Grant Agreement Number: 825225

Safe-DEED

www.safe-deed.eu

D6.3 Personal data trials report final version

Deliverable number	*D6.3*
Dissemination level	***Public***
Delivery date	*30th November 2021*
Status	*Final*
Author(s)	*Ioannis Markopoulos*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825225.

Changes Summary

Date	Author	Summary	Version
07.09.2021	Ioannis Markopoulos	TOC preparation	0.1
11.10.2021	Ioannis Markopoulos	First version	0.2
17.10.2021	Evangelos Kotsifakos	Input received	0.3
05.11.2021	Ioannis Markopoulos	Final draft for peer review	0.4
12.11.2021	Pert Knoth, Mihnea Tufis, Evangelos Kotsifakos	Peer review	0.4
16.11.2021	Ioannis Markopoulos	Final	0.5

Executive summary

As privacy and trust remain key in the data sharing debate, Privacy enhancing technologies[1] (PET) will play a prominent role in the future. Safe-DEED takes a highly interdisciplinary approach, bringing together partners from cryptography, data science, business innovation, and legal domain to focus on improving Security technologies, improving trust as well as on the diffusion of Privacy enhancing technologies to keep up pace with global macrotrends and the data economy, to enable the fastest possible growth.

This document aims at supplementing the final version of the Safe-DEED demonstrator containing reports on the demonstrator applications, the data used, and the trials organized in NOVA (ex. Forthnet) use case.

A strategic decision was taken in the project to implement two versions of the demonstrator:

- The first version is fully functional aiming specifically at the professional community building actions. This demonstrator version is set as confidential according to the project DoA.
- The second version is used to address a wider public audience through the project web site. This demonstrator version is public and has been developed on top of the DoA requirements. It exposes its functionality with predefined scenarios and datasets. The Safe-DEED project considers this as a very useful channel to increase awareness and disseminate the project achievements. Such a strategic decision to implement a public version was taken in the project since the research results are considered to be sound and can easily by adopted by the professional community.

Having first thoroughly evaluated the demonstrator internally within the project, a set of additional activities have been planned to evaluate the demonstrator further:

- within the partners' organizations;
- with third part professionals via respective interview sessions;
- via a public questionnaire in collaboration with the ReachOUT[2] EU project.

[1] https://www.r3.com/gartner-2021-privacy-enhancing-computation/

[2] https://www.reachout-project.eu/

Table of Contents

List of Figures

Abbreviations

Abbreviation / Term	Description
PSI	Private Set Intersection
MPC	Multi-Party Computation
CRM	Customer Relationship Management
QI	Quasi Identifier
GDPR	General Data Protection Regulation
DVC	Data Valuation Component
QDSC	Qualitative information extraction and Data Scoring Sub-Component
DIL	Data Ingestion Layer
DoA	Description of Action
ADAS	Automatic Data Analysis and Scoring
S2VM	Score-to-Value Mapping
CPL	Communication and Presentation Layer
UI	User Interface
PPC	Project Professional Community

1 Introduction

Safe-DEED WP6 aims at providing the means to demonstrate how industries can benefit from big dataspaces and data marketplaces while trying to extract value of data by correlating with external sources and analysts securely and in a GDPR compliant way.

This was achieved via setting-up novel demonstrators incorporating GDPR compliant business processes and anonymized CRM data

Two cycles of trials were performed in NOVA (ex. Forthnet) premises with the participation of LST and RSA towards our effort to define and evolve the demonstrators.

The final version of the demonstrators has been reported in the deliverable D6.2, though additional improvements (e.g., integration of the new PSI application, etc.) have been achieved beyond this point to maintain the up-to-date status of the demonstrator.

The demonstrators aim at showcasing the ability to perform digital transformation creating a data exchange ecosystem in a GDPR complaint way.

This document aims at outlining the actions performed to evaluate the demonstrators with various stakeholders and receive valuable input for their further use.

2 Safe-DEED personal data demonstrators

A strategic decision was taken in the project to implement two versions of the demonstrator:

- The first version is fully functional aiming at being used within the professional community building actions. This demonstrator version is characterised as confidential according to the project DoA.
- The second version is used to address the wide audience through the project web site. This demonstrator version is public and is implemented beyond the DoA requirements. It exposes the complete functionality with all the confidential demonstrator scenarios and datasets to the wide community, but users are not able to upload their own data due to security reasons.

