SCOREwater

Smart City Observatories implement REsilient Water Management

DELIVERABLE 3.5

FUNCTIONAL AND TECHNICAL GUIDELINES FOR IMPLEMENTATION OF PRIVACY ENHANCING TECHNOLOGIES WITHIN THE SCOREWATER PLATFORM

Dissemination level	Public
Type	Report
Issued by	CIVITY
Contributing project partners	Eurecat, Talkpool
Author(s)	Vanmeulebrouk, B., Rubion Soler, E.
Reviewed by	Sanne, J., Wilhelmsson, J., de Roover, S.
Keywords	GDPR, security, anonymisation
Number of pages	23
Number of annexes	1
Date:	2021-04-29
Version:	V 1
Deliverable number	3.5
Work Package number:	WP 3
Status:	Delivered
Approved by coordinator (IVL)	2021-04-30

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Copyright notices

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

REVISION HISTORY

Version	Reason for changes	Name	Date
1	Original release to EU	Bas Vanmeulebrouk	2021-04-30

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

CONTENT

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

LIST OF FIGURES

LIST OF TABLES

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

ABBREVIATIONS

Abbreviation	Definition
AES	the Advanced Encryption Standard
API	Application Programming Interface
GDPR	General Data Protection Regulation
ICT	Information and Communications Technology
IoT	Internet of Things
SDG	Sustainable Development Goals
SLL	Secure Sockets Layer
SME	Small and Medium-sized Enterprise
TLS	Transport Layer Security
VPN	Virtual Private Network

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

PROJECT ABSTRACT

SCOREwater focuses on enhancing the resilience of cities against negative effects of climate change and urbanization by enabling a water smart society that fulfils SDGs 3, 6, 11, 12 and 13 and secures future ecosystem services. We introduce digital services to improve management of wastewater, stormwater and flooding events. These services are provided by an adaptive digital platform, developed and verified by relevant stakeholders (communities, municipalities, businesses, and civil society) in iterative collaboration with developers, thus tailoring to stakeholders' needs. Existing technical platforms and services (e.g. FIWARE, CKAN) are extended to the water domain by integrating relevant standards, ontologies and vocabularies, and provide an interoperable open-source platform for smart water management. Emerging digital technologies such as IoT, Artificial Intelligence, and Big Data are used to provide accurate real-time predictions and refined information.

We implement three large-scale, cross-cutting innovation demonstrators and enable transfer and upscale by providing harmonized data and services. We initiate a new domain "sewage sociology" mining biomarkers of community-wide lifestyle habits from sewage. We develop new water monitoring techniques and data-adaptive storm water treatment and apply to water resource protection and legal compliance for construction projects. We enhance resilience against flooding by sensing and hydrological modelling coupled to urban water engineering. We will identify best practices for developing and using the digital services, thus addressing water stakeholders beyond the project partners. The project will also develop technologies to increase public engagement in water management.

Moreover, SCOREwater will deliver an innovation ecosystem driven by the financial savings in both maintenance and operation of water systems that are offered using the SCOREwater digital services, providing new business opportunities for water and ICT SMEs.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

EXECUTIVE SUMMARY

To support the different cases, SCOREwater collects data which raises the security issue of unauthorized access to personal data. When collecting personal information, the case has to explain why this data has to be collected. If collecting personal information is not needed for actual use cases, the personal information should not be collected at all (data minimisation). Personal data as such will not be uploaded to the SCOREwater platform. Data providers and cases must make sure data is properly anonymised before it is being uploaded to the platform. In case a dataset does contain personal information, they must use the methods described in this deliverable to anonymize the dataset.

The deliverable discusses the perturbation techniques data providers must use to properly remove personal information from datasets. According to the GDPR, data must be pseudonymised, which means any information which can be directly related to persons must be removed. Once data has been pseudonymised, it might still be possible to identify individuals by cross referencing the dataset with other datasets using quasi identifiers. Quasi identifiers are columns which both datasets have in common (for example a timestamp and a location). Anonymisation disables this option by either removing or generalizing these quasi identifiers. Another option to hide personal data is to aggregate data at a higher level. Aggregation has the disadvantage of potentially reducing the usability of the data for other purposes then the one the data were collected for.

Once the data have been properly anonymised by the data provider, they can be uploaded to the SCOREwater platform. Subsequently, it is up to the platform to make sure the data is properly delivered to the users of the data. The following security measures are implemented to prevent unauthorised access to devices used for collecting data (sensors) or actual data on the SCOREwater platform. Preventing unauthorized access to sensors is enforced by a) automatically verifying if the firmware (the software which is operating the device) has not been tampered with, b) securing communications from and to the sensor using certificates and c) close ports which are not being used. This is called sensor integrity.

