



Deliverable D8.8

Evaluation of distributed analysis and federated learning infrastructure solutions and recommendations for adoption

Project Title Grant agreement no	Genomic Data Infrastructure Grant agreement 101081813		
Project Acronym (EC Call)	GDI		
WP No & Title	WP8: Application and Innovation Solutions		
WP Leaders	Alfonso Valencia (BSC), Salvador Capella-Gutierrez (BSC), Marc Van Den Bulcke (SC)		
Deliverable Lead Beneficiary	BSC		
Contractual delivery date	31/10/2023	Actual delivery date	26/03/2024
Delayed	Yes		
Partner(s) contributing to deliverable	BSC, VIB, DKFZ, SC		
Authors	Carles Hernandez-Ferrer (BSC) Dilza Campos (VIB) Laura Portell-Silva (BSC) Salvador Capella (BSC)		
Contributors	Sergi Aguiló (BSC), José M ^a Fernández (BSC)		
Reviewers	Luiz Gadelha (DKFZ) Emilie Cauët (SC)		



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.



Log of changes

Date	Mvm	Who	Description
21/12/2023	0v1	Laura Portell-Silva (BSC)	First draft sent to WP8
15/02/2023	0v2	Carles Hernandez-Ferrer (BSC)	Sent to reviewers
27/02/2024	0v3	Carles Hernandez-Ferrer (BSC)	Updated according reviewers and sent to Coordination
26/03/2024	1v)	Mercedes Rothschild Steiner (ELIXIR Hub)	Final version submitted to EC Portal

Contents

1. Executive Summary.....	3
2. Contribution towards project outcomes.....	4
3. Methods.....	6
4. Description of work accomplished.....	7
4.1 Distributed analysis (PRS) for infectious diseases.....	7
4.2 Cancer data mobilisation across analysis platform.....	8
4.3 5-safes RO-Crate compatible Trusted Research Environment.....	9
5. Results.....	10
6. Discussion.....	10
6.1 Distributed analysis (PRS) for infectious diseases.....	10
6.2 Cancer data mobilisation across analysis platform.....	11
6.3 5-safes RO-Crate compatible Trusted Research Environment.....	11
7. Next steps.....	12
7.1 Distributed/federated analysis (PRS) for infectious diseases.....	12
7.2 Cancer data mobilisation across analysis platform.....	12
7.3 5-safes RO-Crate compatible Trusted Research Environment.....	12
8. Conclusions & Impact.....	13
8.1 Advancements in Distributed and Federated Analysis.....	13
8.2 Cancer Data Mobilization and Integration.....	13
8.3 Advancements in Trusted Research Environments.....	14
8.4 Overall Impact and Future Directions.....	14





1. Executive Summary

The completion of the first iteration of the Global Data Initiative technical demonstrators marks a significant milestone in collaborative efforts defining the next level of technologies that will be incorporated into the project. Three key demonstrators showcased advancements in distributed and federated analysis, cancer data mobilisation, and trusted research environments.

Advancements in Distributed and Federated Analysis: The Polygenic Risk Score (PRS) analysis of infectious diseases demonstrated the potential of collaborative frameworks in leveraging disparate datasets while preserving data privacy. Challenges encountered with platforms like Galaxy underscore the need for evolving technologies to facilitate federated analysis seamlessly, which were identified and will be tested on the next iteration of the demonstrator.

Cancer Data Mobilization and Integration: Utilising Galaxy, cBioPortal, and Beacon (v2) highlighted the importance of interoperable platforms in accelerating translational research. Ongoing discussions about platform performance and clinical data harmonisation emphasise the need for interdisciplinary collaboration in refining data management protocols.

Advancements in Trusted Research Environments: The development of a 5-safes RO-Crate compatible Trusted Research Environment laid the foundation for establishing secure and reproducible research infrastructures. Challenges remain in technical capabilities for workflow reproducibility and reutilization.

The successful execution of the three demonstrators underscores multiple collective efforts from different research and knowledge areas. While multiple obstacles were found, strategies to overcome them were found and applied when required and next steps were defined to achieve the goal of each demonstrator.





