Deliverable D6.1

Data Management Plan

Point of Contact	Samuel Fricker
Institution	FHNW
E-mail	samuel.fricker@fhnw.ch
Phone	+41 79 196 9629

Project Acronym	GEIGER
Project Title	GEIGER Cybersecurity Counter
Grant Agreement No.	883588
Topic	H2020-SU-DS03
Project start date	1 June 2020
Dissemination level	Confidential, only for members of the consortium (including the Commission Services)
Due date	M03
Date of delivery	M04
Lead partner	ATOS
Contributing partners	FHNW, BBB, SKV, HAAKO, CLUJ-IT, E-ABO, SCB, SRA, CL
Authors	Jose Francisco Ruiz (ATOS), Samuel Fricker (FHNW), Jürg Haller (BBB), Euplio Di Gregorio (SKV), Moritz Dietsche (HAAKO), Andrei Kelemen (CLUJ-IT), Heike Klaus (E-ABO), Vlad Florian (SCB), Tony van Oorschot (SRA), Loredana Bartels (CL)
Reviewers	Bettina Schneider (FHNW), Samuel Fricker (FHNW)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 883588 (GEIGER). The opinions expressed and arguments employed herein do not necessarily reflect the official views of the funding body.

Revision History

Version	Date	Author	Comment
0.1	10/08/2020	Jose Francisco Ruiz (ATOS)	ToC and description of subsections
0.2	25/08/2020	Jose Francisco Ruiz (ATOS)	Sections 1 and 2 completed
0.3	01/09/2020	Jose Francisco Ruiz (ATOS)	Completed version for review
0.5	09/09/2020	Samuel Fricker (FHNW)	Clarity and adaptation to project needs
0.6	28/09/2020	Jose Francisco Ruiz (ATOS)	Final draft
0.9	30/09/2020	Samuel Fricker (FHNW)	Release candidate
1.0	30/09/2020	Bettina Schneider (FHNW)	Quality review

Contents

Abbreviations, Participant short names and Glossary

Abbreviations

MSE	Micro or Small Enterprise
DMP	Data Management Plan

Glossary

Partner	A member of the GEIGER consortium.
Participant	A subject external to the GEIGER consortium participating in a GEIGER research study
Researcher	A staff member of a GEIGER Partner collecting or analysing research data
Pilot	The GEIGER project will validate and demonstrate the GEIGER solution in three countries, each featuring a set of studies that together are called a pilot.

Abstract

This deliverable presents the Data Management Plan of the GEIGER project (Grant Agreement No.: 883588), funded by the European Commission. Several areas in the GEIGER project have been identified to require a data management plan that specifies how collected or generated data is going to be processed, used, and stored in the project.

Some activities of GEIGER, involving the pilots and users, will require validation and testing by users. We need to specify how the elicitation of requirements, testing and validation of the use cases will be done. For some activities carried out in the project, it may be necessary to collect personal data (e.g. name, e-mail, etc.) even though the project will avoid collecting unnecessary data unless required by a specific task. This data must be protected under the EU's General Data Protection Regulation (GDPR, Regulation (EU) 2016/679) on the protection of individuals with regard to the processing of personal data. Additionally, any national or local legislation applicable in each case will also be applied. Any personal data collected in the project will require a signed informed consent form from the subjects, following the forms documented in D7.1 about ethics and data protection.

This document is the initial Data Management Plan (DMP). Although no second version is scheduled, we will monitor the project for identifying a need to update the document with additional information or methodologies for ensuring the protection of the data managed in the project. Any updates will be documented in the WP4 deliverables that include the plan and report of GEIGER validation and demonstration.

This document contains initial information about the data the project will generate, when and how it will be used or made accessible for validation or reuse and how it will be stored and protected. The purpose of the Data Management Plan is:

- to provide the handling of research data during and after the end of the project
- what data will be collected, processed, or generated
- the methodology and standards to be used
- whether data will be shared or made open access and
- how data will be stored (also after the end of the project)

Therefore, one of the focus of the datasets will be in the pilots of the projects, identifying what data they will generate, the type (real, fake, synthetic, etc.). Additionally, we will define the methodology to follow by all the partners for working with the data in the project.