Storing data on the SCOREwater platform in a safe and secure fashion entails using certificates to secure communication between client and server, not opening ports on servers which do not have to be exposed and making sure the database is installed on a separate server which cannot be accessed from the internet directly. Data in this database can be encrypted using either symmetric or asymmetric encryption. In case of symmetric encryption, the same key is used to encrypt and decrypt the data. As a consequence, this key must be exchanged via the internet, which poses a security risk. In case of asymmetric encryption, a public key is used to encrypt the data and a private key is used to decrypt the data. Only the public key has to be exchanged via the internet, but it is not a problem if this key is exposed (since the key can only be used to encrypt data and not to decrypt data), making asymmetric encryption a more secure option. Asymmetric encryption is much slower than symmetric encryption though.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

1. INTRODUCTION

Security and privacy issues regarding data attract a lot of attention these days. This is fuelled by recent leaks of privacy sensitive information, for instance the incident in which COVID19 test results from millions of Dutch people were leaked (Verhagen, 2021). The aim of this deliverable is to provide guidelines on how to deal with these issues within SCOREwater.

To support the different cases, SCOREwater collects data which raises the security issue of unauthorized access to personal data. When collecting personal information, the case has to explain why this data has to be collected. If collecting personal information is not needed for actual use cases, the personal information should not be collected at all (data minimisation). Personal data might be collected within the frame of the Amersfoort (data from the thermal walks) and Barcelona (health information extracted from sewage samples and data from questionnaires) cases.

According to the grant agreement and a confidential SCOREwater deliverable from the ethics work package, personal data will not be uploaded to the SCOREwater platform. Datasets containing personal information must be anonymised before actually being uploaded to the SCOREwater platform. This deliverable describes how this should be properly done, it provides data providers and cases with a toolbox containing tools to remove personal information from datasets. These tools should be applied when onboarding a dataset on the SCOREwater platform. In addition to this toolbox, the deliverable discusses the security measures implemented to prevent unauthorised access to the data on the SCOREwater platform or the devices used to collect the data (sensors).

The deliverable starts by discussing the legal framework for using data, the General Data Protection Regulation (GDPR) in chapter 2. Chapter 3 focusses on perturbation techniques which can be used to remove personal information from datasets. These perturbation techniques are the tools data providers have at their disposal to make sure a dataset is properly anonymised before it is being uploaded to the SCOREwater platform. Chapter 4 discusses how to make sure data are secure once they have been made available via a platform such as the SCOREwater platform.

The deliverable focuses on privacy issues and privacy related technical aspects of sharing data. It does not focus on ethical aspects related to using the data: that is up to the actual users of data.

2. LEGAL FRAMEWORK FOR USING DATA

The main legal regulation to take into consideration for a data platform such as the SCOREwater platform is the General Data Protection Regulation (GDPR).

2.1. GDPR

Currently, the European Union allows the free movement on non-personal data within the EU for companies and public authorities. Nevertheless, the personal data are operated under the General Data Protection Regulation (GDPR). The GDPR is effective since May 25, 2018. When it came into effect, it replaced the 1995 Data Protection Directive. Both the Directive and the GDPR are comprehensive data privacy regulations. The GDPR imposes many enhanced privacy protections, including key anonymisation requirements.

Article 4 of the GDPR defines personal data as “any information relating to an identified or identifiable natural person”. Moreover, the GDPR expands the definition of personal data to “one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person”, that is, include all information that could be used to indirectly identify individuals. Some examples are ID numbers, HR records, IP addresses, health records, biometrics, among others.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

The GDPR explicitly recommends pseudonymisation of personal data as one of several ways to reduce risks from the perspective of the data subject. The term "pseudonymisation" is defined under the GDPR as "The processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person". In short, pseudonymisation does not prevent re-identification, but it is a security measure used to reduce the linkability of data to its data subjects by separating the data from direct identifiers.

3. DATA PERTURBATION TECHNIQUES

Before it is decided to start storing privacy sensitive information, it should be considered whether it is actually needed to store the sensitive information. This is called data minimisation. The less data there is to collect, store and share, the easier it becomes to protect the data (Youyang Qu, 2019). If it is not needed to be able to fulfill current or future use cases, not storing privacy sensitive information alleviates you of any issues related to it. If current or future use cases do require storing of privacy sensitive information, there are various techniques which can be applied to make sure sensitive information does not fall into the wrong hands.

Data perturbation is the data security technique that modifies the data in order to preserve the privacy and confidentiality. Using a data perturbation procedure, data is changed in certain ways to disguise the sensitive information while preserving relevant data properties needed to actually use the data. Some of the commonly used perturbation techniques are (Thanga Revathi S, 2017):

1) Reconstruction, e. g. randomization or aggregation

2) Anonymisation, e. g. using k-anonymity

3) Data encryption

3.1. RECONSTRUCTION

Protection against reconstruction refers to the process where the dataset is being modified in order to hide privacy sensitive information. The sensitive data are transformed or masked by adding additional data to the original data.

