



DS-06-2017: Cybersecurity PPP: Cryptography

PRIVILEGE

Privacy-Enhancing Cryptography in Distributed Ledgers

### D6.5 – Data Management Plan

Due date of deliverable: 29 June 2021

Actual submission date: 29 June 2021


Grant agreement number: 780477

Lead contractor: Guardtime OÜ

Start date of the project: 1 January 2018

Duration: 42 months

Revision 2.0

	Project funded by the European Commission within the EU Framework Programme for Research and Innovation HORIZON 2020	
Dissemination Level		
PU = Public, fully open		X
CO = Confidential, restricted under conditions set out in the Grant Agreement		
CI = Classified, information as referred to in Commission Decision 2001/844/EC		

## **D6.5 – Data Management Plan**

### **Editor**

Liis Livin (Guardtime OÜ)

### **Contributors**

Ahto Truu (Guardtime OÜ)

Kristo Klesment (Guardtime OÜ)

Sven Heiberg (SCCEIV)

Nikos Voutsinas (GUNet)

Nikos Karagiannidis (I.O. Research)

29.06.2021

Revision 1.1

*The work described in this document has been conducted within the project PRIViLEDGE, started in January 2018.*

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 780477.*

*The opinions expressed and arguments employed herein do not necessarily reflect the official views of the European Commission.*

*©Copyright by the PRIViLEDGE Consortium*

1. Introduction	1
2. Data Summary	2
2.1 Online voting use-case “iVoting”- Verifiable online voting with ledgers	2
2.2 Health insurance use-case - Distributed ledger for insurance	4
2.3 Diplomas use-case - University diploma record ledger	7
2.4 Decentralized software updates use-case - Update mechanism for Cardano stake-based ledgers	10
2.5 Data related to stakeholder interviews	11
3. General principles	14
3.1 Participation in the Pilot on Open Research Data	14
3.2 Intellectual Property Rights Management and Security	15
4. Allocation of resources	16
5. Data security	16
6. Ethical aspects	16

# 1. Introduction

**The purpose of this Data Management Plan is to provide general principles on how PRIViLEDGE project partners handle collecting and maintaining data.** The main goal of PRIViLEDGE project is to realise cryptographic protocols supporting privacy, anonymity, and efficient decentralised consensus for distributed ledger technologies. PRIViLEDGE's work is guided by four use-cases in different areas: online voting (lead by SCCEIV), health insurance (lead by GT), university diplomas (lead by GRNet and GUNet) and decentralized software updates (lead by I.O. Research). The selected use-cases and related prototypes are diverse and represent the principal application domains of distributed ledger technology (DLT) to which the PRIViLEDGE's technology is integrated into.

The PRIViLEDGE use-cases utilise synthetic data to analyse the functionality of the specific implementation and do not use or release any real-life data with identities. Thus, we surmise that the data(-sets) utilised for testing will not create benefits for other projects or entities on a practical level. Nevertheless, this document specifies the synthetic data-sets used for the use-cases and describes potential data policies that might become applicable if the prototypes are launched into real-life solutions in sections 2.1 - 2.4. The following section 2.5 describes the data handling in dissemination activities, more precisely for the data gathered during the stakeholder interviews. The rest of the document describes general principles the consortium adheres to in relation to open access and IPR (section 3), followed by resource allocation description (section 4), and data security and ethical aspects in the sections 5 and 6.

## 2. Data Summary

This chapter describes the datasets that PRIViLEDGE’s four use-cases and related prototypes use, and data that was acquired during the stakeholder interviews performed for one PRIViLEDGE workshop.

The use-case leads manage their data procedure independently and as mentioned above, the prototypes only use synthetic data. The described future data policy is dependent on the specific potential implementation area, as well as the international, national and/or corporate regulatory framework surrounding it.

### 2.1 Online voting use-case “iVoting”- Verifiable online voting with ledgers

The table below describes the synthetic data-set used by online voting use-case.