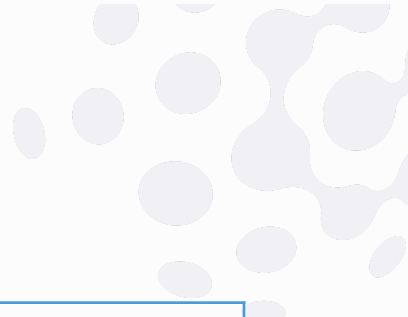
2. Contribution towards project outcomes

With this deliverable, the project has reached or the deliverable has contributed to the following project outcomes:

	Contributed
<p>Outcome 1</p> <p>Secure federated infrastructure and data governance needed to enable sustainable and secure cross border linkage of genomic data sets in compliance with the relevant and agreed legal, ethical, quality and interoperability requirements and standards based on the progress achieved by the 1+MG initiative.</p>	Yes
<p>Outcome 2</p> <p>Platform performing distributed analysis of genetic/genomic data and any linked clinical/phenotypic information; it should be based on the principle of federated access to data sources, include a federated/multi party authorisation and authentication system, and enable application of appropriate secure multi-party and/or high-end computing, AI and simulation techniques and resources.</p>	Yes
<p>Outcome 3</p> <p>Clear description of the roles and responsibilities related to personal data and privacy protection, for humans and computers, applicable during project lifetime and after its finalisation.</p>	No
<p>Outcome 4</p> <p>Business model including an uptake strategy explaining the motivation, patient incentives and conditions for all stakeholders at the different levels (national, European, global) to support the GDI towards its sustainability, including data controllers, patients, citizens, data users, service providers (e.g., IT and biotech companies), healthcare systems and public authorities at large.</p>	No
<p>Outcome 5</p> <p>Sustained coordination mechanism for the GDI and for the GoE multi-country project launched in the context of the 1+MG initiative.</p>	No



GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.



<p>Outcome 6</p> <p>Communication strategy – to be designed and implemented at the European and national levels.</p>	<p>No</p>
<p>Outcome 7</p> <p>Capacity building measures necessary to ensure the establishment, sustainable operation, and successful uptake of the infrastructure.</p>	<p>No</p>
<p>Outcome 8</p> <p>Financial support to the relevant stakeholders to enable extension, upgrade, creation and/or physical connection of further data sources beyond the project consortium or to implement the communication strategy and for capacity-building.</p>	<p>No</p>





3. Methods

The Genomics Data Infrastructure (GDI) represents a federated system necessitating robust infrastructure for either distributed or federated analysis of the data hosted by its nodes. In pursuit of this objective, various methodologies have been explored for GDI, with current considerations encompassing the following:

- Distributed Analysis refers to the process of analysing data that is distributed across multiple locations or nodes, without necessarily involving collaborative learning or model training. The main goal of distributed analysis is to leverage the computational resources available at each location to perform data analysis tasks. The key difference between distributed and federated analysis is that in this case, the data is the same everywhere and that the main reason for the computation can be leveraged across the sites.
- Federated Analysis, on the other hand, refers to the process of analysing data that is distributed across multiple locations or nodes in a privacy-preserving manner. The main goal of federated analysis is to enable collaborative analysis of sensitive data, specific and local to each node, without compromising data privacy or security.

To identify specific methods within GDI aligning with either distributed or federated analysis, a workshop¹ was conducted in Barcelona on October 11, 2023. During this event, diverse partners presented their technologies, showcasing solutions that could contribute to the federated or distributed analysis goals of the infrastructure. The technologies included:

- Galaxy Europe is a powerful and collaborative platform designed for bioinformatics and computational biology research across the European region. It serves as an open-access, web-based framework that enables scientists to analyse and interpret complex biological data seamlessly. Galaxy Europe facilitates distributed analysis allowing researchers to access and share tools, workflows, and datasets while maintaining science reproducibility and transparency.
- The TRE-FX is a conceptual infrastructure based on a federated architecture of multiple Trusted Research Environments (TREs), secure locations in which data is placed for researchers to analyse. The TRE-FX is composed of four basic and interconnected components:
 - The submission layer that is the entry point to the infrastructure, where users can submit queries to the safe data.
 - The controller that sits inside the TRE and decides if a query (or a workflow) can be run or has to be rejected, as well as if results should be returned.

¹ [20231011.JointWorkshop.FederatedAnalysis.Agenda](#)



- The workflow executor that is in charge of running the query (workflows) once they are accepted by the controller.
- The transparency layer ensures minimal and basic information is made public for each query (run workflow) in order to support FAIR reproducible research.