3.1.1. RANDOMIZATION

Randomized data distortion techniques attempt to hide sensitive data by randomly modifying the data values by adding additional noise to mask the data for preserving the privacy of sensitive data. Random noise is added to the data in such a way that the individual data values are distorted while preserving the underlying distribution properties at dataset level.

Using the Value Distortion approach, not the actual value for a cell is returned, but the actual value plus a random value. This random value is drawn from a certain distribution. Commonly used distributions include the uniform distribution over an interval from minus a certain value to plus the same value or a Gaussian distribution with a mean equal to zero and a certain standard deviation.

3.1.2. AGGREGATION

Data aggregation is the process where the raw data points which are being collected are not being published using individual records, but is gathered and expressed in a summary form for statistical analysis instead. For example, raw data points can be aggregated over a given time period or spatial unit (for instance a municipality or a neighborhood) to provide statistics such as average, standard deviation, minimum, maximum, sum and count. The aggregated data should provide insights into the phenomenon under investigation. There are three types of data aggregation:

- Category aggregation: combines all data points for a certain resource into different categories;
- time aggregation: combines all data points for a certain resource over a specified time period;

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

- spatial aggregation: combines all data points for a certain resource over a specific geographical unit.

These different aggregation types can be combined obviously, by summarizing data points for a certain time period for a certain spatial unit for example. When applying data aggregation, make sure that there is a sufficient amount of data points not to disclose sensitive information. For example, if income information from a census is aggregated by calculating the average income per profession on a municipality level, the sensitive information will still be revealed if there is only one person in the municipality with a certain profession. A disadvantage of aggregation is that it becomes difficult to further process the data. Once the average has been calculated, and additional data comes in, one always needs the original data to calculate new averages including the newly collected data.

In the Barcelona case, SCOREwater manages health information extracted from the sewage samples. This information is aggregated at neighborhood level. As a consequence, sensitive information cannot be extracted from this dataset.

3.2. ANONYMISATION

Whereas aggregation combines multiple individual records into one new record (losing the individual record), anonymisation preserves the individual record but makes sure the record can no longer be traced back to a person. Anonymisation is a two-step process: pseudonymisation refers to removing personal data from the dataset and anonymisation refers to preventing identification of persons by cross referencing the dataset with other datasets.

3.2.1. PSEUDONYMISATION

Pseudonymisation is the first step. It entails removing personal data from the dataset that can be attributed to a specific data subject without the use of supplementary data. Examples of data which should be removed to pseudonymise a dataset include names, addresses, email addresses, phone numbers, social security numbers and birth dates. As discussed in paragraph 2.1, the GDPR requires pseudonymisation. But a dataset which has been pseudonymised can still be used to identify individuals if the dataset contains information which can be used to link the dataset to other datasets which do contain personal information. Anonymisation refers to removing this information from the dataset as well. K-anonymity is a widely used method to anonymise data.

3.2.2. K-ANONYMITY

The purpose of k-anonymity is to produce a release of a dataset in which it is impossible to identify individuals while the data can still be used (Sweeney, 2002). In other words: the sensitive information which makes the dataset interesting for research purposes for instance must be present, but this sensitive information cannot be traced back to individuals. Even though names, phone numbers, email addresses and other information which could be used to identify individuals are removed from a dataset, the dataset might still contain information which could be used to identify individuals when cross referencing the dataset with other datasets. If the dataset contains for instance a location and a timestamp, it might be possible to identify individuals by cross referencing the dataset with a dataset from a telecom provider.

Columns that can be used to distinguish between the different records in the dataset are called quasi-identifiers. In themselves, these quasi-identifiers do not have to be sensitive, but they might be used to re-identify individuals in a different dataset. K refers to number of quasi-identifiers we find acceptable. Each quasi-identifier occurs at least k times in a dataset with k-anonymity. A dataset has 5-anonymity when each quasi-identifier occurs at least five times in a dataset. The choice for a value of k is arbitrary. The bigger the number of records in a dataset, the easier it becomes to achieve the desired level of k-anonymity.

Two methods are used to achieve k-anonymity:

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

- Suppression: replace all or certain cells in a column or a row with an empty value. Suppression of a row is applied when a row is an outlier: the row has a quasi-identifier which does not occur often in the dataset, while the quasi-identifiers of other rows do occur often. If this row is removed, k will increase. An entire column may be suppressed when it contains too much different values. This is often applied to timestamp and location columns (Privitar, 2017);
- Generalization: Data generalization is the process of generating summary data with successive layers for a dataset. The purpose of data generalization, regarding privacy preservation, is to hide the characteristic of an individual from its group, such that the adversary will not be able to distinguish this individual from its peers. For a numerical value, a typical way is to replace the value by a range of it, so that the accuracy of the observation of an adversary decreases (Science Direct, 2021). Individual values are replaced with a broader category, making quasi-identifiers less precise to make sure that records with different values are generalized into records with the same value. Examples include categorizing birth dates (January 25, 1986) into age categories (20 to 25) or removing detail from postal codes[1].

