eosc | SIESTA

EOSC-SIESTA

Project overview

Álvaro López García aloga@ifca.unican.es – https://advancedcomputing.ifca.es
IFCA, CSIC-UC

CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

Funded by
the European Union

SIESTA in a nutshell

- HORIZON-INFRA-2023-EOSC-01-06 call
- 5M€, lump-sum scheme
- Duration: 1st Jan 2024 – 31st Dec 2026
- 12 partners (ES, IT, FR, SK, DK, SE, NL)
  - Academic and Research: CSIC, IISAS, INSERM, ISI, CNRS, ULE, SRU, NRU
  - Law: Javier de la Cueva
  - SMEs & Industry: Cendio, interWAY
  - Statistical offices: INE
- https://cordis.europa.eu/project/id/101131957

Objectives

Deliver trusted cloud-based environments for the analysis of sensitive data, built in a reproducible way, and a set of services to ease the secure sharing through state-of-the-art anonymization techniques

1. Enhance the EOSC Exchange services with cloud-based trusted environments for the analysis of sensitive data in the EOSC demonstrating the feasibility of the FAIR principles over them
2. Study, explore and demonstrate the feasibility of FAIR management and processing of sensitive data, showcasing the benefits for society, science and research
3. Deliver tools for the secure anonymization or pseudonymisation of datasets, allowing rightholders to safely release sensitive data through the EOSC Exchange
4. Provide rightholders with best practices and methodologies for the release of sensitive data following FAIR principles
5. Extend the service offer and the capabilities being offered through the EOSC portal, coordinating with the operational and management activities carried out by the EOSC partnership and related projects

SIESTA Concept

- Different access methods based on the data sensitivity: from collaborative development environments (like JupyterHub) for low risk data, to access through security hardened remote desktop solutions with limited capabilities, strict network controls and VPN access for higher sensitivity levels.
- Internal repositories allow the installation of software components and libraries only from trusted sources that have been previously approved.
- User-provided predefined software components that have been endorsed and approved, allowing the offload of user workflow tasks into the platform, accessing sensitive data.
- Assisted anonymization tools for data ingestion and risk-disclosure evaluation tools for data stage out, allowing the improvement of the privacy levels of the shared data.
- A tamper-proof component for keeping track of all relevant transactions, providing auditing mechanisms.
- Integrations with the EOSC Core and Exchange, allowing for instance the inclusion of existing datasets, the generation and storage of anonymized or synthetic data in EOSC compliant repositories or the delivery of trusted and secure thematic data spaces into the EOSC.

SIESTA Concept

- Provide safe and trusted access to sensitive data
- Following tiered model for data sensitiveness
  1. Fully open data. No need to use a trusted research environment.
  2. Very low risk. Pseudonymised data with very low linking risk. Unlikely to cause harm.
  3. Low risk. Strongly pseudonymised datasets with some indirect identifiers.
  4. Average risk. Pseudonymised personal data and confidential organisations information.
  5. High risk. Weak or no de-identification and very sensitive commercial data.
  6. Very high risk. Very sensitive personal data or highly confidential government or commercial data.
- Initially Focus on categories to 2 to 5, with increased level of security, different entry methods, different restrictions
- Infrastructure as Code to ensure reproducibility of the compute environment

Tiered model (data sensitiveness) implications

- Data sensitivity definition is not static, it depends on the context and dynamically defined via policy tooling
  - E.g. depending on data rightholder, audience (who is going to use it),
- Different data sensitiveness (tiered model) coupled with different access models.
  - Provide different access levels: e.g. remote interactive sessions (levels 0-2), remote desktop (level 3), remote desktop with limited capabilities (level 4), execution of trusted code or SMPC (level 5)
- Other platform level security implications
  - e.g. SIESTA audit system has high sensitivity
- SIESTA aims to address these security aspects through an Infrastructure as Code
  - Map platform deployments to certified resource providers (SIESTA involves partners whose compute resources are certified with ISO27001:2017 and National Cybersecurity Schemes)

IFCA CSIC CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS CERTIFIED ISO 9001:2015 COMPANY CERTIFIED ISO 27001:2017 COMPANY ens Esquema Nacional de Seguridad

SIESTA implementation and co-design

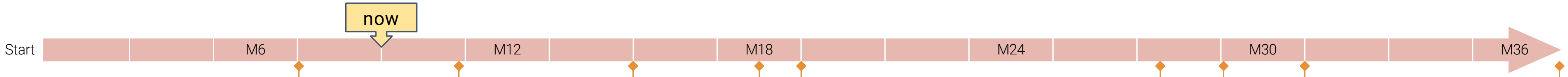
Five complementary cases:

