



OSSDIP: Open Source Secure Data Infrastructure and Processes

Andreas Rauber¹ Martin Weise¹
Martina Landman¹ Moritz Staudinger¹
Cornelia Michlitz¹

¹ TU Wien, Austria

Motivation

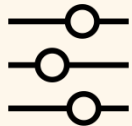


Sensitive data (e.g. personally identifiable information) handling is inevitable in many projects, ensure it is never leaked or misused.



Provide Access for Analysis

Ensure proper usage of data and minimize risk while also allowing maximum usage of data



Data Owner maintains Full Control

Who can access, over which period, which subset of data, answer which research questions and activities



Fingerprinting

Last perimeter of defense, know who leaked the data, track the leak



Data Visiting ≠ Data Sharing

Invite others to come to the data instead of giving away data, once shared it can be considered as gone



Legal + Technical Mechanisms

Processing Agreement, Access Agreement, Monitoring Agreement, Analysis Agreement, NDA



Extensive Monitoring

Activities, input, output, screenshots, video stream, data stream



Project aims and goals



Main Goal: Meeting the conflicting goals of protecting and maintaining control over sensitive data while also allowing access by third parties

Automating processes for data provisioning and safe returns

Ingress and egress of data, automated provisioning on Analyst-VM, request for result export

Monitoring and preserving all actions

Append-only logs (software), in the future write-once technology, recording video stream for forensics

Automated export of metadata supporting FAIR principles

Landing page (portal) information supports findability and accessibility

Analytics and workflow management tools, evaluate with pilot adopters

Expand supported software, pre-configure, optionally Docker integration during access request

Support small number of pilot adopters in the life sciences, interoperability

Requirements for process automation, tools, coordinating interoperability, training material



Project scientific and technical background

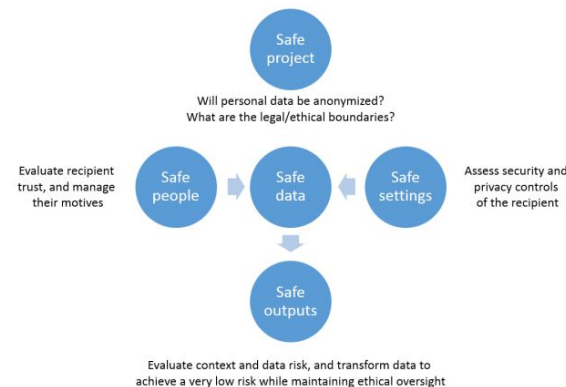


Building on top of the experience gathered from operating a very similar infrastructure in the health care sector for almost a decade.

Deadlock of sharing sensitive data vs. privacy concerns

This challenge has become acute during the early days of the COVID-19 pandemic when evidence-based decision making required the analysis and sometimes integration of highly sensitive (due to privacy or commercial reasons) information such as health data, social science data, movement data from telecommunications operators, or supply-chain logistics data from the retail sector.

But even outside this exceptional situation, academia-industry collaborations as well as industry-to-industry co-operations frequently are hindered by the conflicting needs to keep the data secret that the other party should process or analyze.



Project scientific and technical background

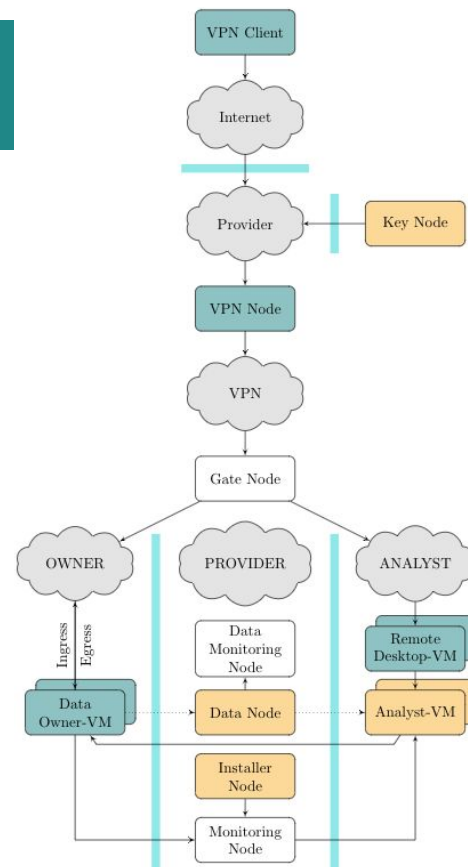


The technology stack includes automation engines, virtual machines, containers, fingerprinting, screen recording and scanning

Our solution: Data Visiting

The Open Source Secure Data Infrastructure (OSSDIP) is centered around the principle of data visiting: data is never passed on to researchers, but researchers visit isolated and completely shielded compute environments (virtual machines in a shielded network) via remote desktop.

All interactions are strictly monitored, the desktop screen is recorded with subsequent share to the data owner and export of data is blocked. The infrastructure never provides direct access to the air-gapped Data Node, instead, data subsets are copied and fingerprinted at the Analyst-VM along with the analysis tools required to directly work with the data set.