Both suppression and generalization can be applied in two ways: global suppression or generalization will always transform all values for a certain column using the same method. Local suppression or generalization allows for different suppression or generalization methods for different records.

Datasets which have been anonymised using the k-anonymity method may still have issues regarding anonymisation in a couple of situations:

- Homogeneity – if all records in the sensitive column are similar, it is enough to know that a certain person is in the dataset to know his or her sensitive data (Privitar, 2017). For instance, if in a health related dataset almost all 70-75 year old males have a certain disease, it is sufficient to know someone belongs to this category to be able to identify the disease he is suffering from;
- Background knowledge attack: use the association between one or more quasi-identifiers and the sensitive attribute. Background knowledge (e. g. knowing that a certain disease is very common amongst males in certain geographical areas) is used to reduce the number of potential values for the sensitive attribute, increasing homogeneity;
- The data in the sensitive column must not reveal information that was generalized in one of the quasi-identifier columns (Privitar, 2017). For example, if you are recording whether or not people are pregnant or have prostate cancer, there is no point in hiding the gender column to increase k since gender can be deduced from the sensitive data.

3.2.3. ANONYMISATION EXAMPLE

Providing a real-world example of pseudonymisation and anonymisation is difficult for obvious reasons: it would entail displaying data which should actually be hidden. Therefore, a real-world example using fictitious data is used: the example is based on “Slim Melden”, an application citizens can use to report problems they notice in public space to the municipality. This application has to record contact details from people reporting problems (in order to be able to ask questions if needed and provide feedback once a problem has been solved) and a timestamp and a location to be able to locate the problem in public space. Table 1 contains the original record of such a report, as it is stored in the database inaccessible to everyone but the system administrators. It is the record before it is pseudonymised.

[1] An example for a Dutch postal code: the postal code of the Civity office is 3707 NH which points to roughly six addresses. Chances are this postal code occurs only once in the dataset. If the two letters are left off, the remainder points to an entire neighborhood which reduces the chances of identifying an individual address.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 1. Pseudonimization, before

Column	Type	Example	Action
ID	Sensitive data	37769-1586228173126-29-80330	-
Name	Personal data	Joe Example	Remove (pseudonymisation)
Phone	Personal data	0612345678	Remove (pseudonymisation)
Email	Personal data	joe@example.com	Remove (pseudonymisation)
Longitude	Quasi identifier	6.80282267396818	Generalize (anonymisation)
Latitude	Quasi identifier	52.2041960413761	Generalize (anonymisation)
Created	Quasi identifier	2021-04-07 09:46:38	Generalize (anonymisation)
Description	Sensitive data	A lot of weeds grow in the garden …	-
Category	Sensitive data	Weeds in plants / gardens	-
Status	Sensitive data	Open	-
Modified	Sensitive data	2021-04-07 11:34:12	-
Closed	Sensitive data	<NULL>	-

Table 2 contains the same record, but now the personal information has been removed. The record can no longer be related to a person directly (and thus complies with GDPR). When cross-referencing this record with other datasets, it would still be possible to identify the person who created the report. Longitude/latitude (the location) and the timestamp at which the report was created could be used to look up the personal data in another file.

Table 2. Pseudonimization, after – anonimization, before

Column	Type	Example	Action
ID	Sensitive data	37769-1586228173126-29-80330	-
Longitude	Quasi identifier	6.80282267396818	Generalize (anonymisation)
Latitude	Quasi identifier	52.2041960413761	Generalize (anonymisation)
Created	Quasi identifier	2021-04-07 09:46:38	Generalize (anonymisation)
Description	Sensitive data	A lot of weeds grow in the garden …	-
Category	Sensitive data	Weeds in plants / gardens	-
Status	Sensitive data	Open	-
Modified	Sensitive data	2021-04-07 11:34:12	-
Closed	Sensitive data	<NULL>	-

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 3 shows the same record once it has been anonymised. The location has been replaced with the name of the neighborhood in which the report was created and the timestamp has been replaced with a less exact one. Assuming there are now multiple records in the same neighborhood in the same week, the record can no longer be related to a specific person. But the sensitive information is still present and usable: the category and the description are still there and from the modified and closed field it can be concluded the municipality started working on the report, but has not closed it yet.