1. Epidemiology: (CSIC, INSERM, ISI) development of an ecosystem with data collected by the SIESTA team from different sources (population, mobility, surveillance, etc.), plus data from collaborative surveillance systems, and with the addition of last generation models able to study propagation patterns of generic infectious diseases.
2. Medical imaging: (SRU, NRU, CNRS) development of (neuro)imaging data analysis pipelines integrated with the EOSC platform that allow expert and non-expert users to carry out analyses on public and non-public neuroimaging (fMRI, EEG, MEG) datasets including demographic, health and questionnaire (tabular) data as covariates.
3. Energy: (-) secured Renewable Energy Community (REC) information hub, able to guarantee a trusted, privacy-compliant, seamless and even cross-border access, reuse and valorization of technical information associated with energy consumption, production and storage.
4. Text anonymization on sensitive data: (ULE) tools that allow anonymizing documents or any information containing text, with special focus on personal data and also on information related to locations, organisations, addresses, emails, finance or any information that could lead to identifying a person or organisation.
5. Demography: (CSIC, INE) tools to improve the anonymization of the data to be shared, the creation of designed populations whose data can be shared without privacy concerns and systems to analyse in a reproducible FAIR way the data without the need of a direct access.

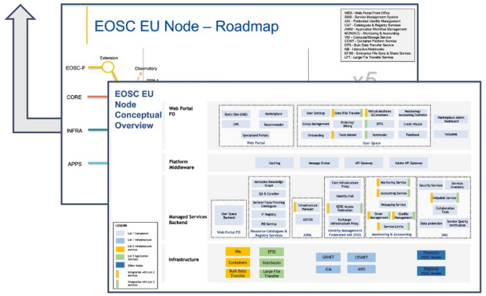
eosc | SIESTA

Results (early) published in SIESTA Zenodo community:
https://zenodo.org/communities/siesta/

Roadmap

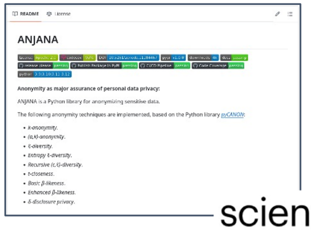


Requirements elicitation
Architecture definition
Platform implementation plans


EOSC EU Node – Roadmap
EOSC-F
CORE
INFRA
APPS
Extension
Observatory
x5
WEB - Web Portal Front Office
SMS - Service Management System
AAI - Federated Identity Management
CAT - Catalogues & Registry Services
AWM - Application Workflow Management
MON/ACC - Monitoring & Accounting
VM - Compute/Storage Service
CONT - Container Platform Service
DTS - Bulk Data Transfer Service
NB - Interactive Notebooks
EFSS - Enterprise File Sync & Share Service
LFT - Large File Transfer Service
EOSC EU Node Conceptual Overview
Web Portal FO
Platform Middleware
Managed Services Backend
Infrastructure
Web Portal
User Space
Caching
Message Broker
API Gateway
Admin API Gateway
Resource Catalogues & Registry Services
AWM
Identity Management Federated AAI (SSO)
Monitoring & Accounting
SMS
VMs
Containers
Bulk Data Transfer
EFSS
Notebooks
Large File Transfer
GRNET
CESNET
ICM
AWS
Thematic EOSC Nodes
Regional EOSC Nodes
LEGEND

User workshop
All hands meeting

Early release (preview) of anonymisation tools
(common privacy models, differential privacy, etc.)
https://github.com/SIESTA-eu/anjana



ANJANA

Anonymity as major assurance of personal data privacy:

ANJANA is a Python library for anonymizing sensitive data.

The following anonymity techniques are implemented, based on the Python library pyCANON:

- *k-anonymity.*
- *(α,k)-anonymity.*
- *ℓ-diversity.*
- *Entropy ℓ-diversity.*
- *Recursive (c,ℓ)-diversity.*
- *t-closeness.*
- *Basic β-likeness.*
- *Enhanced β-likeness.*
- *δ-disclosure privacy.*

scientific **data**

Tension between FAIR guiding principles
and sensitive data: report

1st platform release
EOSC SIESTA anonymisation tools

EOSC Integration

SIESTA risk
control tools

External user workshop

EOSC DevSecOps
practises and processes

2nd platform release
User uptake and validation

Agile project development (Personas, Epics, user stories, requirements) towards first platform prototype