1. Introduction

1.1 Purpose of the Document

The objective of this deliverable D6.1 is to define relevant information and describe the details concerning the management of the data in GEIGER, how it will be collected and used by partners and the type of data. D6.1 defines the creation, processing, storage, and protection of data within the consortium of GEIGER partners.

The project aims to create a cybersecurity solution for MSEs that improves both their technical capabilities (protection against cyberattacks) and training and awareness for different types of employees. It could be described as "cybersecurity-as-a-service" in the sense that we will provide services of cybersecurity aimed to MSEs for increasing their capabilities, knowledge, and resilience.

To achieve this objective, we need to manage from the beginning good documentation and implementation of the data that will be used by use case partners for eliciting requirements and for testing and validating the GEIGER solution. Validation will involve the delivery of the services to stakeholders that will be specified in D1.1 and D1.2.

Several non-functional requirements and goals are central to data management. Privacy and confidentiality are important in this project as we need to give assurance to the users that any data used in the system is processed and used. Also, ethics is another mandatory element that is linked to data management. The ethics aspects of the project are reflected and described in the deliverables D7.1 (Humans) and D7.2 (Protection of Personal Data).

The data we obtain from the use cases are about users and the organisations using the GEIGER solution. Users data could be related to the GEIGER tools or Security Defender training. The data of the organisation could be obtained by tools used for appraising risk, closing vulnerabilities, or respond to incidents. Therefore, we discussed with the use case and technical partners about the type of data they plan to use for the testing and validation of GEIGER in the use cases. The input obtained guided the data management plan and the process to be followed in the project.

1.2 Relation to Other Project Work

This deliverable D6.1 is linked to the ethics deliverables, D7.1 and D7.2, that cover consent for information for the use cases and any ethic constraint. In particular, D7.2 will describe the handling of personal and confidential data and include as an annex the non-disclosure agreement to be signed by third-parties like national CERTs and Micro or Small Enterprises (MSEs) that will collaborate for GEIGER validation and demonstration.

This deliverable D6.1 is also linked to the requirements and architecture definition deliverables, D1.1 and D1.2. The data management plan will be adapted if required by decisions related to requirements and architecture definition. Such adaptations will be documented by the WP4 deliverables that include the plan and report of GEIGER validation and demonstration.

1.3 Structure of this Document

The structure of the document is as follows:

Section 1: this section describes the objective of the document, what we want to achieve and the planning for the data management plan along with the whole duration of the project

Section 2: describes the data management plan of the project, its objectives, the roles and responsibilities of the different actors involved in it, etc.

Section 3: this section describes the strategies and type of data the use case partners plan to use in the project. We conducted a short interview with each use case partner to obtain this information, so from the beginning, we can have a clear idea of the type of data that will be used in the project.

2. Data Management in GEIGER

2.1 IPR Management and Security

GEIGER plans to design and develop a cybersecurity solution for MSEs that not only protect their business from a technical point of view but also to increase the knowledge and expertise in cybersecurity for the employees. Additionally, GEIGER aims to improve the business of MSEs by allowing them to create secure services, improve existing ones, or access to new markets with these services.

The consortium includes partners from the private sector (tools and service providers), public (national CERTs/CSIRTs), and end-users (as MSEs or associations of MSEs) that collaborate to create this innovative solution. The multi-party collaboration implies that some partners may have intellectual property rights on their technologies or data used/managed in the project. For this reason, the consortium will protect the data and always check with the related partners before publishing any data.

We will follow a holistic approach that focuses on protecting the data, the privacy and confidentiality concerns related to the data, and the rules for publication of the data. In the project, we will perform risk analysis for data management for the collection, processing, transmission, and use of personal and confidential information provided in the project.

From the technical point of view the cybersecurity protection of the data in the project, the data flows, will include, among others:

- Secure protocols such as HTTPS and SSL
- Authentication and authorisation module for login in the system and identify the services or solutions exchanging data
- CAPTCHA technologies and similar for protecting the system against bots
- Code analysis for detecting any vulnerability or risk for data leaking

We will also provide protection measures against cyberattacks in the development and testing environment and access control measures to control and monitor the access to the GEIGER framework and data.