Table 3. Pseudonimization, after – anonimization, after

Column	Type	Example	Action
ID	Sensitive data	37769-1586228173126-29-80330	-
Neighborhood	Generalized	Kanaleneiland	Generalize (Anonymisation)
Created	Generalized	2021, week 14	Generalize (Anonymisation)
Description	Sensitive data	A lot of weeds grow in the garden …	-
Category	Sensitive data	Weeds in plants / gardens	-
Status	Sensitive data	Open	-
Modified	Sensitive data	2021-04-07 11:34:12	-
Closed	Sensitive data	<NULL>	-

If the generalized “neighborhood”/”created” combination is still unique (or almost unique, below the desired k) the quasi-identifiers in the record need further generalization (e. g. replace the week in the timestamp with the month). If the categories start containing a lot of records (i. e. k becomes high), the generalization could be reduced (maybe use days instead of weeks).

3.3. DATA ENCRYPTION

Data encryption translates data into another form, or code, so that only people with access to a secret key (the so called decryption key) or password can read it. Encrypted data is commonly referred to as ciphertext, while unencrypted data is called plaintext. Data, or plaintext, is encrypted with an encryption algorithm and an encryption key. The process results in ciphertext, which only can be viewed in its original form if it is decrypted with the correct key. Two main types of data encryption exist:

- asymmetric encryption, also known as public-key encryption;
- symmetric encryption.

3.3.1. SYMMETRIC ENCRYPTION

Symmetric-key ciphers use the same secret key for encrypting and decrypting a message or file. While symmetric-key encryption is much faster than asymmetric encryption, the sender must exchange the encryption key with the recipient before the receiver can decrypt the data. Exchange of the decryption key poses a security risk. AES, the Advanced Encryption Standard and bcrypt are examples of symmetric encryption algorithms. Table 4 contains an example of symmetric encryption using the AES algorithm. The email address is encrypted using the key. The receiver needs the key to be able to decrypt the email address again.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 4. Symmetric encryption/decryption example.

Key	cuchIsh9odNafUrgUmsOard9quorceldEwnAirjOkPyacnoxgi
Value to encrypt	joe@example.com
Encrypted value	oEEJOlNwHFl+XtVig14NRQ==
Decrypted value	joe@example.com

3.3.2. ASYMMETRIC ENCRYPTION

Asymmetric cryptography, also known as public-key cryptography, uses two different keys, one public and one private. The public key, as it is named, may be shared with everyone, but the private key must be protected. The public key is used to encrypt the data and the private key (which is only known by the receiver of the data) is used to decrypt the data. In other words: the public key can only be used to encrypt (lock) data and the private key is used to decrypt (unlock) data. The RSA algorithm is a popular algorithm for public-key encryption that is widely used to secure sensitive data. Table 5 contains an example. The receiver generates a public/private key pair. The public key is sent to the sender of the data who uses it to encrypt the data (in this case an email address). After receiving the data, the receiver uses the private key to decrypt the data.

Table 5. Asymmetric encryption/decryption example.

Public key	MIGfMA0GCSqGSIb3DQEBAQUAA4GNADCBiQKBg...
Value to encrypt	joe@example.com
Encrypted value	I6xaMbq7ar+G6ZPc17z/7JNEToqTTS78KsuEL...
Private key	<hidden>
Decrypted value	joe@example.com

With every doubling of the RSA key length, decryption becomes six to seven times slower (Rai, 2018). Therefore, when large messages have to be RSA encrypted, the performance degrades. In such a scenario, the data can first be encrypted using AES and then the key used for AES encryption/decryption is RSA encrypted before it is sent across the internet.

3.4. HASH OR CHECKSUM

A hash or checksum function provides a unique signature for digital data. A very simple hash function is to count the number of bytes. This is not a very secure hash though (if bytes are added or deleted number of bytes changes, but if the order of the bytes is changed, the number of bytes stays the same), so more advanced hashing algorithms should be used. Table 6 contains an example of two MD5 checksums. The file containing the string “This data has been modified” has a different hash then the file containing the string “This data has not been modified”, marking the two files as being different. Verifying the integrity of software is a common use case for hashes or checksums.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 6. Example checksums

Contents of file	MD5 checksum
This data has not been modified	09756e2b96472ce13f5628049594b417
This data has been modified	14981298d3d06ba81d15517d4a37d820

4. SECURITY AND PRIVACY METHODS AND PROTOCOLS

Security and privacy methods and protocols should be applied at different locations in the SCOREwater software infrastructure. Purpose of these methods and protocols is to ensure that the data users are downloading from the SCOREwater platform has not been compromised at some point.

- Sensor integrity must be guaranteed in order to be able to trust the sensor data coming in to the platform;
- Data must be securely stored in the SCOREwater platform to prevent unauthorized access to the data in case of security breaches.

4.1. SENSOR INTEGRITY

Sensor integrity refers to making sure that the devices uploading data towards your infrastructure are actually trusted sensors. The City of Gothenburg describes a number of principles that can be applied to make sure this is the case (Environment Administration, 2020). Sensor integrity is enforced by a) properly securing keys on the devices b) making sure the firmware of the device has not been tampered with and c) disable ports which are not needed.