It supports both power users capable of writing and submitting RO-Crates as well as software vendors who can generate the crates on the fly.

- The Workflow Execution Service (WfExS) backend engine aims to fetch a workflow from a TRS-enabled WorkflowHub instance, fetch the inputs and workflow execution engine, and execute the workflow in a secure way. It is compatible with RO-Crate and GA4GH cloud workstream specifications (TRS/WES/TES).

A second workshop² was organised in Barcelona on January 22-24, 2024 in order to align the expectations and responsibilities between Pillar II and Pillar III, as well as to evaluate the output of the current federated analysis technological demonstrators and define the next steps for them.

4. Description of work accomplished

The work done in distributed/federated analysis was divided into three different technological demonstrators, encompassing multiple technologies representing different problems in the scientific community. These demonstrators are:

- Distributed/ analysis (PRS) for infectious diseases
- Cancer data mobilisation across the analysis platform
- 5-safes RO-Crate compatible Trusted Research Environment

4.1 Distributed analysis (PRS) for infectious diseases

For this demonstrator, a multisample vcf file from a synthetic cohort (EGA dataset EGAD00001006673) was divided in three, to simulate three different cohorts in different countries. Each country had its own group of sequenced patients and a table with the respective random outcomes (if the individual was infected with Covid-19 and the severity of the disease) as well as two covariates (sex and smoking status). A researcher in a country wanted to test the association of specific genetic variants with prognosis (i.e., severity), taking into account possible confounders (e.g., sex and smoking status).

The three countries initially part of this demonstrator are Belgium, Germany and Spain. Each country has a tilt of the data, representing a local dataset whose privacy has to be secured.

² [20240122-24.GDI.PillarIII+II.Workshop.F2F](#)



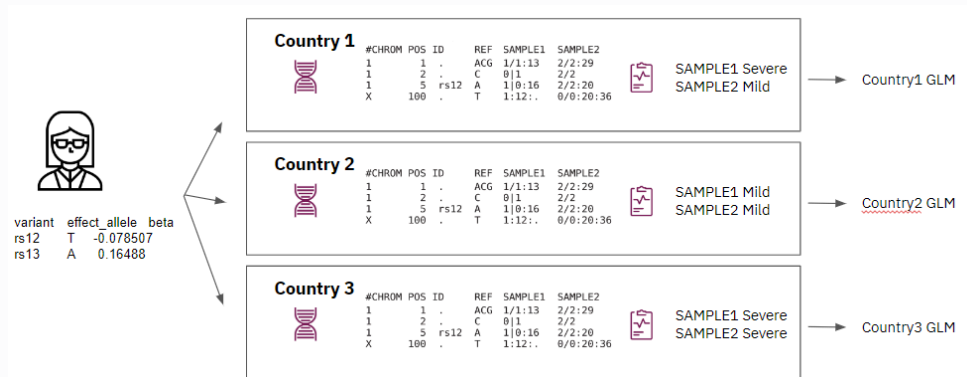


Figure 1. Description of the conceptual orchestration of the demonstrator calculating a PRS score using a distributed architecture within infectious diseases

The demonstrator was built to execute the PRS calculation on a local Galaxy node, in order to preserve data-privacy. The result, three generalised linear models (GLMs), was obtained for each one of the sites. In the ideal case, individual scores are not returned, as it would reveal the risk of an individual versus a certain disease. However an aggregated or pooled GLM to assess relationships between phenotypes and PRS scores would be the right output.

4.2 Cancer data mobilisation across analysis platform

This demonstrator makes use of multiple technologies, encompassing Galaxy, cBioPortal³ (a resource for interactive exploration of multidimensional cancer genomics data sets) and Beacon⁴ (a protocol that allows for data discovery of genomic and phenoclinic data). Its goal is to process a raw dataset from cancer (FastQ files) in a way that allows for discoverability and interrogation. The data used in this demonstrator is a synthetic dataset on colorectal cancer, created at the BSC for the EOsc4Cancer⁵ project.