Safe-DEED project considers this as a very useful channel to increase awareness and disseminate the project achievements.

Such a strategic decision to implement a public version was taken in the project since the research results are considered to be sound and can easily by adopted by the professional community.

2.1 Demonstrator overview

2.1.1 The demonstrator development process

The Safe-DEED demonstrator has undergone a major revision from its initial version, both in terms of User Interface and overall functionality.

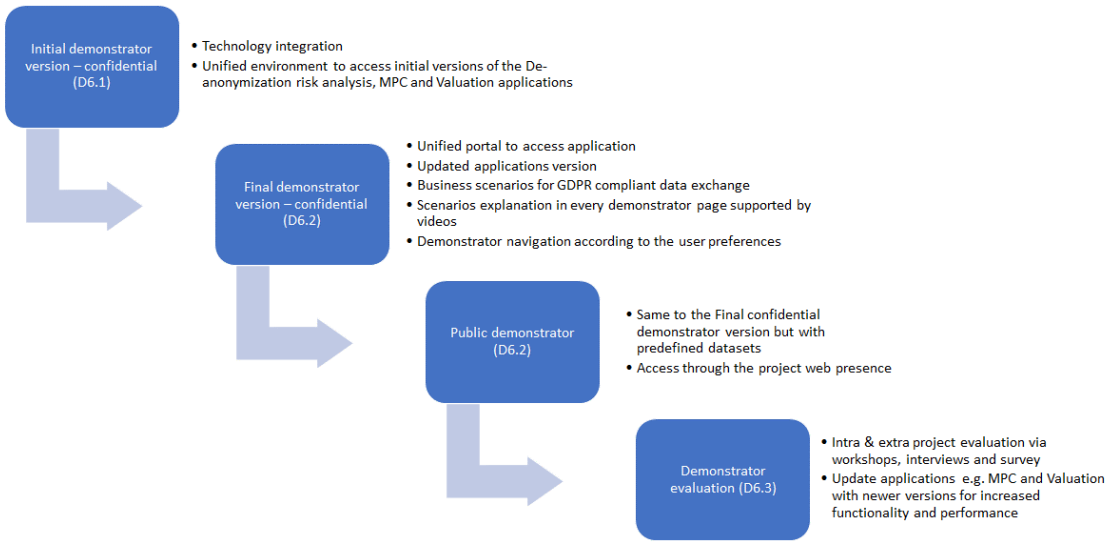

Figure 1: WP6 Demonstrator implementation process

The first version of the demonstrator aimed at bringing together the technology, scientific and business partners. The main goal was to create the initial environment and set the ground for data extraction, exchange, and analysis.

In the second version of the demonstrator, the focus was based on increasing usability, application functionality and most of all incorporation of respective business scenarios.

Last but not least, an analysis of GDPR recommendations were taken into account defining respective roles in the scenarios thus imitating entrepreneurial processes.

The demonstrator integrates applications that are built in different programming languages (Python, Java, C) and with different approaches. This approach safeguards openness and expandability of the demonstrator.

A common Graphical User Interface (GUI) has been implemented and the applications have been tested using real anonymized CRM data from NOVA. Currently the confidential version of the demonstrator is installed in Google Cloud and can be accessed via a secure connection while the public version can be accessed through the project website[3].