4.1.1. CERTIFICATES AND KEYS

Sensor security systems are based on security certificates locked by keys. The keys must be individual for each sensor and well protected in the sensor device, e.g. in dedicated secure elements, which provide state of-the-art protection of keys both from hardware-based attacks (for example probing in the device for keys) or eavesdropping internal communication within the sensor. The solution must also provide protection against software-based attacks by e.g. reading out memory contents or injecting new code into the device. The keys should never be handled in clear text, not during production, start-up or in working memories during run time. They should only be handled in encrypted format. See paragraph 3.3 for a discussion of different encryption methods.

4.1.2. FIRMWARE

Encryption

To protect devices and meters, the firmware image must be encrypted. If the firmware is readable, it allows a hacker to create attack vectors to hack a device, either to provide malfunctions and data, or to provide bots that attack the system from within.

Secure boot/start-up

The device should check that the correct firmware is loaded and executed in the device to ensure correct functions and data, as well as protect from turning devices into bots. This is done by verifying the “hash” (also referred to as “checksum”, see paragraph 3.4). If any part of the firmware is changed, the hash of the firmware is changed. During boot, the firmware is hashed and checked before loaded. The device should be able to report the firmware hash for the running firmware.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Device integrity

As discussed in the previous paragraph, the device firmware is secured by a hash signature. The device should be able to report the firmware version and the hash it uses, proving that the device integrity is maintained.

4.1.3. DISABLED PORTS

To minimise attack vectors, all device ports should be disabled by default. If ports are available, there is a greater risk of data injection by dumping the runtime code to find attack vectors.

4.1.4. COMMUNICATION

The data from the sensor is sent over a communication system. The data must be protected from injection of false data, replay of old data, or information leakage.

Confidentiality

The data should never be sent in clear text over a communication link. A radio link is easy to pick-up, but also cables are easy to eavesdrop on if you have physical access. Therefore, all data should be ciphered with at least AES-128 bit ciphering (Environment Administration, 2020).

Integrity

The data must be integrity protected, proving that the data is not modified or altered during the data transfer. Furthermore, the system must protect toward replay attacks, e.g. a “salt” in messages. A salt is a piece of random data which is added to a message to disguise the actual message. The integrity should be protected with AES128 bit or higher.

Authentication

The smart meters and sensors shall be authenticated, proving that the sender is the correct device. The authentication is done with unique IDs and keys. The authentication shall be of strength AES128 bit or higher

4.2. STORAGE ON THE SCOREWATER PLATFORM

The SCOREwater platform consists of a number of servers which have been deployed in an isolated network. Figure 1 depicts the network infrastructure of the SCOREwater platform in a simplified form. The network contains multiple application servers, but only one is drawn here. To prevent unauthorized access to the data in the SCOREwater platform, the database server and application servers are not directly accessible from the internet. All traffic to the application servers is directed via a reverse proxy server. This proxy server acts as a gateway between the internet and the SCOREwater platform. To be able to access the application and database servers directly, an attacker has to first access the proxy server before launching an attack on the application or database servers.

The proxy server uses Transport Layer Security (TLS) to protect communication between server and the client and serves a certificate that is trusted by all major platforms. All information exchanged between client and server is encrypted. Only the client (for example the web browser of the user) and the server hold the key to be able to decrypt the information. TLS has been extensively tested and thoroughly reviewed and there are actually no viable alternatives available (Quora, 2021). It is important though to make sure a modern version of TLS (preferably 1.3) is being used.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