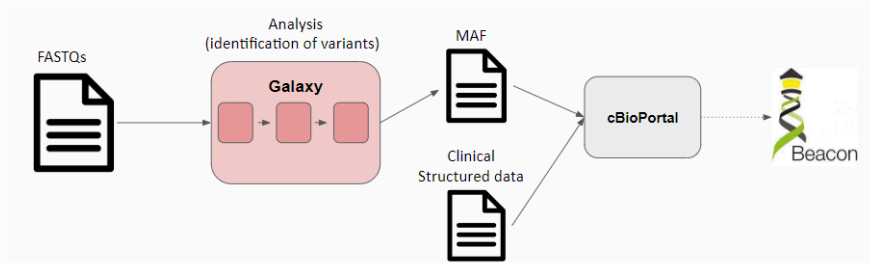


Figure 2. Conceptual orchestration of the demonstrator for cancer data mobilisation for discoverability and interrogation

³ <https://www.cbioportal.org/>

⁴ <https://beacon-project.io/>

⁵ <https://eos4cancer.eu/>



The demonstrator used Galaxy (local in Germany, at ALU-FR) to run a workflow designed to take the raw datasets, several FastQ files (from the at BSC's nextcloud), to process them to obtain aligned and annotated vcf files (one per sample), and to extract the minimum allele frequency (MAF) for each annotated variant. Then, the MAF files (deposited at the BSC's next cloud), paired with clinical structured data for each individual, were uploaded to cBioPortal (downloaded locally and ingested on a local instance).

4.3 5-safes RO-Crate compatible Trusted Research Environment

The goal of this demonstrator is to run any workflow written in CWL/Nexflow (and potentially other workflow languages) with a secure environment subject to appropriate governance approval from a RO-Crate using WfExS.

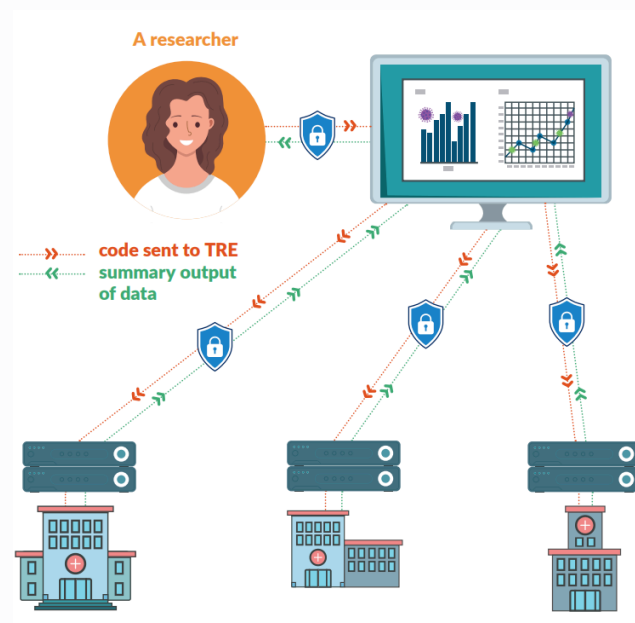


Figure 3. Conceptual design of a trusted research environment and their interconnection with a researcher, depicting the flow of code and secure results.

With TRE-FX as a successful proof of concept of a federation of Trusted Research Environments (TRE) with WfExS at its core, a clear conclusion from its development: any TRE needs to have an internal list of validated, and digitally signed⁶, workflows with all its dependencies internally materialised (workflow itself, containers, etc...), so nothing is pulled from outside during the analysis.

Within this demonstrator of GDI, the members of the BSC has been working on improving the use cases of WfExS as the core of a TRE, where local workflows being referenced through "file" protocol

⁶ <https://github.com/ResearchObject/ro-crate/issues/282>





and used containers are locally available as isolated environments. The containers, local Dockers of Podman, are registered in the TRE's vault.

A natural representation of signed workflows and containers is RO-Crates (following the Workflow Run profile). Members of the development team of TRE-FX, aiming to use signed workflows and containers as stated before, joined BSC developers of WfExS on advancing their development using as starting points signed workflows already existing RO-Crate(s), in order to replicate the same analysis using different inputs. As several milestones in the development of WfExS are related to the generation and consumption of RO-Crates, members of the BSC team on WfExS have been attending the design and development meetings since the very beginning of Workflow Run RO-Crate profile creation.

5. Results

Two out of the three demonstrators were able to record a short video proving the execution and outcome of the demonstrators:

- Distributed/federated analysis (PRS) of infectious diseases⁷
- Cancer data mobilisation across the analysis platform⁸

Although 5-safes RO-Crate compatible TRE was unable to record the short video, as additional efforts are underway to enhance its functionality before presenting a comprehensive demonstration, these endeavours were showcased at the GDI general assembly.


