decode

Privacy Design Strategies for the DECODE architecture - update

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 732546

Project no. 732546

DECODE

DEcentralised Citizens Owned Data Ecosystem

D1.3 Privacy Design Strategies for the DECODE Architecture

Version Number: V1.0

Lead beneficiary: RU

Due Date: January 2018

Author(s): Shehar Bano, Alberto Sonnino (UCL), Eleonora Bassi, Marco Ciurcina, Juan Carlos De Martin, Selina Fenoglietto, Antonio Santangelo (POLITO), Francisco Sacramento Gutierres, David Laniado, Pablo Aragón (EURECAT), Jaap-Henk Hoepman (RU)

Editors and reviewers: Elena Japundžić, Denis "Jaromil" Roio (Dyne)

Dissemination level:		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Approved by: Francesca Bria (Chief Technology and Digital Innovation Officer, Barcelona City Hall)

Date: [31/01/2019]

This report is currently awaiting approval from the EC and cannot be not considered to be a final version.

Table of Contents

1. Introduction

DECODE aims to develop a privacy preserving data distribution platform to foster commons-based sharing economy models, where citizens own and control their data. This asks for a privacy by design-based approach, for which the concept of privacy design strategies have recently been developed.

The General Data Protection Regulation (GDPR), as well as other data protection or privacy protection laws and regulations, define data protection in legal terms. These terms are soft, open to interpretation, and highly dependent on context. Because of this inherent vagueness, engineers find such legal requirements hard to understand and interpret.

The GDPR also mandates privacy by design, without describing clearly what this means exactly, let alone giving concrete guidelines on how to go about implementing privacy by design when actually designing a system. Intuitively, privacy design means addressing privacy concerns throughout the system development lifecycle, from the conception of a system, through its design and implementation, proceeding through its deployment all the way to the decommissioning of the system many years later. In terms of software engineering, privacy is a quality attribute, like security, or performance. To make privacy by design concrete, the soft legal norms need to be translated into more concrete design requirements that engineers understand. This is achieved using privacy design strategies.

Last year we released our initial deliverable (D1.2) with a preliminary recommendation on how the DECODE architecture could and should protect the privacy of the end users of the DECODE architecture.

In this deliverable (D1.3) we elaborate on that initial analysis. We describe the legal constraints (2), describe the initial DECODE architecture (3), describe and apply the privacy design strategies approach to it (4), in particular to the Barcelona CitizenSensing (IoT) and Digital Democracy and Data Commons (DDDC) pilots (5) and discuss privacy in relation to non-discriminatory data mining in the context of DECODE (6). The result is a list of concrete recommendations to guide the design and implementation of the DECODE architecture, data collection and data processing. This deliverable is structured accordingly.

2. Legal constraints and prescribed methods

According to Privacy and Data Protection European legislation, privacy by design measures should assist all the data processing phases (i.e. architecture and data processing) in order to protect and enhance individual rights and the ethical coherence of the entire project (DECODE).

2.1. General Legal Framework on Privacy by Design:

Since the late '90 the principle of privacy by design was introduced by Ann Cavoukian (Cavoukian, 2010) and its fortune in European legal framework is stated from the document "The Future of Privacy" (02356/09/EN – WP168) adopted on December 1st, 2009 by EU Article 29 Data Protection Working Party (WP29) and the Working Party on Police and Justice (WPPJ).

In 2012 it was included in the Proposal of revision of the Directive 95/46/EC, and, finally, it was fixed in Article 25 of the General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, that introduces legal obligation to design strategies.

For a full comprehension of the implications of the PbD principle, it should be interpreted in accordance with the recommendations by WP29 and by the European Data Protection Supervisor (see EDPS opinion on privacy in the digital age: "Privacy by Design" as a key tool to ensure citizens' trust in ICTs), and taking advantages from the standards and principles stressed by the International Standard Organization (ISO 29100).

The GDPR is fully enforceable since 25 May 2018 and replaces Directive 95/46/EC.

According to Article 25 of GDPR "Data protection by design and by default":

1. Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.
2. The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing,

the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons.

3. An approved certification mechanism pursuant to Article 42 may be used as an element to demonstrate compliance with the requirements set out in paragraphs 1 and 2 of this Article.

In order to better understand the meaning of this provision, in the next we enumerate the subjects responsible for aboption of privacy by design measures, mention the period of adoption of the measures and what are the measures in question.

2.2. Subjects responsible for adoption of privacy by design measures

Article 25 provides a method to enforce the fairness and the transparency of the data processing.

According to it, the adoption of privacy by design measures is clearly an obligation of the data controller[1], who is the person or the entity which determines means of the data processing (pursuing the definition set by Article 4 (1,7) GDPR). Hence, the distributed architecture of Decode, and more in general of DLTs, requires to point out carefully how to individuate the data controller[2], or the joint data controllers (according to Article 26, GDPR "Joint controllers" with regarding to the determination of the means of the processing[3]). Furthermore, Recital 78 of the GDPR and scholars (for instance Koops & Leenes, 2014) outline that the adoption of privacy by design measures is not only an obligation of data controllers, but also a recommendation for IT systems producers.

> (...) *When developing, designing, selecting and using applications, services and products that are based on the processing of personal data or process personal data to fulfil their task, producers of the products, services and applications should be encouraged to take into account the right to data protection when developing and designing such products, services and applications and, with due regard to the state of the art, to make sure that controllers and processors*

---

[1] See Articles 4, 1 (7) and 24, GDPR. For how to interpret these provisions see also the Opinion 1/2010 on the concepts of "controller" and "processor", adopted by WP29 (wp169)).