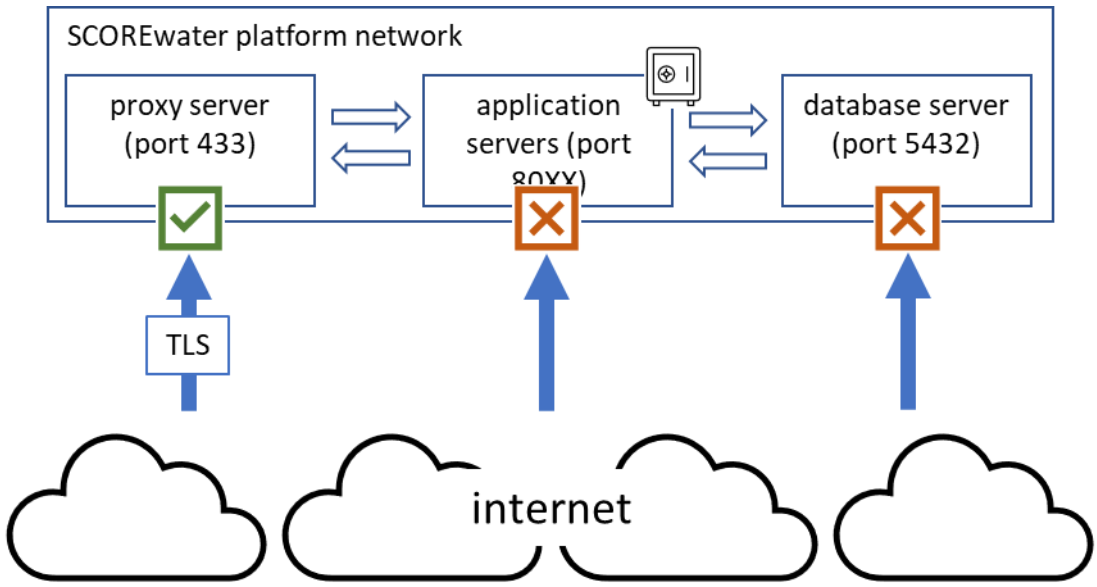

Figure 1. SCOREwater network architecture

To reduce the chance of unauthorized access to the database server in the SCOREwater platform, the servers within the SCOREwater platform network can only communicate to each other on ports that have been opened explicitly. For example: the application servers can only communicate to the database server on port 5432, other ports are closed by default. This gives the same level of security as moving the database to another network and place a firewall in between these networks.

The application servers require sensitive configuration information, for example database passwords or API keys to be able to access third party systems. As this information should not be stored on the server (or anywhere else for that matter) in a readable form, this information is stored in a so called vault. The passwords, API keys and other sensitive configuration settings are stored in the vault in a non-readable format. To obtain access to the information, one needs a key to be able to decrypt it. This key in turn should not be stored in a readable format anywhere, but in a secure location which can only be accessed by a limited number of people, for instance system administrators with so called root privileges.

Endpoints on the SCOREwater application server which are not intended to be used by the general public, but only by a limited number of clients are provided with an additional security measure: a so called whitelist. This whitelist only allows requests coming from certain IP addresses to pass. Requests originating from other clients will be rejected. These requests still use SSL to communicate securely (see Figure 2). This principle has been applied to provide sensors which push data to the SCOREwater platform directly with access to the endpoint created specifically for that purpose.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

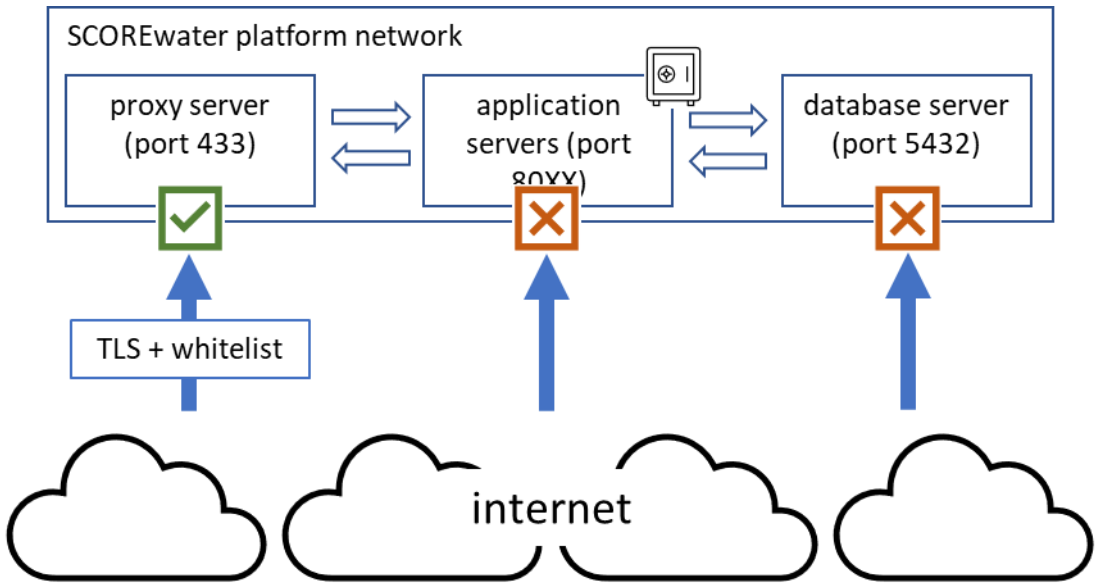

Figure 2. TLS plus whitelist to prevent unauthorized access to the SCOREwater platform network

Management of the SCOREwater platform, like applying updates and configuration changes can only be done via a dedicated VPN. Accessing the servers for maintenance requires Multi Factor Authentication and password policies (strong password, retention) apply.

5. IMPLICATIONS

If a use case justifies the collection of personal information, the dataset containing this personal information must be properly anonymised using one of the many methods described in this deliverable before the data is being uploaded to the SCOREwater platform. Anonymisation must be taken into consideration by the different cases and data providers when onboarding a dataset on the SCOREwater platform. In addition, when replicating the SCOREwater platform set-up to a different environment, the security and privacy methods and protocols which have been applied in the SCOREwater platform instance deployed within the frame of the project and which have been described in this deliverable must be replicated as well in order to prevent unauthorized access to the data.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

6. REFERENCES

Environment Administration, C. o. (2020). *Report number R2020:19 The LoV-IoT project: Air and water monitoring with Internet of Things.* Gothenburg: The City of Gothenburg.