[3] https://demo.safe-deed.eu/

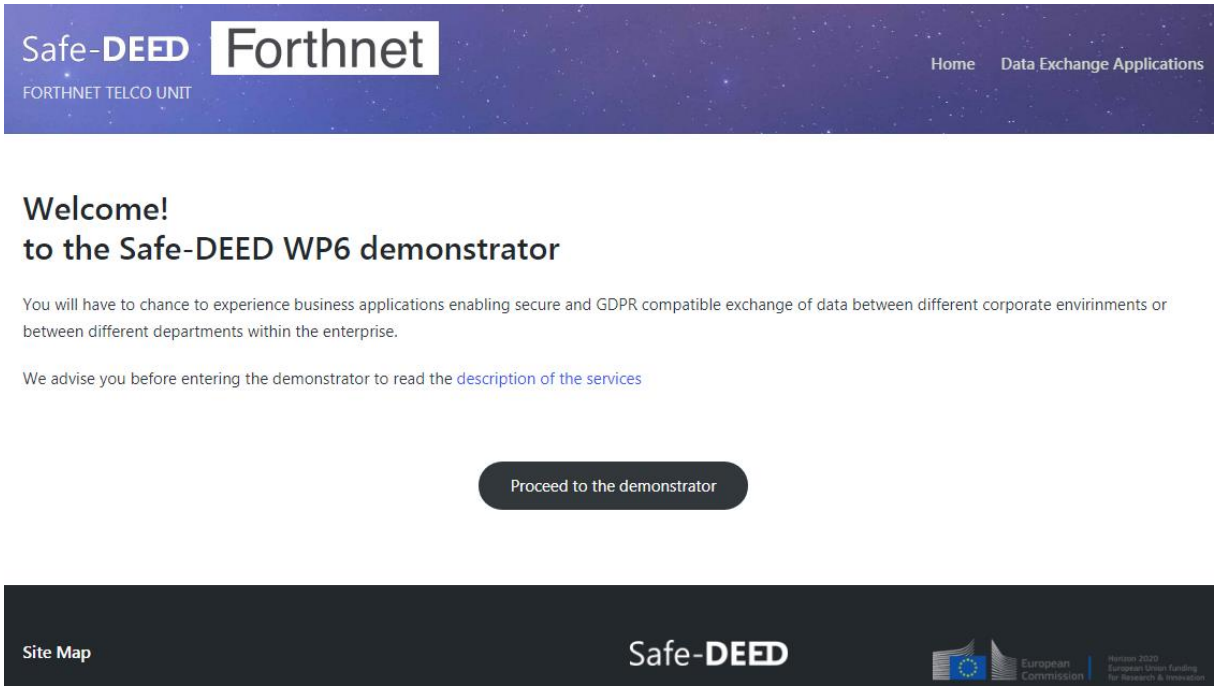

Figure 2: THe WP6 demonstrator welcome screen

In the following paragraphs, we present the final demonstrator version also describing the supported business scenarios.

2.1.2 The demonstrator applications

The demonstrator incorporates the applications presented below:

Private set intersection application

This part of the demonstrator exemplifies the power of private set intersection (PSI) protocols in the context of marketing. In particular, we enable two companies – in our case NOVA, and another company "C" (e.g., a Bank operating in the same territory as NOVA) – to improve their own marketing strategy for customers they have in common.

De-anonymisability risk analysis

It performs a de-anonymisability analysis of the dataset. Despite datasets not containing any personally identifying information (PII), such as name, address, etc., individuals can be identified through their quasi-identifiers (QIs). QIs are the attributes whose combination can serve as a unique identifier for individuals.

Data valuation application

This part of the demonstrator describes the initial implementation of the Data Valuation Component (DVC).

The supported algorithms are:

- selected regression, classification and clustering algorithms;
- a rule-based algorithm for generating the economic value of the input data set.

2.1.3 Business and GDPR compliant scenarios for data exchange

GDPR is an EU law with mandatory rules for how organizations and companies must use personal data in an integrity friendly way. Personal data means any information which, directly or indirectly, could identify a living person. Name, phone number, and address are examples of personal data.

Processing data means collecting, structuring, organizing, using, storing, sharing, disclosing, erasing and destruction of data. Each organization that processes personal data (which is every organization with employees and customers) must ensure that the personal data it uses fulfils the requirements of the GDPR.

With the digital transformation trend that most on the enterprises and organisations are planning, there is a temptation. It's the allure of just focusing on the deluge of information available and the potential for business advancement, if only one can successfully aggregate, interrogate and monetise it. Digital transformation and data protection, on the other hand, seem to be at odds. After-all, the principles of adding more control to data usage can feel like roadblocks on the path to becoming data driven.

Further examination reveals a quite different relationship between digital transformation and data protection. In many ways, they are co-dependent.