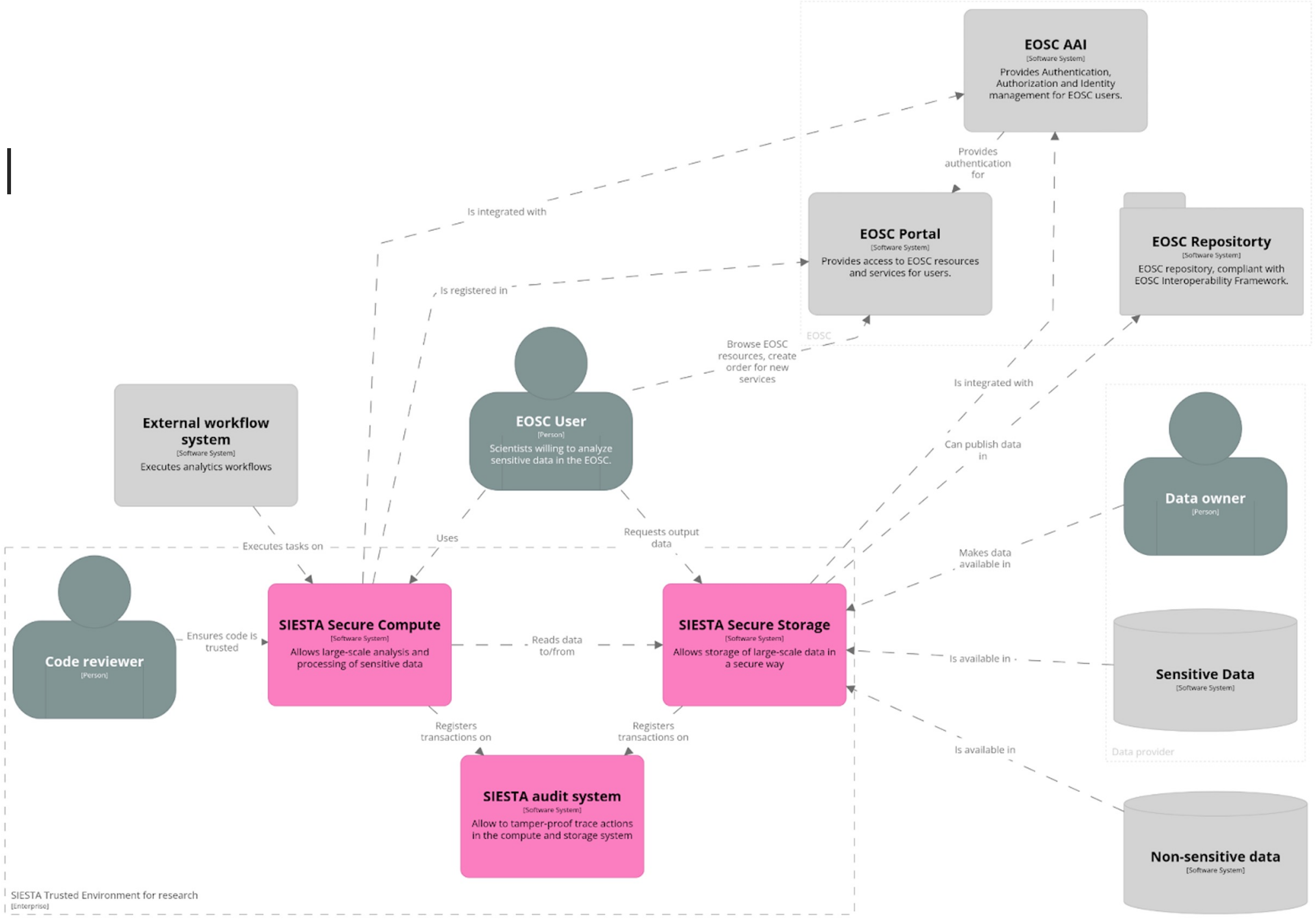
Foster platform and tools usage and use case uptake
Good practices and guidelines
Dissemination and KPI maximization, KER (re)definition

Initial high level architecture

EOSC-SIESTA follows the C4 model and notation for its architecture definition

Code and static diagrams:
https://zenodo.org/records/13347733

Online diagrams:
https://structurizr.com/share/94201/c50c8b32-ebb1-4963-a4f0-51193eee0fba



EOSC AAI
[Software System]
Provides Authentication, Authorization and Identity management for EOSC users.