[2] See: Finck (2018), p 17; Ibanez, O'Hara, Simperl (2018), p. 4

[3] "1) Where two or more controllers jointly determine the purposes and means of processing, they shall be joint controllers. They shall in a transparent manner determine their respective responsibilities for compliance with the obligations under this Regulation, in particular as regards the exercising of the rights of the data subject and their respective duties to provide the information referred to in Articles 13 and 14, by means of an arrangement between them unless, and in so far as, the respective responsibilities of the controllers are determined by Union or Member State law to which the controllers are subject. The arrangement may designate a contact point for data subjects.

2) The arrangement referred to in paragraph 1 shall duly reflect the respective roles and relationships of the joint controllers vis-à-vis the data subjects. The essence of the arrangement shall be made available to the data subject.

3) Irrespective of the terms of the arrangement referred to in paragraph 1, the data subject may exercise his or her rights under this Regulation in respect of and against each of the controllers."

> *are able to fulfil their data protection obligations. The principles of data protection by design and by default should also be taken into consideration in the context of public tenders.* (GDPR, Rec. (78))

Scholars stressed the importance of Article 25 in combination with Recital 78 pointing out the potential in terms of transparency, fairness and accountability for information systems as ecosystems. The Privacy by design method offers organizations a way to operationalize legal requirements[4].

These considerations require us to address privacy by design issues by clearly distinguishing the proper roles and responsibilities of different nodes of Decode's architecture[5] (see Berberich & Steiner, 2016).

Article 25 of GDPR does not properly address any obligation of privacy by design on the data processor, who is the "natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller" (pursuing the definition provided by Article 4 (1,8) GDPR). The determination of the means of the data processing, and by consequence, of the adequate design of the processing, is an obligation of the controller. Moreover, Article 28 (1) provides that

> *Where processing is to be carried out on behalf of a controller, the controller shall use only processors providing sufficient guarantees to implement appropriate technical and organisational measures in such a manner that processing will meet the requirements of this Regulation and ensure the protection of the rights of the data subject.*

This provides us with two remarks. On the one hand, the controller shall use only processors who can ensure to implement the processing following the principle of privacy by design as defined by Article 25. On the other hand, the designed data processor has the duty to follow the requirements of Article 25, regarding the instructions provided by the data controller for the data processing in order to "meet the requirements of this Regulation and ensure the protection of the rights of the data subject". These instructions shall be included in the agreement between the controller and the processor, according Article 28 (3) GDPR, that states:

Processing by a processor shall be governed by a contract or other legal act under Union or Member State law, that is binding on the processor with regard to the controller and that sets out the subject-matter and duration of the processing, the nature and purpose of the processing, the type of personal data and categories of data subjects and the obligations and rights of the controller. That contract or other legal act shall stipulate, in particular, that the processor:

> *(a) processes the personal data only on documented instructions from the controller, including with regard to transfers of personal data to a third country or an international organisation, unless required to do so by Union or Member State law to which the processor is subject; in such*

4 See Kurtz, Semmann, Böhmann, 2018.

5 For these roles within DECODE, see the analysis provided in D1.8 and D1.9.

a case, the processor shall inform the controller of that legal requirement before processing, unless that law prohibits such information on important grounds of public interest;

(b) ensures that persons authorised to process the personal data have committed themselves to confidentiality or are under an appropriate statutory obligation of confidentiality;

(c) takes all measures required pursuant to Article 32;

(d) respects the conditions referred to in paragraphs 2 and 4 for engaging another processor;

(e) taking into account the nature of the processing, assists the controller by appropriate technical and organisational measures, insofar as this is possible, for the fulfilment of the controller's obligation to respond to requests for exercising the data subject's rights laid down in Chapter III;

(f) assists the controller in ensuring compliance with the obligations pursuant to Articles 32 to 36 taking into account the nature of processing and the information available to the processor;

(g) at the choice of the controller, deletes or returns all the personal data to the controller after the end of the provision of services relating to processing, and deletes existing copies unless Union or Member State law requires storage of the personal data;

(h) makes available to the controller all information necessary to demonstrate compliance with the obligations laid down in this Article and allow for and contribute to audits, including inspections, conducted by the controller or another auditor mandated by the controller.

With regard to point (h) of the first subparagraph, the processor shall immediately inform the controller if, in its opinion, an instruction infringes this Regulation or other Union or Member State data protection provisions.

2.3. hen shall privacy by design measures be adopted?

Privacy by design technical and organizational measures shall be adopted "both at the time of the determination of the means for processing and at the time of the processing itself" (Article 25(1) GDPR). This distinction aims to underline the necessity of pursuing privacy preserving goals in the design phase and also due to all the processing, introducing a sort of double responsibility both in design and implementation.

In addition, the requirement of implementing the privacy preserving measures "in an effective manner and to integrate the necessary safeguards into the processing" implies that the privacy by design principle implies an on-going monitoring process in order to ensure the effectiveness of the measures adopted. This is also a requirement set up by Article 24 (1) on the obligation of the data controller, which specifies that: "Those measures shall be reviewed and updated where necessary".