According to IDC[4] "More than a third of organizations (37%) have already started integrating and executing a digital-first approach, and 7% say they're already an enterprise-wide digital business. Still, almost half (45%) of IT and business leaders surveyed say their companies are in the very early stages of becoming a digital business – either gathering information or just beginning to formulate a digital-first strategy…. It's no surprise that digital-first enthusiasts have already jumped on the big data/analytics bandwagon. Big data/analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers. It also drives process efficiencies and employee productivity, which are primary goals of digital-first initiatives".

One of the main goals of WP6 is to provide concrete business scenarios assisting the incorporation of advanced privacy solutions to emerging business processes thus assisting enterprises' and organizations' digital transformation as well as the creation of the digital ecosystem.

Three main scenarios have been designed and presented:

Scenario 1 – Private Set Intersection (PSI)

The scenario demonstrates the PSI application allowing the user to exchange data between them and another partner.

The following architecture has been chosen:

- The PSI library (respectively its wrapper);
- The demonstrator UI;
- NOVA's and external company's CRM database.

The demonstrator UI can either directly connect to the CRM database or provide a field to import data from a CSV file. Then, as soon as the demonstrator UI has access to the data, one company starts the PSI library as a server and the other company as a client. Once the connection between the two companies is established, the PSI library runs and outputs the intersection of the two sets to the party that has initiated the interaction. The demonstrator UI then displays the resulting set to the user.

[4] https://cdn2.hubspot.net/hubfs/1624046/Digital%20Business%20Executive%20Summary_FINAL.pdf

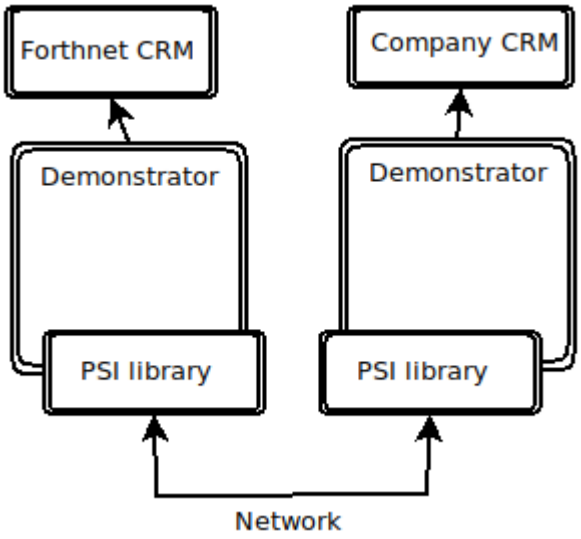

Figure 3: PSI Trials Diagram

Scenario 2 – De-anonymisability

Scenario 2.1 - De-anonymisation risk analysis

The goal of this trial was to apply a battery of de-anonymisation tests on NOVA's data in order to raise privacy red flags.

Scenario 2.2 – K-anonymisation with PrioPrivacy Tool

The goal of this trial was to reduce the de-anonymisation risks in case NOVA would decide to release, exchange or sell their dataset to a third party(-ies).