Provides authentication for

EOSC Portal
[Software System]
Provides access to EOSC resources and services for users.

EOSC Repository
[Software System]
EOSC repository, compliant with EOSC Interoperability Framework.

EOSC

Is integrated with

Is registered in

Browse EOSC resources, create order for new services

Is integrated with

Can publish data in

External workflow system
[Software System]
Executes analytics workflows

EOSC User
[Person]
Scientists willing to analyze sensitive data in the EOSC.

Data owner
[Person]

Executes tasks on

Uses

Requests output data

Makes data available in

Code reviewer
[Person]

Ensures code is trusted

SIESTA Secure Compute
[Software System]
Allows large-scale analysis and processing of sensitive data

Reads data to/from

SIESTA Secure Storage
[Software System]
Allows storage of large-scale data in a secure way

Is available in

Sensitive Data
[Software System]

Data provider

Registers transactions on

Registers transactions on

Is available in

SIESTA audit system
[Software System]
Allow to tamper-proof trace actions in the compute and storage system

Non-sensitive data
[Software System]

SIESTA Trusted Environment for research
[Enterprise]

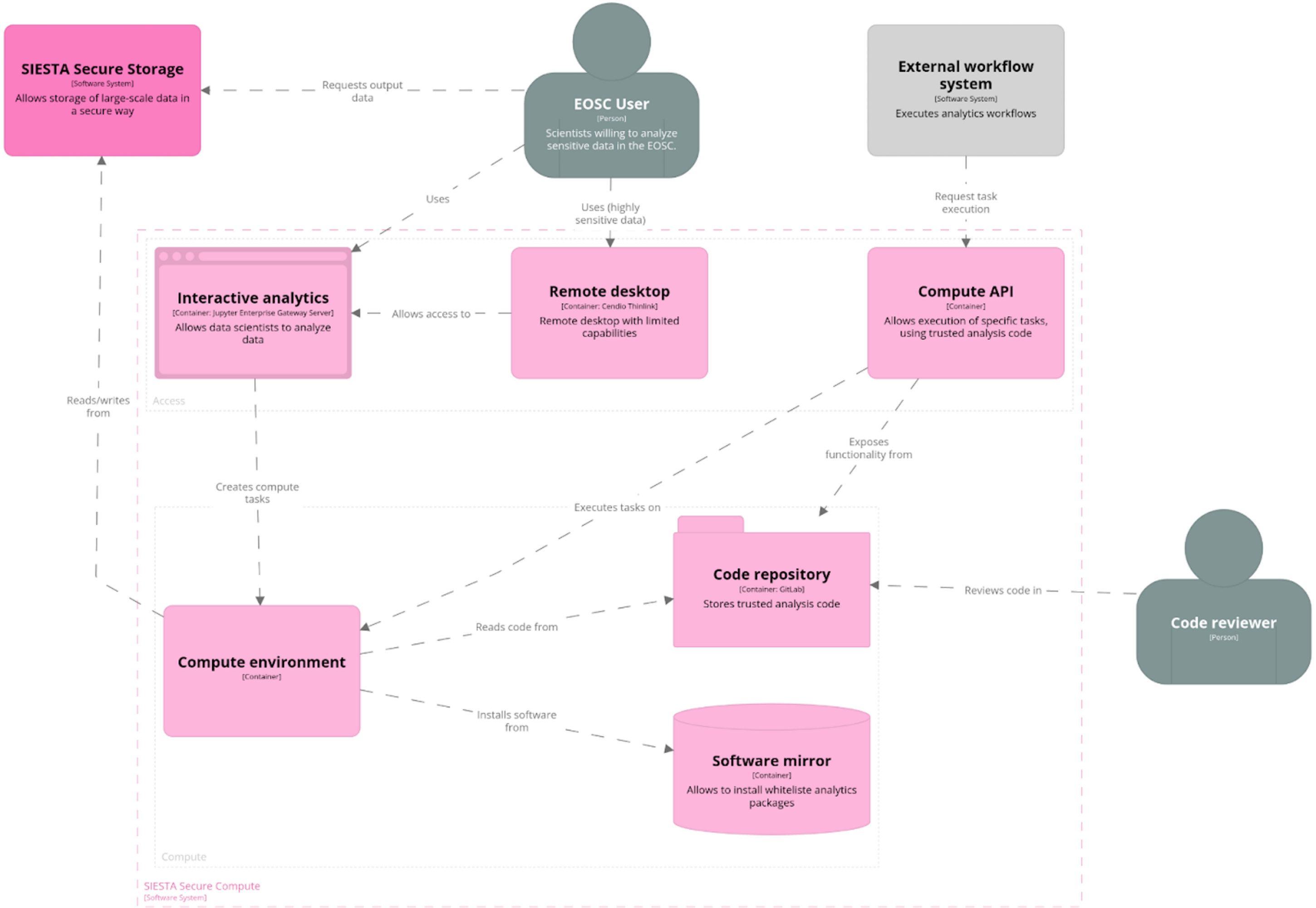
[System Landscape] SIESTA Trusted Environment for research
Tuesday, March 7, 2023 at 1:00 PM Central European Standard Time

