eosc | SIESTA

SIESTA : Cloud confidentiel pour la recherche

Vincent Legoll

vincent.legoll@iphc.cnrs.fr

JCAD – septembre 2025

cnrs

IPHC Institut Pluridisciplinaire Hubert CURIEN STRASBOURG

Université de Strasbourg

Funded by the European Union

Cloud confidentiel pour la recherche

SIESTA :
Secure **I**nteractive **E**nvironments for **S**ensi**T**ive data **A**nalytics

https://eosc-siesta.eu

Traduction :
Environnements **H**yper-sécurisés et interactifs **P**our l'**A**nalyse de **D**onnées sensibles

Projet Européen EOSC: European Open Science Cloud

- Budget total : ~5M€
- Durée : 36 mois

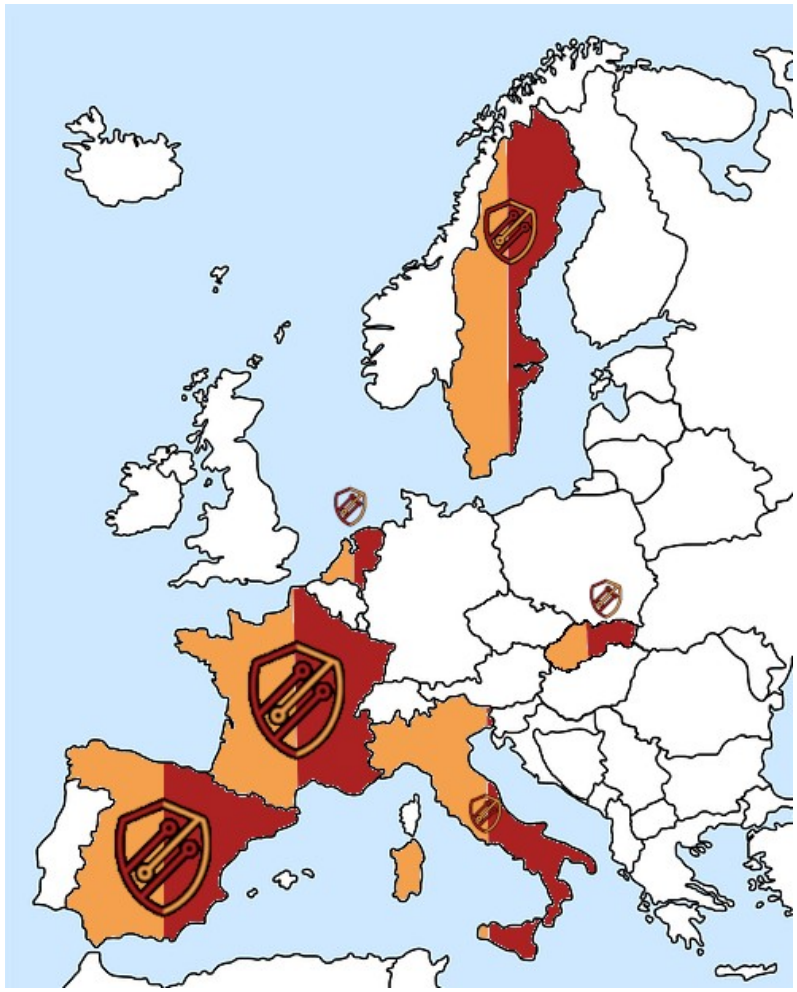
Partenaires

12 partenaires :

- Institutions de recherche
- Entreprises (service, logiciel, avocats)

6 pays :

- **Espagne**
  - CSIC – IFCA – IFISC - UIB
  - INE - Instituto Nacional de Estadística
  - JdCA – Javier de la Cueva (cabinet d'avocats)
  - Predictia – Société de R&D (spinoff University of Cantabria)
  - ULE - Universidad de León
- **France**
  - CNRS
  - INSERM
- **Hollande**
  - NRU/SRU - Donders Institute for Brain, Cognition, and Behaviour - Radboud University
- **Italie**
  - ISI - International Statistical Institute
- **Slovaquie**
  - IISAS - Institute of Informatics Slovak Academy of Sciences
  - IWAY - Interway (Société de Services)
- **Suède**
  - CEN - Cendio (ThinLinc, services et FOSS)



Introduction

- Les données sensibles sont partout
- Ces données intéressent la recherche
- Protégées par des législations diverses
- Sécurité : stockage, transferts, et pendant l’utilisation ?
- Cloud => Machines partagées => risques de fuites
- Accès par les opérateurs de l’infrastructure
- Cloud Confidentiel => les données restent cryptées en mémoire
- Système d’enclaves sécurisées inaccessibles aux tiers
- Technologies jeunes => matériel récent => logiciel récent