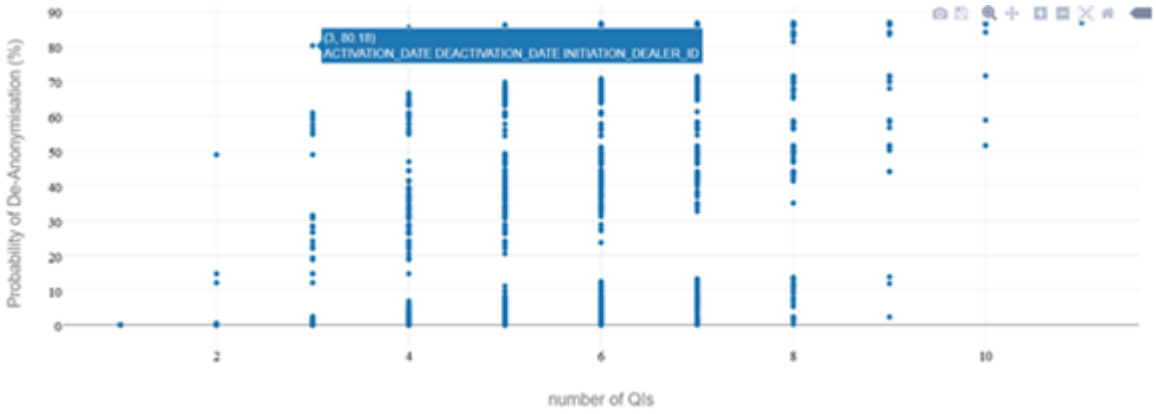

Figure 4: De-anonymisation risk analysis of Assets

Scenario 3 - Data valuation

In this scenario there are two types of outputs being generated: the one is the "pdf output" and the second the "profiles data set" page.

For the profile files, they include the following sections:

1. Dataset info: size, shape, duplicate percentage;
2. Variables types: how many columns of each type;
3. Warnings: if columns contain a large proportion of only 1 value, or large proportion of 0s, etc.;

4. Variables: a summary of the descriptive stats for each column, including histogram of the distribution, number of unique values in that column, number of missing values in that column, mean, std, max, min. Then, if you toggle the details of each column, you can see detailed quantile stats, descriptive stats, the histogram of the distribution, common values for each given column and extreme values (top-5 max and min);
5. Correlations: a set of correlation matrices based on different correlation coefficients between all pairs of columns. This informs the suggestion to discard from model design those columns that have a high correlation coefficient;
6. Missing values: a histogram of how many values are present in each of the columns;
7. Sample: a sample from the head and tail of the dataset.

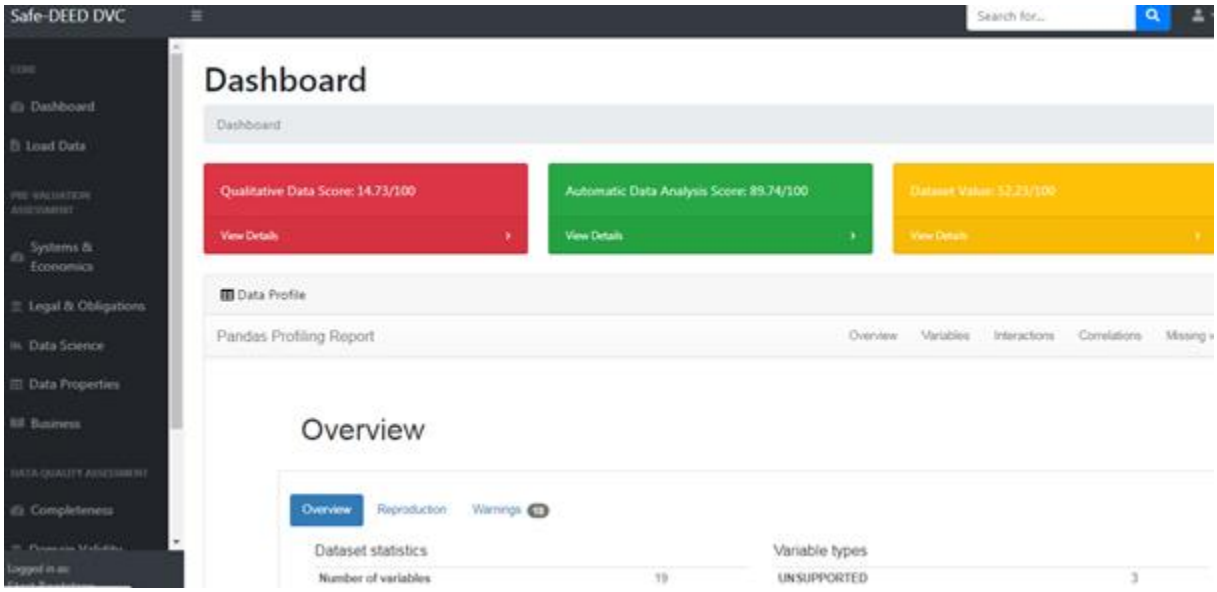

Figure 5: Output of the Data Valuation Application

3 Demonstrator evaluation methodology and target groups

The Safe-DEED project has deployed a set of multidisciplinary activities to disseminate the project outcomes, ranging for building a professional community to organizing workshops.