<b>Dataset name</b>	<b>Test data</b>
Dataset description	Test election configurations for 15000, 20000, 25000 and 30000 voters.
Dataset status	The dataset is generated on-demand by the CI infrastructure.
Responsible entity for dataset	SCCEIV
<b>Security and privacy considerations</b>	
None, since the data is auto-generated and synthetic.	
<b>Data release plan</b>	
This data shall not be released as it can always be re-generated.	
<b>Data type</b>	<b>Election configuration</b>

Data description	Data to populate voter lists, candidate lists and lists of election administrators with respective credentials.		
Purpose of the data	To support semi-automatic end-to-end testing of the Tivledge prototype.		
Maintenance and aggregation of data	Temporary data, not persistent, not maintained.		
Relation to project objective	Necessary to test the UC1 pilot voting system.		
File types	JSON files with proprietary schema		
(Data provider) Origin of the data	Size (xByte)	Access for Partners	Access for the public
N/A	Variable	None	None

**In the future the iVoting use-case will be implemented through the TIVI voting platform** used for online voting for binding electoral events. The purpose of the platform is to make it possible to hold elections, find out the voting result and audit the correctness of the voting result in the context of ballot secrecy.

- The data is located in the AWS environment, in the <AWS region> region and in the respective infrastructure of any participant in the ledger.
- The platform processes personal data on an election basis. In the context of a single election, only the personal data of those and only those users is processed on the basis of
  1. Those who have the right to vote in the given election or
  2. Those who try to vote in the given election even though they do not have the right to vote.

The platform allows multiple elections to be held at the same time, in which case personal data related to different elections is managed in a logically separate manner.

- The platform processes the personal data that is essential for organizing elections with electronic voting.
  1. Unique identifiers of all persons with the right to vote to interface with the identification service of the <Organization>.
  2. The corresponding name for each unique identifier.
  3. Encrypted and digitally signed ballot.
  4. Cryptographic commitments of the ballots on the ledger.

- 5. The fact if the ballot was revoked or sent to tally.
- The platform processes personal data that is inevitably collected during voting.
  - 1. For each authenticated user, regardless of their voting rights:
    - o the outcome of authentication,
    - o time of access,
    - o unique identifier,
    - o name,
    - o IP address and
    - o identification information provided by the web browser.
  - 2. For each digitally signed and encrypted ballot
    - o the time of retrieval.
- In the context of ballot secrecy, the following is emphasized: although the platform stores information on voters, their voting times, possible re-voting and the voting result, the platform does not permit it as possible to reconcile the identity of a particular voter with his or her plain-text vote preference. The platform fully complies with the requirement to ensure ballot secrecy. Personal data will be stored for up to 1 year from the moment of inclusion in the voter list.
- Personal data will only be transferred to third parties in the event of a legal obligation and on the basis of an official formal notice. The personal data processed will not be transferred outside the European Union or the European Economic Area (EEA) through the platform.
- The platform's technical solution protects data from unauthorized access, modification, disclosure, removal or infringement. To ensure data security, personal data is treated as confidential, only encrypted communication is used, a cryptographical voting protocol is used that guarantees the secrecy of the vote, access to personal data is restricted to those employees and contractors who need this information to process it and who are subject to contractual confidentiality obligations, and the personal data repository is protected by the necessary IT technical and organizational protection measures.

## 2.2 Health insurance use-case - Distributed ledger for insurance

The tables below describe the synthetic data-sets used by health insurance use-case.