So that, to be compliant to this provision, it seems necessary to define and design the architecture both of the entire system and the data collection and processing, in order to be able to pinpoint critical phases and requirements, and then to choose which are the measures to implement. Within the DECODE Project this implies that the design and the implementation of such measures shall cover both the DECODE general architecture and the Pilots' activities, data collection and processing.

Furthermore, it is important to stress that in order to foster the transparency and the accountability of the data processing, the GDPR set up a close relation between, on the one hand, designing and adopting privacy by design measures for protection of the rights of data subjects and, on the other hand, assessing risks and adequate measure to face them. Article 35 GDPR provides the data controller with the obligation to "carry out an assessment of the impact of the envisaged processing operations on the protection of personal data" (Article 35 (1) GDPR). On this account, it's important to point out that the data processor has the obligation to "assists the controller in ensuring compliance with the obligations pursuant to Articles 32 to 36 taking into account the nature of processing and the information available to the processor", pursuing Article 28 (3,f).

According to Article 35 (7,d), this assessment, the DPIA (data protection impact assessment)

> *"shall contain at least": "the measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data and to demonstrate compliance with this Regulation taking into account the rights and legitimate interests of data subjects and other persons concerned"* (Article 35(7,d) GDPR).

Then, including details on when specific privacy preserving measures are adopted within the DPIA pursuing Article 35 GDPR can be considered a good practice.

2.4. What kind of measures decide to adopt?

Article 25 expressly provides an obligation to adopt both technical and organizational measures.

It is important to stress that these measures shall be adopted "taking into account the state of the art" (Article 25(1) GDPR). This wording implies the duty for the data controller to stay updated on technical advancement in privacy technologies as well as on standards of diffusion and recommendations. Data controllers have the same duty in order to guarantee "to implement appropriate technical and organisational measures in such a manner that processing will meet the requirements of this Regulation and ensure the protection of the rights of the data subject" (Article 28 (1)).

With regard to the technical measures, Article 25 mentions pseudonymisation and data minimisation as examples of privacy by design measures, but the list is longer and includes security measures, encryption, anonymisation, aggregation, third parties limitation access, tools for ensuring data subject's informed consent and data subject's

right (among others: the right to erasure[6], the right to access, the right to be forgotten and data portability)[7]. These measures will be further analysed in the next sections of this document.

With regard to the organizational measures, it is possible to mention (this list is not exhaustive):

- the adoption of automatic means for the collection of the informed data subject's consent, and for the withdrawal of her given consent to the processing (pursuing Article 6 GDPR);
- the adoption of fair and appropriate measures to provide any information and communication to the data subject pursuing Articles 13-22 on data subject's rights "in a concise, transparent, intelligible and easily accessible form, using clear and plain language" (Article 12 (1));
- the use of "privacy friendly" icons for the interaction with the data subject. Article 12 (7) provides that "The information to be provided to data subjects pursuant to Articles 13 and 14 may be provided in combination with standardised icons in order to give in an easily visible, intelligible and clearly legible manner a meaningful overview of the intended processing. Where the icons are presented electronically, they shall be machine-readable";
- the implementation of user interfaces for "privacy friendly" interactions with data subjects[8];
- the adoption of automatic means (or protocols) for the exercise of the data subject rights (or at least for some of them, e.g. data portability, and the right to obtain copy of the data referring to the data subject in the exercise of the right to access provided by Article 15 GDPR);
- the engagement of a data processor who can ensure to process the personal data in an adequate manner (Article 28)
- the duty of the data controller to "maintain a record of processing activities under its responsibility" (Article 30 (1) GDPR) and to make it available to the supervisor authority under request;
- the adoption of security measures under Article 32 GDPR (it is important to stress the correspondence between Article 25 and Article 32);
- where necessary, the designation of a DPO (data protection officer) according to Article 37 GDPR;
- the adoption of appropriate measures to notify a data breach to the supervisory authority without undue delay (Article 33 GDPR).
- the adoption of a DPIA, pursuing Article 35;
- the adherence to approved codes of conduct as referred to in Article 40 or approved certification mechanisms as referred to in Article 42 GDPR;

---

[6] See also Pagallo, Bassi, Crepaldi, Durante (2018, forthcoming).

[7] On the relation between privacy by design and privacy by default measures for blockchain technologies, see Finck (2018), p. 27.

[8] In this perspective, and DECODE Project could be an experimental context, can be useful designing and implementing technologies to store and record all the crucial information on the processing and on the interaction between data controller, data processor and data subjects (e.g.: information according to Articles 13-14 GDPR received by the controllers, consent manifestations collected by them, possible additional obligations assumed by them and irrefutable evidences of the above).

Nonetheless. in some cases, for instance when we are dealing with data minimization, it is not so easy to distinguish between technical and organizational measures. This one of the reasons why we are focusing on Privacy Design Strategies, in order to proactively enforce a more pervasive privacy by design principle.

Moreover, all these measures are obviously not alternative but compatible. These are strategies, patterns and tools that can work simultaneously to gain specific goals in preserving and in enforcing privacy rights (see Hoepman, 2014; Colesky et al., 2016), following the idea of modularity of design.

Scholars introduced some other distinctions and classifications. Some measures could be described as self-enforceable (e.g. encryption and third parties' limitation access), while others are directed to conduct or to change user's behaviour (e.g. automatic notifications for updating data subjects on the data processing) (Pagallo, 2012). Moreover, some scholars distinguish between measures adopted by code and measures adopted by policy or by communication (this is the case of DPIA pursuing Article 35 GDPR) (see for instance Koops & Leenes, 2014).