Beyond the respective activities that were reported by WP8, WP6 has launched the final set of the evaluation activities.

The evaluation methodology consists of:

- Intracompany evaluation workshops;
- Demonstration and interviews with domain professionals;
- Liaising with external project to launch a survey.

3.1 Intracompany demonstrations and workshops

Two workshops were scheduled on May 26th and 27th 2021 internally to NOVA.

The workshops targeted professionals from marketing, product development and analysis departments.

In each workshop, a cohort of three professionals at mid and upper management level participated.

The duration of each workshop was two hours and the workshops were conducted physically with the presence of all participants.

Ioannis Markopoulos (NOVA) was the moderator.

The agenda was:

- Safe-DEED project presentation;
- Presentation of other related initiatives i.e. GAIA-X initiative, dataspaces initiative by Digital Europe and NOVA's participation in the TRUSTS project;
- GDPR processes in NOVA;
- Safe-DEED demonstrator presentation and data exchange scenarios elaboration;
- Discussion on the usability of the demonstrator and the applications;
- Evaluation outcome.

The outcome of both workshops is summarized in the following:

- No matter if GDPR has been applied for approximately two years in the company there are still unclear areas with respect to everyday processes;
- Initiatives, such as Safe-DEED were perceived as very important by the participants to increase awareness and provide information via real business scenarios and applications;
- The demonstrator (both public and confidential) should become part of NOVA's intranet and be used in relation to the GDPR processes training;
- Demonstrator videos and narration are very important;
- The personal choice on how to navigate the demonstrator is very useful;
- The range of the scenarios in the demonstrator is significant to start thinking beyond the everyday processes.

The following proposals were made aiming to improve the demonstrator:

- It would be beneficial if an enterprise could create additional scenarios applicable to own everyday operations;
- Also it would be beneficial if the demonstrator was extended to allow the creation and upload of sample datasets meaningful to the trainee (e.g. via a wizard);
- No matter if the videos are very helpful, sometimes the terminology and the figures are difficult to be understood without external assistance e.g. QA and the doted histograms. It was proposed to improve the analysis e.g. by adding explanations;
- With respect to the PSI, there was a fear that business confidential information could be revealed even if data were not transferred. One should preapprove the questions. Alternatively an AI based threat alert functionality could exist. Third party certification of the application would be a plus;
- With respect to the valuation application it can be perceived as datasets quality analysis functionality;
- A set of training material can be implemented with hands on examples and collaborative training e.g. one team "against" another.

3.2 Workshops/Interviews with external professionals

Following intracompany evaluation of the demonstrator, a follow-up evaluation was scheduled in 2021 in the form of interview.

The interviewees were:

- 2 representatives (1 high level management and 1 mid-level management) of a multinational consultancy firm (Greek branch). The first interview was conducted face to face while the second via an online service;

- 1 representative at the level of partner of a multinational consultancy firm (Greek branch). The interview was conducted via an online service;
- 1 representative of a pharmaceutical industry in Greece (mid-level management via and online service);
- 2 representatives of a British SME with knowledge on large scale demographic data collection and analysis

Each interview duration was approximately one hour. Interviews agendas were:

- Safe-DEED project presentation
- Presentation of other related initiatives i.e. GAIA-X initiative, dataspaces initiative by Digital Europe and NOVA's participation in the TRUSTS project;
- Safe-DEED demonstrator presentation and data exchange scenarios elaboration;
- Discussion on the usability of the demonstrator and the applications;
- Evaluation outcome.

The outcome of the interviews can be summarized as follows:

- Initiatives, such as Safe-DEED are very important to increase awareness and provide information via real business scenarios and applications

  (interviewee quote: "It is important to have consistent framework scenarios enabling data exchange otherwise each one will have his own ground truth and interpretation…");

- Providing and effect assessment of risk in relation to anonymised data being de-anonymised based on a range of data points is very helpful

  (interviewee quote: "It is usual that we rely on the internal anonymisation practices and ignore the information that exists widely. Such initiatives make us think and improve the process…");

- The demonstrator could become part of sales process, project implementation and platforms[5] towards the digital transformation of large enterprises

  (interviewee quote: "I want to include this in my sales speech to show that the transformation is tangible and customisable for the company...");

- The use of video explanation is extremely helpful to augment the individuals' personal exploration of the demonstrator

  (interviewee quote: "Video explanations are great. People do not want to read too much anymore. With the videos we attract their attention…");

- The range of the scenarios in the demonstrator is significant to start thinking beyond the everyday processes towards creating a data ecosystem

  (interviewee quote: "We may use part of the scenarios. Seeing more opens our mind.").