6. Discussion

6.1 Distributed analysis (PRS) for infectious diseases

Galaxy is a web-based platform used by tens of thousands of scientists across the world to analyse large biomedical dataset, including genomics. It has a friendly user interface and has been cited in thousands of scientific papers, being a successful example on how to democratise bioinformatics for non-experts in computer science.

Along the development of the demonstrator, bottlenecks with the use of Galaxy as a platform for distributed analysis of sensitive data in a privacy-preserving manner were noted:

1. Currently, it is not yet possible to run workflows without importing the data first, meaning that it was available into the user history and, therefore, available for any Galaxy workflow.

⁷  plink_project.mp4

⁸  GDI-cBioPortal.mp4



2. Another bottleneck faced using Galaxy was the disclosure of individual-level disease risks to the analyst, since each line containing the sample key was incremented with the total score for disease severity.
3. It could not be demonstrated the use of Galaxy for a fully federated analysis, as for now it is not possible yet to analyse data located in different Galaxy instances and retrieve only the aggregated results from these different locations as a single model.

These three issues alone and combined could potentially lead to privacy breaches.

The next iteration of the workshop will include the use of a new 3rd party tool, different from Galaxy, to explore calculating the same PRS in a real federated fashion and comparing its performance with Galaxy, as well as assert the differences in respect to data security and privacy respecting capabilities. Concurrently, discussions with Galaxy were initiated during the Barcelona event to enhance the integration of Galaxy with GDI's foundational infrastructure "starter kit". This integration is crucial as it aligns with the goal set forth by this demonstrator, leveraging AAI, DRS, REMS, and htsgat to achieve optimal outcome.

6.2 Cancer data mobilisation across analysis platform

The main discussion of the demonstrator is the use of Galaxy and how to upload clinical data in cBioPortal. For now, Galaxy has been the workflow manager to pass from raw data to MAF, the variant annotated format needed by cBioPortal. There has been discussion over the performance of Galaxy and the use of other workflow managers, such as Netxflow. Nevertheless, we need people with expertise in these areas to find the best way to connect the different platforms and run the workflows.

Another discussion is the robustness in the clinical data model of cBioPortal. There are only three mandatory variables in this platform, and the other clinical variables are described as recommendations. This leads to the fact that most of the clinical data in cBioPortal is completely different and is not possible to harmonise them to a future sharing of data. So, there is a need of finding common data models to have a minimal data common in all datasets or only use the genomic data from cBioPortal, that is already standardised with the MAF format, to share the data.

6.3 5-safes RO-Crate compatible Trusted Research Environment

This demonstrator was focused as a pure technical demonstrator compared to the previous two, having a scientific requirement behind. But, despite it, a preliminary version of an RO-Crate following the Workflow Run profile was obtained under the hood of this demonstrator within GDI. The results of the work of this demonstrator was the consolidation of RO-Crates generation in a prospective way from already existing WfExS workflows. Some relevant issues were detected during the





demonstrator, not being able to provide proof of the current state of work done: WfExS was missing some technical capabilities to be properly integrated in the proposed scenario by TRE-FX.

Furthermore, the demonstrator highlighted a range of issues requiring community and internal discussion within the GDI project, notably regarding workflow verification. This could be addressed through container and/or RO-Crate signature verification, as well as leveraging trusted third-party platforms like WorkflowHub. Additionally, there is a need for a defined strategy to handle scenarios where a TRE revokes a workflow due to distrust in one or more components.

7. Next steps

The next steps regarding federated analysis in GDI involve integrating the different demonstrators with the essential components of the GDI starter-kit, to make sure that these technologies will be ready for the nodes that deploy the starter kit in the future.

7.1 Distributed/federated analysis (PRS) for infectious diseases

Additionally, we will expand our reach by incorporating Portugal and Norway into the infectious diseases demonstrator.

A forward-looking proposal entails the inclusion of a new demonstrator using DataSHIELD⁹, a powerful tool designed for federated analysis, as well as Flower¹⁰, a powerful python framework for federated learning , and we are considering its incorporation into the TRE demonstrator.

7.2 Cancer data mobilisation across analysis platform

The next steps that will be done jointly with EOSC4Cancer is to try to send the genomic data from Galaxy to cBioPortal in the most automatic way possible, as now the upload of data is manual. Then, demonstrate it with more diverse test data. Another extension that would benefit the demonstrator is connecting cBioPortal with EGA, by retrieving the initial data from the archive and passing it to cBioPortal or from a study in cBioPortal to find the corresponding study at EGA.