3. High level architecture description

DECODE provides a distributed and privacy-aware architecture for decentralised data governance and federated identities. Key components of this architecture are as follows:

3.1. DECODE Hubs and Nodes

A device that runs the DECODE OS is called the DECODE HUB. The DECODE NODE is an interface that can be used to configure smart rules. It also abstracts common choices regarding granting access to data and mapping its use. The DECODE OS provides a cross-platform, securely connected base operating system that grants the integrity of execution of the NODE rules and applications. It can run on different hardware (e.g., a GNU/Linux/BSDbased computer, a mobile device, or a file server) or can be virtualized inside other host operating systems.

3.2. Smart Rules

DECODE's fully decentralised architecture offers a flexible and extensible data governance, which enables fine-grained control of different regimes of data ownership and privacy. Smart rules are a set of algorithmic protocols expressed in a formal language that implement this flexibility. Data User Providers, as personal data owners, can use smart rules to define how data should be managed in terms of access, value attribution and other parameters, and legal/contractual obligations and other constraints. Smart rules follow a defined structure to manage access to subsets of data (e.g., personal data or for specific uses granted to specific subjects). Smart rules can also be used to revoke authorisation for access or change the legal status and the conditions of use and exploitation of the data. More in general Smart Rules can be used to embed privacy preserving rules or additional privacy clauses and commitments (see D1.9 and its Annexes).

Smart rules can be expressed in a declarative language, which may be visually represented, which is then compiled in a functional language and executed. Executed smart rules can provide basic functions of governance and identity management: publish/subscribe access to events and functions to interface with external APIs; core functions to store and access the blockchain according to ABC entitlements; library functions to interfaces with external applications.

Smart rules enable providers and app developers to define rules about operation of the system or the regulatory environment. Such abstraction between people's choices and its enforcement creates a rich landscape for flexible and decentralised creation of new applications and services.

The smart rules, as well as all the platform specifications, protocols, ontology, semantic specifications, will be released under a Free and Open Source Software (F/OSS) license. The initial set of rules will be gradually extended with community participation as

requirements evolve, with the goal to eventually emerge as the standard language for managing data access and valorisation in distributed and decentralised architectures.

3.3. Distributed Ledger

A distributed ledger is a decentralized data repository that is resistant to malware and hacking, and provides privacy and transparency through ABC (Attribute Based Cryptography) and other privacy-enhancing technologies. Cryptographic primitives enforce strict access control that maintain the ledger's security and accuracy. A fully decentralized platform is realized by combining smart rules with distributed ledger technologies, which enforces by design flexible and extensible data governance.

The platform supports multiple, diverse contexts of data ownership and privacy. A heterogeneous set of data streams can be collected and fed to the platform: civic datasets and open linked data, private data, personal data with undisclosed identity and personal data associated with an electronic identity.

Data confidentiality is enforced by encryption, which also allows pseudonymization and/or anonymization since association with an e-identity depends on permissions specified by the data owner. Thus personal data can be consensually and anonymously used for collective intelligence, or for personalised services and applications (if authorised by data subjects). User authorizations are managed by defining ontologies and indexes over data streams collected by sensors, IoT objects or personal devices.

We implement the distributed ledger as Chainspace[9]—a distributed ledger platform for high-integrity and transparent processing of transactions within a distributed or decentralized system. Unlike application specific distributed ledgers, such as Bitcoin (Nakamoto, 2008) supporting a currency, or certificate transparency (Laurie et al, 2013) supporting certificate verification, Chainspace offers extensibility though supporting smart contracts, like the Ethereum platform (Wood, 2014). However, the Chainspace system is exposed to enough information about contract and transactions, in order to support and provide higher scalability through automatic sharding. The platform is agnostic as to the smart contract language, or identity infrastructure. In fact DECODE also implements a Zenroom VM for smart contract execution. Privacy features can be integrated in the system through modern zero-knowledge proofs or SNARKs.

3.4. Improved Distributed Ledger Security

Light clients, also known as Simple Payment Verification (SPV) clients, are nodes which only download a small portion of the data in a blockchain, and use indirect means to verify that a given chain is valid. Typically, instead of validating block data, they assume that the chain favoured by the blockchain's consensus algorithm only contains

9 See https://gogs.dyne.org/DECODE/wip/src/5477da0ac089795ecc13c11b64a3591e9cfae121/distributed-ledger/ChainspaceDL_april17.pdf

valid blocks, and that the majority of block producers are honest. By allowing such clients to receive fraud proofs generated by fully validating nodes that show that a block violates the protocol rules, and combining this with probabilistic sampling techniques to verify that all of the data in a block actually is available to be downloaded, we can eliminate the honest-majority assumption, and instead make much weaker assumptions about a minimum number of honest nodes that rebroadcast data. Fraud and data availability proofs are key to enabling on-chain scaling of blockchains (e.g., via sharding or bigger blocks) while maintaining a strong assurance that on-chain data is available and valid. We present, implement, and evaluate a novel fraud and data availability proof system.

To instantiate a blockchain, we make use of sparse Merkle trees, and represent the state as a key-value map. In a UTXO-based model, the keys in the map are transaction output identifiers e.g., hash(hash(d)||i) where d is the data of the transaction and i is the index of the output being referred to in d. The value of each key is the state of each transaction output identifier: either unspent (1) or nonexistent (0, the default value). The state would need to keep track of all data that is relevant to block processing, including for example the cumulative transaction fees paid to the creator of the current block after each transaction.