Dataset name	<b>FHIR test data</b>
--------------	-----------------------

D6.5 - Data Management Plan

Dataset description	Data to populate the FHIR database for integration tests.		
Dataset status	Pseudo-randomly generated synthetic test data.		
Responsible entity for dataset	Guardtime		
<b>Security and privacy considerations</b>			
None, since this is pseudo-randomly generated synthetic test data.			
<b>Data release plan</b>			
This data shall not be released.			
<b>Data type</b>	<b>FHIR records in JSON format</b>		
Data description	FHIR records generated pseudo-randomly using the Synthea model and generation scripts ( <a href="https://synthetichealth.github.io/synthea/">https://synthetichealth.github.io/synthea/</a> ).		
Purpose of the data	To populate the FHIR database for integration tests.		
Maintenance and aggregation of data	Temporary test data. Will not be maintained.		
Relation to project objective	Testing of the application prototype.		
File types	FHIR records in JSON format.		
(Data provider) Origin of the data	Size (xByte)	Access for Partners	Access for the public
N/A	Variable	No	No
<b>Dataset name</b>	<b>MPC test data</b>		



D6.5 - Data Management Plan

Dataset description	Data to feed into the MPC protocol for performance tests		
Dataset status	Randomly generated test data.		
Responsible entity for dataset	Guardtime		
<b>Security and privacy considerations</b>			
None, this is randomly generated test data.			
<b>Data release plan</b>			
This data shall not be released.			
<b>Data type</b>	<b>Sets of randomly generated integer values</b>		
Data description	Sequences of tuples of randomly generated integer values.		
Purpose of the data	To feed into the MPC protocol for performance tests.		
Maintenance and aggregation of data	Temporary test data. Will not be maintained.		
Relation to project objective	Testing of the application prototype.		
File types	Text files.		
(Data provider) Origin of the data	Size (xByte)	Access for Partners	Access for the public
N/A	Variable	No	No

**In the future, when the verifiable reporting features are integrated into Guardtime’s offerings for the health insurance market, the underlying patient data will be managed by the healthcare**

service providers who already process this data as part of their normal operations. Only the following data types will leave the premises of the original data holders:

- Cryptographic commitments on sets of patient records. The commitment scheme used is unconditionally hiding and thus the commitments should not be considered PII under the GDPR. Applicability of additional local data protection regulations needs to be evaluated separately for each new market when the product is introduced.
- Cryptographic shares of input data in the multi-party computation protocol. The sharing scheme used is unconditionally hiding and thus the shares should not be considered PII under the GDPR. Applicability of additional local data protection regulations needs to be evaluated separately for each new market when the product is introduced.
- Cryptographic proofs of correctness of the multi-party computations. The proof scheme used is perfectly zero-knowledge and thus the proofs should not be considered PII under the GDPR. Applicability of additional local data protection regulations needs to be evaluated separately for each new market when the product is introduced.
- Aggregate values included in the generated reports. The structure and dissemination of these reports will be approved by all participants according to their local patient privacy regulations before the reports are implemented.

### 2.3 Diplomas use-case - University diploma record ledger

The table below describes the synthetic data-set used by diplomas use-case.

Dataset name	<b>Diploma Templates</b>
Dataset description	Diploma templates hold the static data of the corresponding certificates such as the issuing institution(s) and department(s) and the details of the curricula.
Dataset status	Synthetic data were used that follow the schemas and the semantics of actual diplomas.
Responsible entity for dataset	GUNET
<b>Security and privacy considerations</b>	
None to consider since synthetic data were used.	

D6.5 - Data Management Plan

Data release plan			
No plans to release the dataset.			
Data type	<b>Diploma template sets in JSON format</b>		
Data description	Diploma templates of fictitious undergraduate courses in JSON format.		
Purpose of the data	The use-case prototype joins the diploma's template (reference data) with the holder's personal information and the certificate specific data (grade, nomination date etc), to construct the diploma's object.		
Maintenance and aggregation of data	There is no need for a maintenance plan, considering that data are being used for testing purposes only.		
Relation to project objective	To provide a realistic dataset for the design, implementation, and demonstration of the use-case prototype.		
File types	JSON structure		
(Data provider) Origin of the data	Size (xByte)	Access for Partners	Access for the public
N/A	Variable	No	No