2.2 Personal Data Protection and Security

The GEIGER project works closely with end-users from the beginning of the project as they are the main stakeholders of the GEIGER solution. For this reason, one of the first activities was to analyse how the stakeholders plan to provide information for GEIGER and their plan for validating and demonstrating the GEIGER solution in the pilots. We elicited specific details about data to be used or provided by them in the project and the plan to use real users in any of these phases. After conducting these interviews, all of them confirmed they plan to minimise the use of real data or users, de-personalise data provided to the system and stakeholders, and use synthetic data where appropriate. In the next section, each use case partner describes in detail what data they will provide.

GEIGER data management will follow the following procedure for protecting the data flows or any data leak to limit the risk of cyberthreats:

- Keep potentially critical data anonymised,
- Encrypt data if a partner requires it,
- Limit the use of USB flash drives,
- Control access to the data in the project, and
- Label files in a structured way to have homogenisation of the data collected.

2.3 Objective of the DMP

The objective of the data management plan is to define how the data generated in the project is managed. The outputs are a) the type of data the end-users of the project will use and b) the dataset for each of them.

Even though the data planned to be used will be anonymised and, where appropriate, replaced by synthetic data, we think it is necessary to keep structured information of the data and how it is managed.

2.4 Roles and Responsibilities

After discussing with the technical and use case partners, we identified initially two different roles that will be related to the data generated and used in the project.

Data owner: this role is related to the use case partners. They will generate real or synthetic data that will be used for validation and demonstration of the GEIGER solution.

Data user: the role uses the data provided by the use case partners. Several entities in the GEIGER ecosystem are data users. The tool provider entities will use the data for analysis of the cybersecurity status of the MSE, for providing recommendations or reactions, and for sharing data with stakeholder entities (e.g. CERTs). The stakeholder entities will use the data for wide-area cybersecurity monitoring, offering insights regarding cyber threats and recommendations for protection against these threats.

2.3 GEIGER Dataset

Data used in a project can be of different types. Also, the way data is obtained, processed, and stored can vary and be specific according to the domain. Therefore, in a DMP, the first consideration is to define the file format to be used. Many well-known file formats are containers for standard file formats, which can be used for facilitating the process for obtaining, analysing, and providing results of the data. Also, file formats can be of different types, either compressed or uncompressed, have redundant information, and imply different file sizes.

For each dataset of the use cases, it will be specified the following information:

Dataset identifier	*Identifier of the dataset*
Description	*Description of the dataset in relation to the project*
Partners	*The partners of the consortium related to the pilot*
Data types	*The type of data that will be used*
File formats	*The format of the files that will be used*
Type of data	*If the data is personal, confidential, synthetic, etc.*
Data production methods	*How the data will be generated in the use case*
Expected size of data	*The different size of the data to be provided*
Objective of the use of data	*How the data will be used in the project*
Potential for reuse in the project	*If the data can be reused for other activities in the project*
Dissemination in the project	*How the data will be shared among the project partners*
Open Access	*Whether the data will be made open access*

3. Data Summary

Here we define the different data sources (datasets) we will have in the project. We describe in each subsection the information regarding each of the pilots and the data they will generate following the template described in the previous subsection.

3.1 Apprentices Pilot in Switzerland

Dataset identifier	DT_SWZ
Description	This dataset describes the type of data provided in the pilot in Switzerland for vocational school and MSEs. The data will be used for the validation and demonstration of the GEIGER solution with its different components and services, focusing on technical and educational aspects.
Partners	BBB, SKV, HAAKO, E-ABO, CL Other: Swiss national CERT NSCS, third-party MSEs who consent to participate in the GEIGER research.
Data types	Technical: information about each MSE and its systems for protection, recommendations, incident reports, and information and feedback about the performance of the GEIGER framework. The GEIGER framework also includes the component for information exchange with external entities and the GEIGER indicator. Educational: security defender education material, education status, and feedback from end-users.
File formats	Technical: information collected, processed, and generated by the tools. The format will follow the one specified by each tool owner for their tool. The formatting will be text and be both human and machine-readable, e.g. XML. Documents: training courses and feedback. The formatting will follow those of leading Office solutions, e.g. Microsoft Office, Adobe PDF, and jpeg or PNG, mp3, and mp4-encoded rich media.
Type of data	Technical: real and synthetic data Educational: real data
Data production methods	The use case partners and third-party MSEs that consent to participate in the GEIGER research will produce real data; the technical partners may extend that data with synthetic data. Technical: the MSEs will provide a simulation of their day-to-day environments to have a good representation of their systems and cybersecurity practices. The technical partners may simulate attacks in these systems together with the MSEs to check the response of the GEIGER solution. Educational: anonymous feedback for improving the courses and certification status of security defenders.
Expected size of data	To be specified at validation and demonstration planning
Objective of the use of data	The data will be used for validating and demonstrating the GEIGER solution and for improving the GEIGER framework and the Security Defenders educational material.

