific IDs to directly identifiable IDs must be available for future reference in an Index Service. Note that the linkage spine is sensitive data in itself and should only be provided for approved linkage requests. Given that each

Index Service has its own registry of linkage spines and also allow for overlap queries, a federation consisting of one or more Index Services can both facilitate Dataset linkage and help answer Discovery Service queries related to overlaps between data stored at different Secure Data Zones in the federation. Importantly, with Project, Dataset, and Linkage spine registries, none of the actual data needs to be linked to answer such overlap queries. We have therefore in the Index Service included a Pseudonymization service that produces and stores project-specific Linkage spines in its own Linkage spines registry (Figure 6.2.5). This Pseudonymization process should be used as part of the process of making project-specific data available in Project environments.

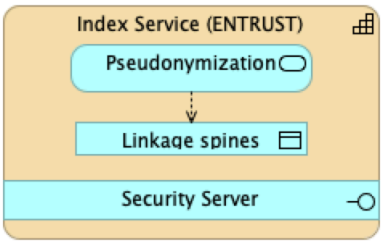

Figure 6.2.5 Index Services, comparing the original DARE UK (left) and ENTRUST (right) versions.

6.3 Operating procedures

To be included in future work in WP14 and WP15.

6.4 Interface definitions

To be included in future work in WP14 and WP15.

6.5 Glossary

See Appendix 1: EOSC-ENTRUST_Architecture_Glossary

7. Discussion

We have presented analyses of the existing interoperability capabilities of selected TRE Providers and the initial combined set of TRE requirements from the four ENTRUST Providers. We framed these analyses as a mapping to the latest version of the DARE UK Federated Architecture Blueprint. The DARE UK Blueprint is designed based on the Five Safes framework to allow research on sensitive data within a network of participating TREs.

Moreover, the DARE UK Blueprint is technology agnostic to accommodate a diverse set of existing TRE implementations and is open and supports FAIR research on sensitive data according to the EOSC principle “open as possible and closed as necessary”. Whereas these design principles align almost perfectly with our goals for the ENTRUST Blueprint, the DARE UK Blueprint was developed in the context of UK legal frameworks and practices regarding sensitive data research. A main goal for mapping ENTRUST TREs to the DARE UK Blueprint was therefore to investigate to what extent the DARE UK Blueprint could support existing TRE architectures across Europe.

To ensure sufficient depth in the analysis, we focused on four TREs from three countries (Norway, Finland, Netherlands) instead of involving all ENTRUST Providers. We will work with the Providers (WP10/WP11) to collect broader data on more TREs for future versions of the ENTRUST Blueprint and ENTRUST TRE Provider Catalogue. Nevertheless, our mapping identified a fundamental difference in the DARE UK Blueprint and the architectures of three of the four TREs. This difference, which is in who (i.e. which actor) approves output from Project environments, likely reflects legal or organisational differences between UK TREs and the Norwegian and Finnish TREs. Many UK TREs hold their own data (i.e. the TRE is the Data Controller of data requested for output approval); in the Finnish and Norwegian TREs, the data owner often is a separate organisation and output control is the responsibility of the PI of the project with permission to analyse the data. This practice is well established for research projects, but we note that the ongoing work in the EHDS on secondary use of health data may introduce other constraints. We have provided an updated version of the DARE UK infrastructure layer that allows for the Norwegian and Finnish architectures.

The mapping of Driver requirements focused on requirements missing from the SATRE framework or other requirements related to interoperability. Most of the Driver requirements mapped directly to capabilities in the DARE UK Blueprint, but we found five requirements to be missing corresponding capabilities in the DARE UK Blueprint. Notably, each of these requirements were listed by at least two Drivers. We note that capabilities for supporting some of these requirements, such as safe user training and certification and general user training have been defined as out of scope for the DARE UK Blueprint. Other requirements, such as data access requests and pseudonymisation may be implicitly supported by other more general capabilities in the DARE UK Blueprint. However, we have chosen to suggest an ENTRUST version based on the DARE UK infrastructure layer where capabilities supporting essential requirements are explicitly modelled.

8. Conclusions

We have described the first version of the architecture specifications of the EOSC ENTRUST Blueprint & Interoperability Framework, built on the DARE UK Federated Architecture

Blueprint. The proposed architecture addresses gaps that we identified through extensive mapping of four TRE provider architectures and the initial combined set of Driver requirements against the DARE UK Blueprint. Our analyses suggest that our current architecture proposal should fit existing European TRE provider architectures and support all Driver requirements while maintaining the Five Safes principles for sensitive data research.

9. Next steps

The current architecture proposal will be evaluated by the Drivers as part of a dedicated Blueprint validation workshop planned for January 21-22, 2025. In the meantime, we will open our proposed architecture for broad comments and suggestions for improvement through a dedicated project in github (or similar). This project page will also be used to establish and refine content for the other components of the ENTRUST Blueprint currently missing from this first version; that is, template legal agreements, operating procedures, and interface definitions. Furthermore, we will work with the ENTRUST work package on trusted researcher identities (WP16) to determine how to best incorporate the AAAI infrastructure capability into the ENTRUST Blueprint. Finally, we will work with the Providers (WP10/WP11) to collect broader data on how Provider TREs map to the ENTRUST Blueprint capabilities. We aim to present and discuss these data on the Requirements and Capabilities workshop planned for May 2025.

10. Impact

This document has delivered the first version of a service blueprint that allows technical interoperability between TRE based on the EOSC Interoperability framework (ENTRUST project objective 2.2). The document is released along with the first versions of the Blueprint training package[58], the Machine readable TRE provider catalogue[59], and the federated analytics demonstrators[60]. Being the first version it is still too early to have objective data on its impact. However, we note that our First EOSC-ENTRUST Evaluation & Adoption Workshop held in Helsinki on September 24-25, 2024, had representatives from TEHDAS2 and from the Genomics Data Infrastructure (GDI) projects, as well as a visitor from the Pawsey Supercomputing Research Center, suggesting that our work is visible and of interest to people outside the ENTRUST project community.

[58] Christine Stansberg, Haneef Awan, Deliverable D13.2 Training package for EOSC-ENTRUST Year one Blueprint & Interoperability Framework [in review].
[59] Miikka Kallberg, Rob Baxter, Stefanie Kirschenmann, Heikki Lehväslaiho, Deliverable D13.3 Machine-readable First Edition of the EOSC-ENTRUST TRE Provider Catalogue [in review].
[60] Philip Quinlan, Laia Codó, Jonathan Couldridge, Eguenio Gonzalo, Stian Soiland-Reyes, Jose Mª Fernandez, Tim Beck, Deliverable D19.1 Deployable Demonstrator of Digital Objects & Workflows Developments