Current status and plans of data resource and workflows



Versioned, cite-able and reproducible access provision

- Concept is applicable (proven in non-sensitive settings) for semi-structured data
- All software used and deployed under open.source license (can be replaced by commercial tools)
- *Metadata will be published following FAIR guidelines (e.g. RDA-WGDC recommendations) via ECRIN*

Relational data provided as CSV file or SQL database

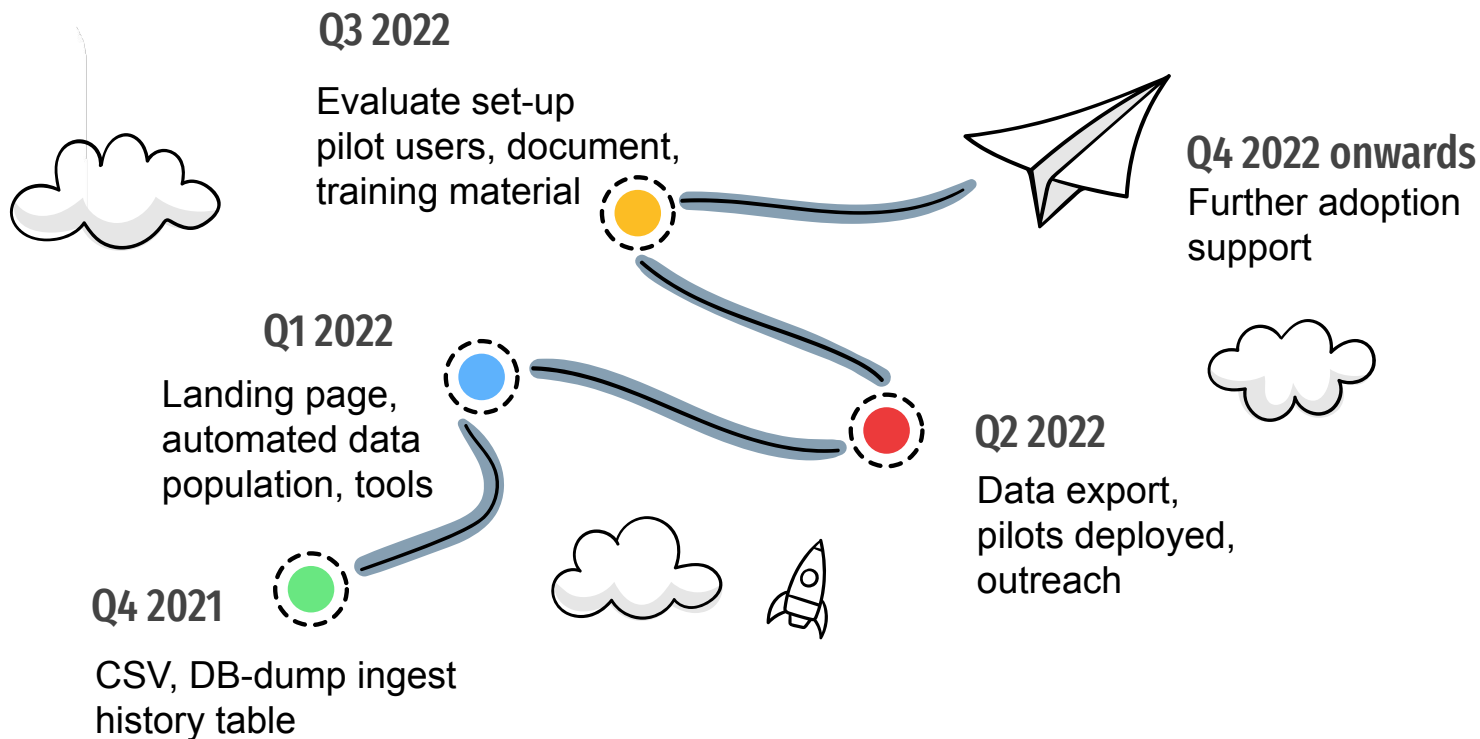
- Processes for data provisioning and data access requests defined
- Pre-defined Analyst-VMs can be automatically deployed
- *Configuration of more analysis tools that are used in the life sciences*
- *Providing pre-packaged tool sets*

Reduce effort to ingress data, maximize utility to researchers

- Process of *safe return* is defined but supported via mostly manual processing steps, *automate*
- *Automate access request process for researchers*
- *Provide better analytics support and increase flexibility (via diverse tools)*



Workplan for implementation including timelines



Input needed from EOSC-Life WP's?



Pilot users and adopters

Who?

- ... **has a pilot setting** that needs such a solution?
- ... is interested in jointly **setting up a test instance** to explore the solution, identify limitations, and suggest improvements?
- (...would like to be in **contact** to follow the development, provide feedback, establish contact to other potentially interested pilot partners?)

What?

- ... **data** do you have that could be used for the pilots (structured, preferably not in the peta-byte range, from simple to more complex)?
- ... (OS) **tools** would you like to use on the analyst machines?
- ... limitations do you see of the current approach that should be addressed?
- ... is the current approach to this problem, and how much “better” will the proposed solution need to be to make it a **viable alternative**?





Open questions?



This project has received funding from the European Union's Horizon
2020 research and innovation programme under grant agreement No 824087