Potential for reuse in the project	The data could be used for improving all aspects of GEIGER (technical and educational) at different stages of the project. It could also be used for disseminating results of the project to stakeholders.
Dissemination in the project	Technical: the data will be provided in a protected and controlled environment. The data subjects define the ecosystem entities that may access the data, including technical partners and national CERT. The data will be stored in our secure repository and controlled its access. Documents: text will be stored and managed in our secure and controlled repository. Data access is restricted only to the consortium, and the usage of the data will be only for the abovementioned reasons. If to be used for any other reason, the consortium will be informed and decide according to it.
Open Access	Due to the confidential nature of the data about companies and security defenders education, open access will not be provided.

3.2 Accountants Pilot in the Netherlands

Dataset identifier	DT_NTH
Description	This dataset describes the type of data provided in the pilot in The Netherlands involving the education of accountants and dissemination of the GEIGER solution by these accountants. The data will be used for the validation and demonstration of the GEIGER solution with its different components and services, focusing on technical and educational aspects.
Partners	SRA
Data types	Technical: information about each MSE and its systems for protection, recommendations, incident reports, and information and feedback about the performance of the GEIGER framework. The GEIGER framework includes the GEIGER indicator. Educational: security defender education material, education status, and feedback from end-users.
File formats	Technical: information collected, processed, and generated by the tools. The format will follow the one specified by each tool owner for their tool. The formatting will be text and be both human and machine-readable, e.g. XML. Documents: training courses and feedback. The formatting will follow those of leading Office solutions, e.g. Microsoft Office, Adobe PDF, and jpeg or PNG, mp3, and mp4-encoded rich media.
Type of data	Technical: real and synthetic data Educational: real data
Data production methods	The data will be generated and provided by the use case partner. Technical data may be generated by third-party MSEs that consent to participate in the GEIGER research. Technical: the MSEs will provide a simulation of their day-to-day environments to have a good representation of their systems and cybersecurity practices. The

	technical partners may simulate attacks in these systems together with the MSEs to check the response of the GEIGER solution. Educational: anonymous feedback for improving the courses and certification status of security defenders.
Expected size of data	To be specified at validation and demonstration planning
Objective of the use of data	The data will be used for validating and demonstrating the GEIGER solution and for improving the GEIGER framework and the Security Defenders educational material.
Potential for reuse in the project	The data could be used for improving all aspects of GEIGER (technical and educational) at different stages of the project. It could also be used for disseminating results of the project to stakeholders.
Dissemination in the project	Technical: the data will be provided in a protected and controlled environment. The data subjects define the ecosystem entities that may access the data, including technical partners and national CERT. The data will be stored in our secure repository and controlled its access. Documents: text will be stored and managed in our secure and controlled repository. Data access is restricted only to the consortium, and the usage of the data will be only for the abovementioned reasons. If to be used for any other reason, the consortium will be informed and decide according to it.
Open Access	Due to the confidential nature of the data about companies and security defenders education, open access will not be provided.