Initial high level architecture

EOSC-SIESTA follows the C4 model and notation for its architecture definition

Code and static diagrams: https://zenodo.org/records/13347733

Online diagrams: https://structurizr.com/share/94201/c50c8b32-ebb1-4963-a4f0-51193eee0fba



SIESTA Secure Storage
[Software System]
Allows storage of large-scale data in a secure way

EOSC User
[Person]
Scientists willing to analyze sensitive data in the EOSC.

External workflow system
[Software System]
Executes analytics workflows

Requests output data

Uses

Uses (highly sensitive data)

Request task execution

Interactive analytics
[Container: Jupyter Enterprise Gateway Server]
Allows data scientists to analyze data

Allows access to

Remote desktop
[Container: Cendio Thinlinc]
Remote desktop with limited capabilities

Compute API
[Container]
Allows execution of specific tasks, using trusted analysis code

Reads/writes from

Access

Creates compute tasks

Exposes functionality from

Executes tasks on

Code repository
[Container: GitLab]
Stores trusted analysis code

Reviews code in

Code reviewer
[Person]

Compute environment
[Container]

Reads code from

Installs software from

Software mirror
[Container]
Allows to install whiteliste analytics packages

Compute

SIESTA Secure Compute
[Software System]

[Container] SIESTA Secure Compute

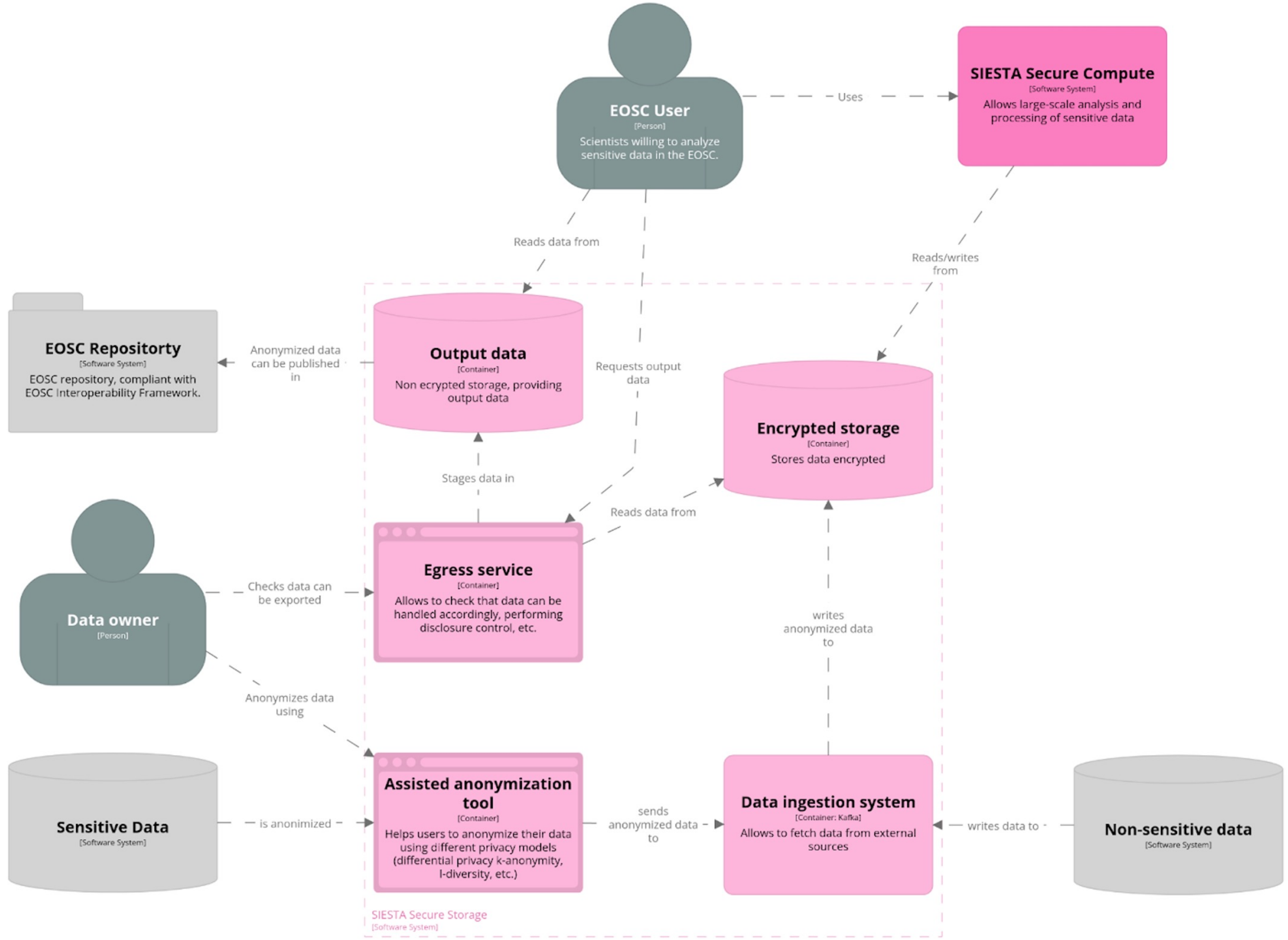
Tuesday, March 7, 2023 at 1:00 PM Central European Standard Time