Législations & réglementations

- Globales (EU GDPR)
- Locales (par pays, par détenteur de jeux de données)
- Différences
- Peu permissives (interdictions)
- Bureaucratie

SIESTA : Centralisation, simplification

- Tiers de confiance reconnu et éprouvé
- Signataire interposé
- Procédures allégées

Objectif données sensibles

1) Fournir des environnements cloud de confiance

2) Étudier la faisabilité des principes FAIR

3) Fournir des outils d'anonymisation

4) Accompagner les utilisateurs pour appliquer les principes FAIR

5) Intégrer la plateforme à « EOSC EU node »

Cas d’usage

- Épidémiologie (population, déplacements)
- Imagerie médical
- Énergie (population, consommation, habitat)
- Anonymisation de données textuelles (documents d'investigation criminelle, type CERT)
- Démographie (population, habitat, données financières)

La SIEST'Tech

- AAI : Authentification et Autorisation Infrastructure
  - Intégration EOSC (Keycloack)

- Traitement, pipelines de calcul
  - VMs (OpenStack)
  - Conteneurs (Kubernetes)

- Stockage, transferts
  - Frontend : NextCloud, iRods, …
  - Backend : Volumes + cryptage **logiciel** (dans l'enclave sécurisée) (OpenStack cinder + LUKS)

- Sécurité logicielle : provenance, signatures, attestation, SBOM (Harbor, Trivy)

- Développement et suivi : IaC, tickets (Gitlab, ansible, kubespray, confluence)

La SIEST'Tech - suite

- (pseudo-)anonymisation (ANJANA, pyCANON, BIDScramble)
- Sécurité et observation réseau (Cilium)
- Monitoring et audit, journaux inaltérables (Wazuh, Trilian)
- Gestion & distribution de secrets (OpenStack Barbican)
- Expérience & interfaces utilisateur
  - Diversité : API REST, CLI, portail web, interactive (ThinLinc)
  - non-experts (compétences techniques requises => délégation)

Réalisations

3 Clouds :

- IFCA – CSIC – Espagne
- IISAS – Slovaquie
- IPHC – France

- Problèmes rencontrés
  - Pile logicielle très récente requise (Noyau Linux, QEmu, OpenStack, K8s)
    - Difficile à mettre en place sur nos infrastructures cloud de production => PoCs séparés
  - Manque de compétences techniques => besoin de simplicité :
    - CLI (quasi) impossible
    - Compromis : délégation de responsabilité

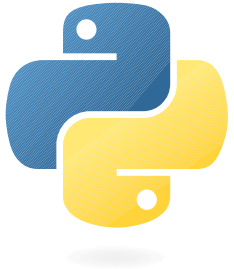
Note : le projet n’est qu’à mi-parcours

Réalisations - suite

Développements logiciels :

- pyCANON : outils python d'analyse d'anonymat de jeux de données tabulaires
  - https://github.com/IFCA-Advanced-Computing/pycanon
- anjana : outils d'anonymisation/pseudonymisation
  - https://github.com/IFCA-Advanced-Computing/anjana
- BidScramble : outils d'anonymisation de données en neuro-imagerie au format BIDS
  - https://github.com/Donders-Institute/bidscramble

- 100 % Python
- FOSS (Apache 2 & GPL3)



Comment faire le FAIR ?

- FAIR n’est pas incompatible avec données sensibles

- Les métadonnées ouvertes permettent :
  - **F**indable
    - Catalogue des jeux de (méta-)données
    - Moteur de recherche
  - **A**ccessible
    - Ne veut pas dire « ouvert à tout public »
    - Les règles et procédures doivent être documentées, accessibles
  - **I**nteroperable
    - Formats ouverts
    - Standards
    - Taxonomies
  - **R**eusable
    - durée de vie gérée, connue
    - Stockage pérenne

Ouverture

EGI / EOSC : d'autres projets

- EOSC ENTRUST & TITAN: des projets concurrents
- EUCAIM : EU Cancer Image (thématique)

- Et dans le privé ?

  Tous (quasi) les fournisseurs majeurs de ressources cloud ont une offre CC, mais on en connaît le(s) problème(s) : Coût, souveraineté, etc.

Conclusions

- Le besoin est important
- Les principes FAIR ne sont pas un obstacle
- Encore du travail à fournir
- Simplicité d’utilisation : un enjeu majeur

Questions

???

P.S. : L’équipe SCIGNE est à votre disposition jusqu’à mercredi...

Voilà, c’est fini...

Il ne reste plus qu’une chose à **FAIR** :

une petite **SIESTA** !