3.3 Entrepreneurs Pilot in Romania

Dataset identifier	DT_ROM
Description	This dataset provides the information that will be provided by the Romanian pilot, including the MSE association, national CERT, and MSEs. The data will be used for the validation and demonstration of the GEIGER solution with its different components and services, focusing on technical and educational aspects.
Partners	CLUT-IT, CERT-RO, SCB, PT Other: third-party MSEs who consent to participate in the GEIGER research.
Data types	Technical: information about each MSE and its systems for protection, recommendations, incident reports, and information and feedback about the performance of the GEIGER framework. The GEIGER framework also includes the component for information exchange with the national CERT and the GEIGER indicator. Educational: security defender education material, education status, and feedback from end-users.
File formats	Technical: information collected, processed, and generated by the tools. The format will follow the one specified by each tool owner for their tool. The formatting will be text and be both human and machine-readable, e.g. XML.

	Documents: training courses and feedback. The formatting will follow those of leading Office solutions, e.g. Microsoft Office, Adobe PDF, and jpeg or PNG, mp3, and mp4-encoded rich media.
Type of data	Technical: real and synthetic data Educational: real data
Data production methods	The use case partners and third-party MSEs that consent to participate in the GEIGER research will produce real data; the technical partners may extend that data with simulated or synthetic data. The feedback of educational data will be provided by the partners of the consortium always following anonymisation techniques.
Expected size of data	To be specified at validation and demonstration planning
Objective of the use of data	The data will be used for validating and demonstrating the GEIGER solution and for improving the GEIGER framework and the Security Defenders educational material. Other data will be used for enhancing other aspects such as usability of the dashboard, results of the training material, etc.
Potential for reuse in the project	The data could be used for improving all aspects of GEIGER (technical and educational) at different stages of the project. It could also be used for disseminating results of the project to stakeholders.
Dissemination in the project	Technical: the data will be provided in a protected and controlled environment. The data subjects define the ecosystem entities that may access the data, including technical partners and national CERT. The data will be stored in our secure repository and controlled its access. Documents: text will be stored and managed in our secure and controlled repository. Data access is restricted only to the consortium, and the usage of the data will be only for the abovementioned reasons. If to be used for any other reason, the consortium will be informed and decide according to it.
Open Access	Due to the confidential nature of the data about companies and security defenders education, open access will not be provided.

3.4 Archiving and Preservation

We plan to store all data in the repository of the project. The repository is a protected system with restricted access control that can only be accessed by the members of the consortium. It also provides functionality for retrieving, modifying, and deleting data and logging who accesses the data.

The data will be maintained until the end of the project. The consortium will decide at the end of the project about potential further use of the data. Such further use will be based on returning the data from the project to the data subjects, allowing these subjects to decide about their future use of the data according to the rules of the GDPR.

4. Summary and Conclusions

This data management plan (DMP) is based on the characteristics of the GEIGER project. The DMP considers the generation and usage of the data of each pilot, and the type of that data. The DMP describes the data that will be generated in each pilot and how that data will be used.

The DMP reports that the generation and use of personal information will be minimised in the pilots, using synthetic data as an alternative where appropriate.

The DMP complements the ethics deliverables D7.1 H and D7.2 POPD. The ethics deliverables further describe the procedures and consent forms for involving human participants in the GEIGER research and managing personal data. In particular, D7.2 will describe the handling of personal and confidential data and include as an annexe the non-disclosure agreement to be signed by third-parties like national CERTs and MSEs that will collaborate for GEIGER validation and demonstration.

The DMP relates to the WP1 deliverables D1.1 Requirements and D1.2 Architecture. The WP1 deliverables will refine the definition of the information processed by the GEIGER solution for the pilots, refining the data definitions in this DMP. D1.2 will describe the technical solution for managing the technical data in the protected repository of the GEIGER solution.

The DMP will be used as an input for WP4 GEIGER Validation and Demonstration. The validation and demonstration report deliverables D4.1 and D4.2 may refine the data and procedures described by the DMP to account for the details of the GEIGER research study protocols. The data to be provided at those stages will be managed by each data producer and protected as specified.

When necessary, the DMP will be updated to cover emerging needs or constraints from the use cases' technical, architectural, or functional point of views. Any such update intends to address the needs of the project and provide security and privacy of the data managed on it. Updates about the data management plan will be provided as additional information in the WP4 deliverables that will describe the plan and report of GEIGER validation and demonstration. Any confidential information related to data management will be stated in the periodic reporting documents of the project as they are also confidential and, that way, we will be able to continuously adapt the plan to the needs of the projects and any future opportunity that may happen.