We define a function transition that performs transitions without requiring the whole state tree, but only the state root and Merkle proofs of parts of the state tree that the transaction reads or modifies. These Merkle proofs are effectively expressed as a sub-tree of the same state tree with a common root.

A faulty or malicious miner may provide an incorrect state root. We can use the execution trace to prove that some part of the execution trace was invalid. We define a function “VerifyTransitionFraudProof" and its parameters which verifies fraud proofs received from full nodes. If the fraud proof is valid, then the block that the fraud proof is for is permanently rejected by the client. In summary, the fraud proof verifier checks if applying the transactions in a period of the block’s data on the intermediate pre-state root results in the intermediate post-state root specified the block data. If it does not, then the fraud proof is valid.

A malicious block producer could prevent full nodes from generating fraud proofs by withholding the data needed to recompute the tree’s root and only releasing the block header to the network. The block producer could then only release the data—which may contain invalid transactions or state transitions—long after the block has been published, and make the block invalid. This would cause a rollback of transactions on the ledger of future blocks. It is therefore necessary for light clients to have a level of assurance that the data matching the tree’s root is indeed available to the network.

We propose a data availability scheme based on Reed-Solomon erasure coding, where light clients request random shares of data to get high probability guarantees that all the data associated with the root of a Merkle tree is available. The scheme assumes there is a sufficient number of honest light clients making the same requests such that the network can recover the data, as light clients upload these shares to full nodes, if a full node who does not have the complete data requests it. It is fundamental for light clients to have assurance that all the transaction data is available, because it is only necessary to withhold a few bytes to hide an invalid transaction in a block.

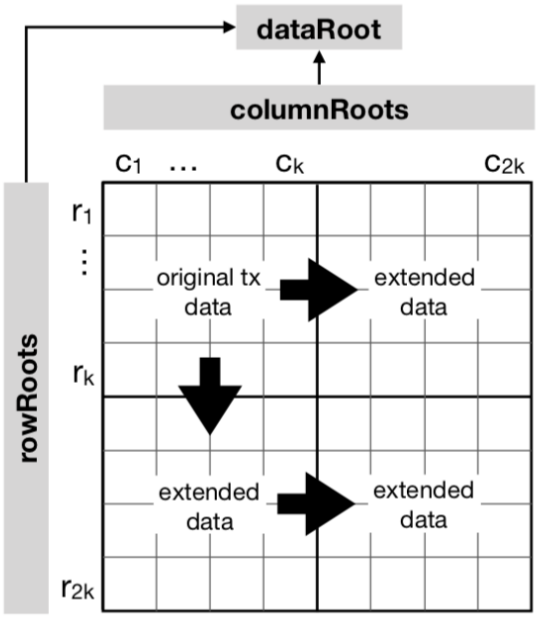
To provide some intuition, we describe a strawman data availability scheme, based on standard Reed-Solomon coding.

A block producer compiles a block of data consisting of k shares, extends the data to 2k shares using Reed-Solomon encoding, and computes a Merkle root over the extended data, where each leaf corresponds to one share.

When light clients receive a block header with a Merkle root, they randomly sample shares from the Merkle tree that the root represents, and only accept a block once it has received all of the shares requested. If an adversarial block producer makes more than 50% of the shares unavailable to make the full data unrecoverable (recall in recalled in Section 2.3 that Reed-Solomon codes allow recovery of 2t shares from any t shares), there is a 50% chance that a client will randomly sample an unavailable share in the first draw, a 25% chance after two draws, a 12.5% chance after three draws, and so on, if they draw with replacement. (In the full scheme, they will draw without replacement, and so the probability will be even lower.)

Note that for this scheme to work, there must be enough light clients in the network sampling enough shares so that block producers will be required to release more than 50% of the shares in order to pass the sampling challenge of all light clients, and so that the full block can be recovered.

The problem with this scheme is that an adversarial block producer may incorrectly construct the extended data, and thus the incomplete block is unrecoverable from the extended data even if more than 50% of the data is available. With standard Reed-Solomon encoding, the fraud proof that the extended data is invalid is the original data itself, as clients would have to re-encode all data locally to verify the mismatch with the given extended data, and thus it re- quires O(n) data with respect to the size of the block. Therefore, we instead use multi-dimensional encoding, so that proofs of incorrectly generated codes are limited to a specific axis—rather than the entire data—reducing proof size to O(√d n) where d is the number of dimensions of the encoding. For simplicity, we will only consider two-dimensional Reed-Solomon encoding.



A 2D Reed-Solomon Encoded Merkle tree can be constructed as follows from a block of data:

1. Split the raw data into shares of size shareSize each, and arrange them into a k × k matrix; apply padding if the last share is not exactly of size shareSize, or if there are not enough shares to complete the matrix.
2. Apply Reed-Solomon encoding on each row and column of the k × k matrix to extend the data horizontally and vertically; i.e., encode each row and each column. Then apply a third time a Reed-Solomon encoding horizontally, on the vertically extended portion of the matrix to create a 2k × 2k matrix, as shown in the Figure above. This results in an extended matrix Mi for block i.
3. Compute the root of the Merkle tree for each row and column in the 2k × 2k matrix, where each leaf is a share.
4. Compute the root of the Merkle tree of the roots computed in step 3.