The following recommendation were provided by the participants to improve the demonstrator:

[5] E.g. Mydex's Inlued platform, Respective interview statement "We could see how this approach would add value in the work we do with our Inclued platform and also as part of a set of tools for Data Trusts and Shared Data Spaces in ensuring they are correctly protected and designed"

- The demonstrator could be offered **as a white label product** to consultants or projects managers to include it in their portfolio;
- It would be beneficial if the demonstrator users could **create additional simple scenarios** applicable to special operations needs e.g. via using a respective wizard;
- Further improvements could include the **integration of GAIA-X scenarios**, terminology and standards;
- **Include specific examples** on use cases **to demonstrate** real-world applications and the **benefits compared to current approaches**. Reduced Risk, Cost and increased compliance;
- Provide descriptions answering to considerations with respect to **scaling capabilities for high demanding transactions both in terms of real timeliness and analytical processing**.

3.3 Liaison with other projects

In order to extend demonstrator, reach the Safe-DEED project collaborated with the HORIZON 2020 ReachOUT[6] project.

The ReachOUT project aims at connecting research projects with beta testers and early users. They assist managing beta-testing campaigns, discovering new software and trying them in their environment. Comprehensive surveys and response analysis tools were provided in order to supplement the tool evaluation.

Safe-DEED defined a questionnaire accompanying the WP6 demonstrator evaluation process.

Responses are summarized in the following:

- All features seem interesting (de-anonymisation risk analysis, aggregated data analysis, data valuation and data exchange);
- Additional use cases may be useful to include. Also, scenarios with increased interactivity;
- The demonstrators target knowledgeable professionals.

4 Conclusions, overall evaluation and demonstrator usage beyond the end of the project

This deliverable aims at supplementing the two previous deliverables i.e., D6.1 and D6.2 which outlined the evolution of the WP6 demonstrator and focuses on analysing the feedback received during the evaluation actions.

The focus of the consortium, beyond the PPC actions and the dissemination actions (e.g., webinars co-organised with the EU project TRUSTS[7], etc.), was to involve diverse professional expertise in the evaluation process.

To this end, digital transformation consultants, company data analysts, and SMEs with data collection and analysis focus were employed.

In addition, effort was placed to interview highly ranked employees/executives and decision makers in order to evaluate the business potential of the demonstrator as well.

In brief the evaluation outcome can be summarized in the following:

- Initiatives, such as Safe-DEED are very important to increase awareness and provide information via real business scenarios and applications;

[6] https://www.reachout-project.eu/

[7] https://www.trusts-data.eu/

- The demonstrator could become part of a sales process, project implementation and external platforms towards the digital transformation of large enterprises;
- The range of the scenarios in the demonstrator is significant to start thinking beyond the everyday processes towards creating a data ecosystem.

Beyond the project completion the demonstrator will be incorporated in NOVA's training material with respect to GDPR.

Evaluation feedback has been extensively communicated and discussed among the consortium partners. Discussions targeted both the applications' owners and the overall demonstrator improvement and dissemination as a whole. Safe-DEED believes that this will be a comprehensive basis to:

- Improve partners and Project Professional Community (PPC) digital transformation process;
- Disseminate applications more effectively than just presenting the respective functionality, since relevant business aspects are referred as well;
- Contribute to the creation of the digital ecosystem that all Safe-DEED partners envisage.

There was a wide consensus that the Safe-DEED partners will discuss ways to further exploit and improve the demonstrator beyond to the project conclusion. The Kaizen[8] methodology will be adopted for continuous improvement of the demonstrators, taking into account the market maturity increase in the data ecosystems domain and the respective business/user needs evolution.

[8] https://www.kaizen.com/