7.3 5-safes RO-Crate compatible Trusted Research Environment

In 2024 WfExS developers are finishing the capabilities to both reproduce an already executed workflow with the very same inputs described in the input RO-Crate, and to replicate the scientific analysis allowing to replace one or more of its inputs. Furthermore, a new discussion group, aiming to understand how this technology can and will be translated to GDI, will be created together with the EN-TRUST project at its start.

⁹ <https://www.datashield.org/>

¹⁰ <https://flower.ai/>





Special recognition must be given to the development of Galaxy, as it played a pivotal role in two out of the three demonstrators. Collaborative efforts between EuroScienceGateway and GDI have been directed towards enhancing Galaxy and its Pulsar nodes to interface with and comprehend the GA4GH TES (Task Execution Service)¹¹. Given that Galaxy serves as both a workflow manager and dispatcher, advocating for the integration of GA4GH WES (Workflow Execution Service)¹² would be redundant. The incorporation of GA4GH TES into Galaxy and its Pulsar nodes stands as the crucial component needed for distributed analysis within GDI's federation, while also accommodating WfExS.

8. Conclusions & Impact

The completion of this first interaction of the GDI demonstrators marks a significant milestone towards collaborative research efforts in the distributed and federated analysis of clinical and genomic data.

8.1 Advancements in Distributed and Federated Analysis

The exploration of distributed/federated analysis methodologies, exemplified by the Polygenic Risk Score (PRS) analysis of infectious diseases, underscores the potential of collaborative frameworks in leveraging disparate datasets while preserving data privacy. Despite encountering challenges with existing platforms like Galaxy, such as data import requirements and privacy vulnerabilities, the demonstration highlights the need for evolving technologies capable of facilitating distributed/federated analysis seamlessly across diverse geographical locations.

8.2 Cancer Data Mobilization and Integration

The utilisation of Galaxy and cBioPortal in the mobilisation and integration of cancer data underscores the importance of interoperable platforms in accelerating translational research. Discussions surrounding platform performance and clinical data harmonisation underscore the ongoing need for interdisciplinary collaboration in refining data management protocols and establishing standardised data models. The envisioned integration of cBioPortal with repositories like the European Genome-phenome Archive (EGA) represents a crucial step towards facilitating seamless data sharing and enhancing accessibility to genomic resources for researchers worldwide.

¹¹ GA4GH TES API: Bringing Compatibility to Task Execution Across HPC Systems, the Cloud, and Beyond: [Link](#)

¹² GA4GH Workflow Execution Service (WES): [Link](#)





8.3 Advancements in Trusted Research Environments

The development of a 5-safes RO-Crate compatible Trusted Research Environment (TRE) lays the foundation for establishing secure and reproducible research infrastructures. While significant progress has been made in generating RO-Crates and integrating workflows within the TRE framework, challenges related to technical capabilities and workflow reproducibility persist.

8.4 Overall Impact and Future Directions

The successful execution of the demonstrators within the GDI underscores the collective efforts of its partners in advancing data-driven research methodologies and fostering collaborative innovation. It must be highlighted that none of the demonstrators presented here utilised the GDI's Starter Kit, as their primary goal was to identify technologies from which GDI could benefit through adoption. However, it is worth noting that Galaxy is a technology slated for integration into the Starter Kit. Subsequent iterations of the demonstrators will incorporate the usage of the Starter Kit from Pillar II as the technologies they test mature sufficiently. Moving forward, continued investment in technological infrastructure, interdisciplinary collaboration, and policy frameworks will be essential to realising the full potential of global data initiatives like GDI.

In conclusion, the completion of the first iteration of the GDI demonstrators signifies a crucial step towards realising the vision of a global data ecosystem characterised by collaboration, innovation, and responsible data stewardship. By harnessing the power of distributed and federated analysis, the GDI initiative sets the stage for transformative advancements in research, healthcare, and beyond. In parallel to D8.8, the D7.4 describes a series of use cases that GDI will have to cover. The goal of these demonstrators is to converge on those use-cases and provide the technologies to make them a reality under the hood of GDI. Consequently, mature software will be provided to Pillar II for its inclusion into the Starter Kit or other components, according to its discretion.