Dataset name	<b>Diploma Data</b>		
Dataset description	Dataset of holders' personal information along with the corresponding certificate specific data (grade, nomination date etc).		
Dataset status	Synthetic data were used that follow the schemas and the semantics of actual diplomas.		
Responsible entity for dataset	GUNET		
<b>Security and privacy considerations</b>			

None to consider since synthetic data were used.			
<b>Data release plan</b>			
No plans to release the dataset.			
<b>Data type</b>	<b>Diploma sets in JSON format</b>		
Data description	Fictitious diplomas dataset consisting of holder’s personal information and diploma’s data in JSON format.		
Purpose of the data	The use-case prototype joins the diploma’s template (reference data) with the holder’s personal information and the diploma’s data (grade, nomination date etc), to construct the diploma’s object.		
Maintenance and aggregation of data	There is no need for a maintenance plan, considering that data are being used for testing purposes only.		
Relation to project objective	To provide realistic dataset for the design, implementation, and demonstration of the use-case prototype.		
File types	JSON structure		
(Data provider) Origin of the data	Size (xByte)	Access for Partners	Access for the public
N/A	Variable	No	No

**Now and in the future, the Higher Education Institutes, being the issuers of the diplomas, are anticipated to keep their authoritative role over the diplomas data** and maintain the existing processes designed to protect the privacy of personal data, as GDPR dictates.

The goal for the use-case is to enhance the security aspects of the diploma issuing and verification process in the digital realm by leveraging cryptographic technologies. The incorporation of DLTs will not alter the existing data management plans and will not be used to hold PII. DLTs however will be used to enhance the immutability of the digital records and the transactions that take place in the process of issuing and verifying credentials and will convey the corresponding cryptographic proofs between the issuer, the holder and the relaying (verifying) party.

In that context the solution attempts to expand the diploma holders’ control over their data. The plan for the future is to provide minimal and selective disclosure of information and prevent the issuer from tracking the holder’s activity.

## 2.4 Decentralized software updates use-case - Update mechanism for Cardano stake-based ledgers

The table below describes the synthetic data-set used by decentralized software updates use-case.

Dataset name	<b>Test protocol updates data</b>
Dataset description	Randomly generated update events used in our property-based testing framework that validated the update mechanism built.
Dataset status	Not stored. Deleted after each test run.
Responsible entity for dataset	I.O. Research
<b>Security and privacy considerations</b>	
These are synthetic data meaningless from a privacy and security perspective. No personal identification is possible. These are only useful for testing the “update logic” of the prototype, i.e., the correct state transitions of the maintained update state based on submitted update events.	
<b>Data release plan</b>	
The data-set will not be released to the public.	
Data type	<b>Hashes, public keys, numbers, text</b>
Data description	Hashes of imaginary update proposals Public keys of imaginary stakeholders in a testnet Minimum info per proposal to test update logic (voting period duration, version dependency, priority).

Purpose of the data	To validate the update logic implemented by the update system.		
Maintenance and aggregation of data	Data are discarded after each test run.		
Relation to project objective	Required for the validation of the prototype.		
File types	In-memory random generated data from Quickcheck haskell library <a href="https://hackage.haskell.org/package/QuickCheck">https://hackage.haskell.org/package/QuickCheck</a>		
(Data provider) Origin of the data	Size (xByte)	Access for Partners	Access for the public
N/A	N/A	No	No

**In the future, this mechanism will influence the protocol update mechanism of the Cardano *public* blockchain.** It is in the nature of all public blockchains to enable public immutable data and this is the main policy followed for the update payload carried within transactions as well.

We need to point out that there are *no personal data involved* in the whole updating process. Only public keys are involved which issue the corresponding transactions. Moreover, one of the main requirements for this update mechanism has been transparency and auditability, which is a prerequisite for true decentralization. This means that anyone should be able to answer *when, why* and *how* the blockchain has evolved the ways it has. Therefore, we do not use privacy in our voting process, although voting takes place through public keys (i.e., aliases). Therefore, the tally of results can be verified at any time by anyone since they are immutably stored in the blockchain.