Initial high level architecture

EOSC-SIESTA follows the C4 model and notation for its architecture definition

Code and static diagrams: https://zenodo.org/records/13347733

Online diagrams: https://structurizr.com/share/94201/c50c8b32-ebb1-4963-a4f0-51193eee0fba



EOSC User
[Person]
Scientists willing to analyze sensitive data in the EOSC.

Uses

SIESTA Secure Compute
[Software System]
Allows large-scale analysis and processing of sensitive data

Reads data from

Reads/writes from

EOSC Repositorty
[Software System]
EOSC repository, compliant with EOSC Interoperability Framework.

Anonymized data can be published in

Output data
[Container]
Non ecrypted storage, providing output data

Requests output data

Encrypted storage
[Container]
Stores data encrypted

Stages data in

Reads data from

Data owner
[Person]

Checks data can be exported

Egress service
[Container]
Allows to check that data can be handled accordingly, performing disclosure control, etc.

writes anonymized data to

Anonymizes data using

Sensitive Data
[Software System]

is anonimized

Assisted anonymization tool
[Container]
Helps users to anonymize their data using different privacy models (differential privacy k-anonymity, l-diversity, etc.)

sends anonymized data to

Data ingestion system
[Container: Kafka]
Allows to fetch data from external sources

writes data to

Non-sensitive data
[Software System]

SIESTA Secure Storage
[Software System]

[Container] SIESTA Secure Storage

Tuesday, March 7, 2023 at 1:00 PM Central European Standard Time

Exploring collaborations

INFRAEOSC and beyond

- Ongoing collaboration with sister projects EOSC-ENTRUST and TITAN
  - EOSC Symposium unconference session
    "Open as possible, restricted as necessary; EOSC sensitive data exchange"
    https://indico.cern.ch/event/1408259/timetable/#b-563302-unconference-open-as
- Identified potential synergies in INFRAEOSC context
  - AI4EOSC
    - Leveraging AI4EOSC federated learning platform to develop AI/ML models over sensitive data
    - Close collaborations with Flower.ai: Code contributions, participation in pilot programmes, MONAI FL and NVIDIA FLARE: Model compatibility
  - RAISE
    - (exploring) work together on analysis of sensitive data, synthetic and exemplary datasets, etc.
- Potential further collaborations outside EOSC:
  - University of Cantabria:
    - Ad-hoc privacy preserving algorithms (homomorphic encryption, SMPC)
  - EUCAIM (Cancer Image Federation) project funded by DIGITAL:
    - Synergies to be explored: de-identification, distributed analysis of sensitive data,..)

eosc | ENTRUST

eosc | TITAN

Funded by the European Union

AI4 | eosc

eosc | RAISE
Research Analysis Identifier System

MONAI+ Flower

EUCAIM
CANCER IMAGE EUROPE

eosc | SIESTA

Project coordination:

- siesta-po@listas.csic.es

Thank you for your attention

Funded by
the European Union