A complete and deep technical explanation is available here: http://www0.cs.ucl.ac.uk/staff/M.AlBassam/publications/fraudproofs.pdf

3.5. Distributed Ledgers and Privacy

Distributed ledger technologies pose compliance problems with privacy laws, including the General Data Protection Regulation (GDPR), Regulation (EU) 2016/679. It is a matter of fact that data recorded on a distributed ledger could be personal data. Also encrypted data or hashes can be personal data. The issue of GDPR compliance on DLT is highly debated in these months and many different opinions face each other. So, obligations provided by the GDPR have to be considered when personal data processing is performed using DLT, but the legal framework is not completely clear.

Likely, on the 24th of September 2018, the Commission Nationale de l'Informatique et des Libertés (CNIL: the French privacy supervisory authority) published "Blockchain. Solutions for a responsible use of the blockchain in the context of personal data"[10]. a document that allows to work on a more solid ground. Particularly, the CNIL seems to allow storing of personal data in distributed ledger using some different techniques, depending on the circumstances (e.g., commitment, fingerprint generated by a hash function with a key, encryption, etc.).

The path to follow seems clear: the controller, before adopting a distributed ledger technology for processing personal data, has to perform a data protection impact assessment (DPIA) according to Article 35 of the GDPR and, eventually, a consultation of the competent supervisory authority according to Article 36 of the GDPR. Steps in this direction have already been started for the DDDC pilot that, to a certain extent, implies using distributed ledger technologies for the processing of personal data.

10 See https://www.cnil.fr/en/blockchain-and-gdpr-solutions-responsible-use-blockchain-context-personal-data.

4. Privacy design strategies

As explained in the introduction, the GDPR defines data protection in more vague legal terms. Engineers find such legal requirements hard to understand and interpret. In particular, the GDPR also mandates privacy by design, without describing clearly what this means exactly, let alone giving concrete guidelines on how to go about implementing privacy by design when actually designing a system.

To understand the privacy by design approach, and see how it can be made more concrete, one needs to know a little bit about how IT systems are usually developed. System development typically proceeds through a number of distinct phases namely: definition, design, development, implementation operation, evaluation and decommissioning. These correspond to the system life cycle (see figure 1).

To make privacy by design concrete, the soft legal norms need to be translated into more concrete design requirements that engineers understand. And tools to elicit and implement these requirements need to be available. For the design and development phase, such tools are available. In particular, so called *privacy enhancing technologies* (PETS) have been developed in the last thirty years ago (starting with the seminal work of David Chaum in the eighties). And also for the design phase, privacy design patterns have started to emerge. (We will discuss them briefly further on in this report). Unfortunately, until recently concrete tools to address privacy during the early design phases of a system, i.e. during the concept formulation and definition phase, were missing. This is why *privacy design strategies* have been developed.

As described in (Colesky et. al. 2016) a privacy design strategy specifies a distinct architectural goal in privacy by design to achieve a certain level of privacy protection. It is noted that this is different from what is understood to be an architectural strategy within the software engineering domain. Instead our strategies can be seen as goals of the privacy protection quality attribute (where a quality attribute is a term from software engineering describing non-functional requirements like performance, security, and also privacy).

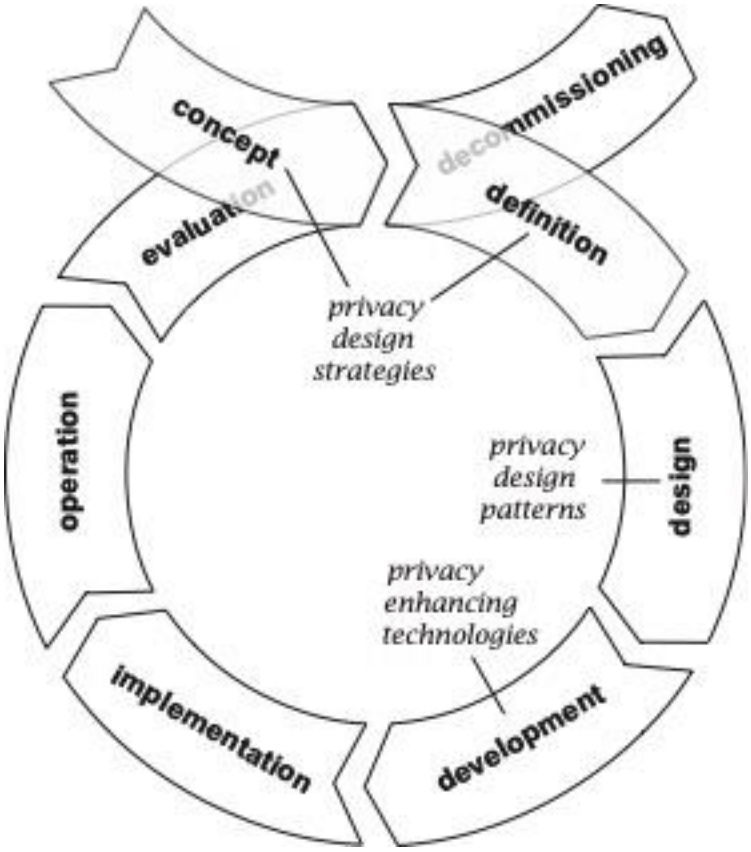


Figure 1 System lifecycle

In the description of the privacy design strategies we frequently refer to *processing* of personal data. Engineers[11] should be aware that the legal concept of processing is broader than what a typical engineer understands processing to mean. In what follows we use the legal interpretation of processing, which includes creating, collecting, storing, sharing and deleting personal data.