Moreover, since we propose a liquid democracy scheme via delegation to experts, transparency is also necessary for the accountability of the delegates. All in all, since our mechanism is embedded within a public blockchain, essentially as a specialized validation layer for update payload and we strive for true decentralization and self-sustainability of the blockchain, we follow the data policies common to all public blockchains, i.e., open data and use of aliases with zero processing and storage of personal data.

### 2.5 Data related to stakeholder interviews

This section describes the data collected during PRIVILEGE workshop (March-April 2021) which consisted of 15 expert interviews with stakeholders. The aim of the workshop was to investigate

## D6.5 - Data Management Plan

PRIVILEGE's use-cases' suitability for determined application domain and potential users, find matches/mismatches from the value propositions prepared for the end-users, and to establish mutually beneficial and sustainable relationships with the interviewees.

Within this workshop 4 interviews were conducted by iVoting use-case (UC1), 4 interviews by health insurance use-case (UC2), 3 interviews by Diplomas use-case (UC3), and 4 by decentralized software updates use-case (UC4).

The following data-sets were created through this action:

- In total 15 interview transcripts.
- Four videos for UC1 interviews.

The following tables describe how the data collected and created through the interviews are handled. The original data (transcripts and videos) of the interviews remains private to the public and other partners<sup>1</sup>. The generalized results of the interviews have been published as a [report](#) where the analysed information is presented in an aggregated and personal information perceiving manner.

Dataset name	Open-ended interviews
Dataset description	15 transcriptions of the open-ended interviews.
Dataset status	Private
Responsible entity for dataset	SCCEIV (4 interview transcripts), GT (all interview transcripts), GUNET (3 interview transcripts), I.O.Research (4 interview transcripts).
<b>Security and privacy considerations</b>	
The interview transcripts include personal and potential commercial information which is considered private. Therefore, this data-set is not publishable in its raw form. Only conclusions will be published with the consent of the interviewees.	
<b>Data release plan</b>	

---

<sup>1</sup> Besides the responsible person from each use-case who conducted the interviews for his/her use-case and the WP5 lead (GT) who compiled the report and had access to all transcripts.

D6.5 - Data Management Plan

The data-set will not be released to the public.			
<b>Data type</b>	<b>Qualitative, textual</b>		
Data description	Transcription text in word or pdf format, including both questions and answers of the interview.		
Purpose of the data	To reflect answers to stated interview questions. To draw conclusions relevant to use-cases exploitation.		
Maintenance and aggregation of data	SCCEIV, GT, GUNET, I.O. Research. Each interviewer stores their conducted interview's transcripts individually. GT stores all interview transcripts until the project is finished.		
Relation to project objective	To investigate suitability and applicability of PRIVILEGE'S use-cases among determined stakeholders.		
File types	Word, pdf.		
(Data provider) Origin of the data	Size (xByte)	Access for Partners	Access for the public
Interviewers	Variable	No	No

<b>Dataset name</b>	<b>Four UC 1 videos</b>		
Dataset description	4 interviews in MP4 format, including capture of the full interview session.		
Dataset status	Private		
Responsible entity for dataset	SCCEIV		
<b>Security and privacy considerations</b>			
The MP4 videos include personal and potential commercial information which is considered private. Therefore, this data-set is not publishable.			



<b>Data release plan</b>			
The data-set will not be released to the public.			
Data type	<b>Recorded audio-visual data</b>		
Data description	MP4 format, including capture of the full interview session.		
Purpose of the data	Capture the discussion for easier transcription.		
Maintenance and aggregation of data	SCCEIV stores the videos for the period of deliverable preparation.		
Relation to project objective	To investigate suitability and applicability of PRIVILEGE'S use-cases among determined stakeholders.		
File types	MP4		
(Data provider) Origin of the data	Size (xByte)	Access for Partners	Access for the public
Interviewer	~300MB per session	No	No

### 3. General principles

#### 3.1 Participation in the Pilot on Open Research Data

The PRIVILEGE project participates in the Pilot on Open Research Data launched by European Commission along with the Horizon 2020 programme. The consortium strongly believes in the concepts of open science, and in the benefits that the European innovation ecosystem and economy can draw from allowing reusing data at a larger scale. Therefore, all data produced by the project can potentially be published with open access – though this objective will obviously need to be balanced with the other principles described below.