Privitar. (2017, 04 7). *K – anonymity: An Introduction*. Retrieved from Privitar: https://www.privitar.com/blog/k-anonymity-an-introduction/

Quora. (2021, 04 23). *Are there any SSL/TLS alternatives used in the industry?* Retrieved from Quora: https://www.quora.com/Are-there-any-SSL-TLS-alternatives-used-in-the-industry

Rai, D. (2018, 03 10). *RSA Encryption and Decryption in Java*. Retrieved from Devglan: https://www.devglan.com/java8/rsa-encryption-decryption-java

Science Direct. (2021, 04 08). *Data generalization*. Retrieved from Data generalization: https://www.sciencedirect.com/topics/computer-science/data-generalization

Sweeney, L. (2002). Acieving k-anonymity privacy protection. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 571-588.

Thanga Revathi S, N. R. (2017). Data privacy preservation using data perturbation. *International Journal of Soft Computing and Artificial Intelligence*, 10-12.

Verhagen, L. (2021, 01 26). Datalek bij GGD: gegevens van miljoenen Nederlanders in criminele handen. *de Volkskrant*.

Youyang Qu, M. R. (2019). Privacy Preservation in Smart Cities. In K. Z. Danda B. Rawat, *Smart cities, cybersecurity and privacy* (pp. 75-88). Elsevier.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

ANNEX 1 – STOCKTAKING

A final Annex of stocktaking was included in all Deliverables of SCOREwater produced after the first half-year of the project. It provides an easy follow-up of how the work leading up to the Deliverable has addressed and contributed to four important project aspects:

1. Strategic Objectives
2. Project KPI
3. Ethical aspects
4. Risk management

STRATEGIC OBJECTIVES

Table 7 lists those strategic objectives of SCOREwater that are relevant for this Deliverable and gives a brief explanation on the specific contribution of this Deliverable.

Table 7. Stocktaking on Deliverable's contribution to reaching the SCOREwater strategic objectives.

Project goal	Contribution by this Deliverable
Ambition 2a: Interoperable ICT platform for the water sector	The deliverable facilitates making water data publicly available in a secure fashion
Ambition 2b: Harmonizing water knowledge systems	
Ambition 4: Standardized open water data available to all, sparking data-as-a-service market	
Strategic Objective 5, Identify and mitigate key barriers to implementation of smart, resilient water management.	The deliverable identifies security and privacy issues and describes how to work around those.

PROJECT KPI

Table 8 lists the project KPIs that are relevant for this Deliverable and gives a brief explanation on the specific contribution of this Deliverable.

Table 8. Stocktaking on Deliverable's contribution to SCOREwater project KPIs.

Project KPI	Contribution by this deliverable
10 - Standardization barriers identified and mitigation options demonstrated	The deliverable identifies security and privacy issues and describes how to work around those.
12 - Technological barriers identified and mitigation options demonstrated	

ETHICAL ASPECTS

Table 9 lists the project's Ethical aspects and gives a brief explanation on the specific treatment in the work leading up to this Deliverable. Ethical aspects are not relevant for all Deliverables. Table 9 indicates "N/A" for aspects that are irrelevant for this Deliverable.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 9. Stocktaking on Deliverable's treatment of Ethical aspects.

Ethical aspect	Treatment in the work on this Deliverable
Justification of ethics data used in project	N/A
Procedures and criteria for identifying research participants	N/A
Informed consent procedures	N/A
Informed consent procedure in case of legal guardians	N/A
Filing of ethics committee's opinions/approval	N/A
Technical and organizational measures taken to safeguard data subjects' rights and freedoms	N/A
Implemented security measures to prevent unauthorized access to ethics data	N/A
Describe Anonymisation techniques	N/A
Interaction with the SCOREwater Ethics Advisor	N/A

RISK MANAGEMENT

Table 10 lists the risks, from the project's risk log, that have been identified as relevant for the work on this Deliverable and gives a brief explanation on the specific treatment in the work leading up to this Deliverable.

Table 10. Stocktaking on Deliverable's treatment of Risks.

Associated risk	Treatment in the work on this Deliverable
8 - Lack of consensus on scientific, technological or business model approach	This risk did not occur
10 - Data from Cases are sparse and are not enough to apply all methods and tools	Data which needed pseudonymisation/anonymisation was indeed sparse. Fictitious examples based on real world cases have been used which also alleviates us of any ethical issues related to showing actual data in the deliverable.
13 - Failure architecture implementation and modules integration	This risk did not occur

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

scoreWATER

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751