We now proceed to briefly describe the eight privacy design strategies (see figure 1). More information can be found in Hoepman, 2014 and Colesky et al, 2016. Each strategy is first described using a brief definition. Subsequently, the strategy is refined by one or more *tactics* that each describe a different way in which the overarching strategy can be achieved. We then present some examples by showing how the strategy impacts the DECODE architecture.

[11] When writing "engineer" we mean the large class of professionals that are tasked to engineer something. In particular this includes "developers", "systems developers", "software developers" etc.

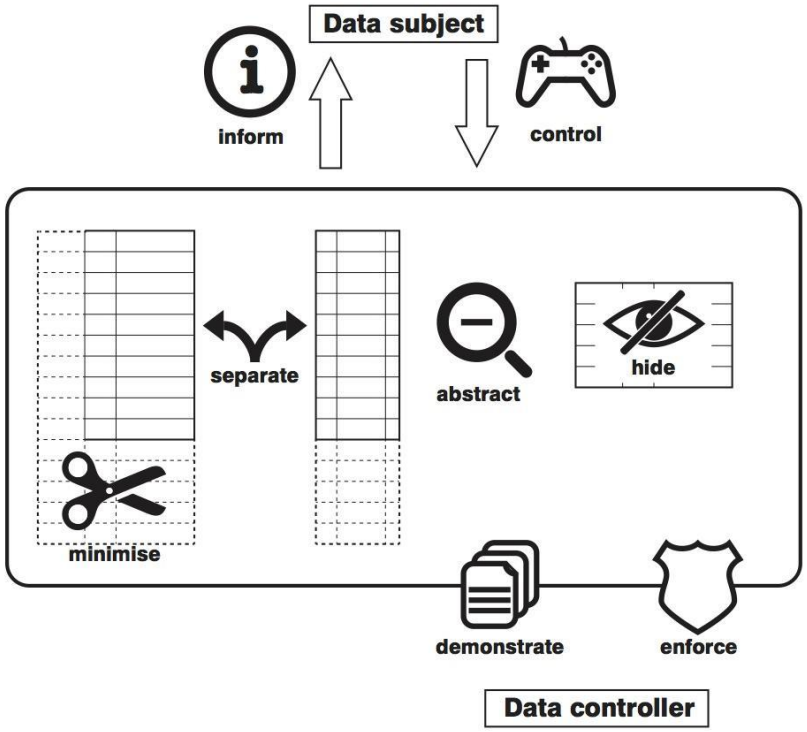


Figure 2 Privacy design strategies

4.1. Minimise

Definition: Limit the processing of personal data as much as possible.

The most obvious approach to protect privacy is by minimising the amount of personal data you process. In the ideal case you do not process any personal data at all, although in practice you almost always will (if only because the visitor of your website exposes her IP address to your web server).

Minimisation is an important strategy because personal data that you do not store or process cannot be abused, misinterpreted, breached, subpoenaed, leaked, sold, etc. In other words, it protects you against errors, malicious employees, incompetent third party processors, overly inquisitive governments and law enforcement agencies, or greedy shareholders and investors. Also, minimisation ensures that your users do not need to trust you to process their data responsibly. Instead of relying on rules or regulations, the system design itself prevents problems, simply because the data is not there.

Associated tactics:

Exclude: refraining from processing a data subject's personal data, partly or entirely, akin to blacklisting or opt out.

Select: decide on a case by case basis on the full or partial usage of personal data, akin to whitelisting or opt-in.

Strip: removing unnecessary personal data fields from the system's representation of each user.

Destroy: completely removing a data subject's personal data.

Example / impact on DECODE:

Explicit selection or exclusion of data items should be done by the DECODE nodes according to strict rules, that depend on the specific application at hand.

As an example to satisfy the "*destroy*" tactic, data items processed by DECODE should always be tagged with an expiry date. This should happen at the smallest granularity of data items where a distinction between expiry dates is relevant. Any node that encounters a data item whose expiry date lies in the past should discard that data item (and any copies thereof under that node'' control). This can also be enforced by the smart rules supported by the DECODE architecture.

The minimise strategy is also captured by 'datensparsamkeit'[12], a German word that is hard to translate into English. It is a concept from privacy laws that is an opposite philosophy to "capture-all-the-things".

4.2. Separate

Definition: Prevent correlation of personal data by separating the processing logically or physically.

A much less obvious strategy to make systems more privacy friendly, but one with a profound impact, is to use the concept of separation. Instead of collecting and processing all personal data in one central location, the idea is to separate the processing either logically or physically. This prevents the unwanted combination of several pieces of personal data, each collected for a different purpose, to be combined into rich personal profiles. Viewed in this manner, separation is a natural strategy to implement *contextual integrity*. In the extreme case separation may even guarantee that the service provider itself (i.e the data controller) does not get access to the personal data at all!

This makes the separate strategy a strong alternative approach for those (many) cases where it is hard to minimise the personal data you collect. For example when designing a personalised service, or when the primary functionality of the service simply processes a lot of personal data (like in health care, or in the financial industry). In those cases, the separate strategy offers a different approach that still protects the data.

Associated tactics:

Distribute: partitioning personal data so that more access is required to process it.