## 3.2 Intellectual Property Rights Management and Security

Special attention has been given to knowledge management and protection issues from the beginning, and during the whole lifetime of the project. All details regarding management and protection of knowledge created within PRIViLEDGE is specified in the Consortium Agreement (CA), following the DESCA Horizon 2020 document. The CA addresses (a) confidentiality of the information disclosed by partners during the project, ownership of results resulting from the execution of the project, (b) legal protection of results resulting from the execution of the project through patent rights, (c) commercial utilisation of results, also taking into account joint ownership of the results, management of evolution of results (new innovations added after the end of the project), (d) patents, know-how and information related to the use of knowledge owned by one of the partners, resulting from work carried out prior to the agreement, and (e) sublicenses to third parties within clearly defined limits.

The definition of the distribution of the EU funds and the Intellectual Property (IP) has been implemented to the CA from the Grant Agreement. The principal basis of the information and know-how exchange is free access rights. The general outline of the IP rules agreed by PRIViLEDGE partners is as follows:

- Partners' pre-existing knowledge (background) has been specified in the CA.
- Knowledge that is generated within PRIViLEDGE project shall remain the property of the partner that generated it. If more than one partner generates that knowledge and it is not possible to separate their contributions, the knowledge will be jointly owned.
- Access rights to knowledge that is needed by a partner for the execution of its part in PRIViLEDGE has been granted to the partner on a royalty-free non-transferable basis.
- A partner will not publish any knowledge provided by another partner and identified as confidential, without the other partner's prior written approval. However, if open source software licenses apply, the open source software license rules will apply for publishing knowledge.
- To meet the need of both industrial partners with commercial and IP interests and research partners in the project, which have a major role in ensuring results are widely disseminated, dissemination assets will be submitted to the Project Coordinator and the relevant WP leader for dissemination and distributed to other relevant partners who may object within a small time period which is agreed upon in the CA. Otherwise the dissemination may proceed.
- Each organisation is free to develop (and promote) its own innovative approaches introduced to PRIViLEDGE after the end of the project without limitations.

An IPR Repository with all software licenses used has been created to aid partners in identifying possible licensing issues, selecting the best licensing models for the software developed during the project and ensuring on the one hand that no IPR issues hinder the exploitation potential of the software assets produced during PRIViLEDGE, on the other hand that IPR is fully honoured. The relevant developments regarding IPR are periodically reviewed and communicated to all partners.

## 4. Allocation of resources

In general, collecting the data and storing would require both working hours and data storage that could create cost for the partners. Also, other things e.g. licensing might create costs. PRIViLEDGE allocated working hours to handle the synthetic data used by the use-cases, and for collecting and processing the data of the stakeholder interviews. Nevertheless, as PRIViLEDGE does not release any data, there are no potential future costs that we could consider and describe at this point.

## 5. Data security

Each use-case partner will be responsible for its own utilised data, including storage and data recovery. Nevertheless, as mentioned under section 2, the use-cases have no plan to maintain the synthetic data used for testing the prototypes. The original data collected in the form of transcripts and videos for the stakeholder interviews will only be maintained by each interviewer and GT until the successful finalization of the project and will be permanently deleted after that.

## 6. Ethical aspects

The PRIViLEDGE partners will comply with the GDPR legislation. Moreover, all the partners follow the ethical principles described in detail in ARTICLE 34 of the PRIViLEDGE Grant Agreement.