[12] https://martinfowler.com/bliki/Datensparsamkeit.html

Isolate: processing parts of personal data independently, without access or correlation to related parts.

Example / impact on DECODE:

DECODE's use of a federated distributed storage based on blockchain technology, where the DECODE nodes each independently contribute their resources (storage and compute cycles) in a controlled manner to the overall system implements the "*distribute*" tactic.

DECODE in the end provides a platform on top of which many different applications may run simultaneously. Isolation can be achieved by tagging data items with the application they were collected or created for, to avoid reuse of data items for other applications. The smart rules (that typically are at the core of such applications) can be used to force this isolation (by checking the tags of all data items they process), but we note this is a type of isolation 'by convention' only, that in theory can be bypassed,

4.3. Abstract

Definition: Limit as much as possible the amount of detail of personal data being processed.

While `minimise' forces one to decide whether or not to process a particular piece of personal data, `abstract' addresses the more subtle question of the level of detail in which to process personal data. The less detailed a personal data item is, the more we `zoom out', the lower the privacy risk is.

Associated tactics:

Summarise: extracting commonalities in personal data by finding and processing correlations instead of the data itself.

Group: inducing less detail from personal data prior to processing, by allocating into common categories.

Perturb: add noise or approximate the real value of a data item.

Example / impact on DECODE:

There may be a way for DECODE to apply the Summarise tactic if it provides a query engine for aggregated data that removes all details and only provides aggregated information. Perturbation can perhaps be used in an ad-hoc fashion to the storage of raw data in some cases.

The use of this strategy can also be explored within the Decidim Petitions. When users are signing a petition, any additional data such as their date of birth, postal code, gender can be abstracted as groupings or aggregations of data instead of gathering raw data.

General examples of this strategy are recording the age instead of date of birth, postal code instead of specific address, etc.

4.4. Hide

Definition: protect personal data, or make them unlinkable or unobservable. Prevent personal data becoming public. Prevent exposure of personal data by restricting access, or hiding its very existence.

In contrast to the minimise strategy that forces one to decide whether or not to process certain personal data at all, this important strategy focuses on the protection of this data once it is decided the data is really needed. Recall that the adequate protection of personal data is often a legal requirement (at least it is in the GDPR). It also addresses the requirement to prevent inadvertent collection of metadata when ordinary data is being collected.

Hiding personal data can be achieved by protecting it (you know it is there but you cannot access it), making it unlinkable (you know the data, but not to which person it belongs), or making it unobservable (you are not even aware of the existence of the data).

Associated tactics:

Restrict: preventing unauthorized access to personal data.

Mix: processing personal data randomly within a large enough group to reduce correlation.

Encrypt: encrypt data (in transit or at rest)

Obfuscate: preventing understandability of personal data to those without the ability to decipher it.

Dissociate: removing the correlation between different pieces of personal data.

Example / impact on DECODE:

Within DECODE access to data, including to personal data, is restricted based on the concept of entitlements. Only an actor possessing the necessary entitlements can access a resource. Entitlements are implemented based on the concept of attribute based credentials (ABC), that in fact provide another layer of privacy protection due to the fact that they are unlinkable to individuals (unless, of course, the attributes used identify a person by name or number). Finally, all personal data in DECODE is encrypted.

4.5. Inform

Definition: provide data subjects with adequate information about which personal data is processed, how it is processed, and for what purpose.

Transparency about which personal data is being processed, how they are processed and for which purpose, is an essential (though not sufficient) prerequisite for better privacy protection. It allows users to take informed decisions about using a system and agreeing to the processing of their personal data. Moreover, it allows society at large to

verify whether organisations are processing our personal data responsibly. The inform strategy is closely related to the control strategy to be discussed next.

Associated tactics:

Supply: making available extensive resources on the processing of personal data, including policies, processes, and potential risks.

Notify: alerting data subjects to any new information about processing of their personal data in a timely manner.

Explain: detailing information on personal data processing in a concise and understandable form.

Example / impact on DECODE:

DECODE aims to implement this strategy in several ways. First of all, there is a strong focus on open source hard and software, both for the DECODE nodes as the interconnecting infrastructure. Second, transparency and auditability are core technical values. This is exemplified by the use of open blockchain technology for the mediation of all transactions involving personal data. Finally, DECODE aims to deploy declarative and intelligible smart-rules that can be well related to legal taxonomy, and that can be used to provide adequate information to data subjects and all the necessary notifications on the data processing.

4.6. Control

Definition: provide data subjects mechanisms to control the processing of their personal data.

Providing control is a fundamental principle to protect the privacy of users. The main goal of privacy is not to totally prevent the processing and sharing of personal data. Not at all! But users want to have control and have a say in how their personal data is processed and shared. Together with the inform strategy, control forces you to address both consent as well as data subject access rights that are both cornerstones of the GDPR.

Associated tactics:

Consent: only processing the personal data for which explicit, freely-given, and informed consent is received.

Choose: allowing for the selection or exclusion of personal data, partly or wholly, from any processing.

Update: providing data subjects with the means to keep their personal data accurate and up to date.

Retract: honouring the data subject's right to the complete removal of any personal data in a timely fashion.

Example / impact on DECODE:

The main element of control provided to users of the DECODE platform is by expressing entitlement conditions for the access to individual pieces of personal data. A graphical user interface and an intuitive language to express these conditions and the associated entitlements and smart rules will be developed to support the user to exert his or her control.

4.7. Enforce

Definition: commit to a privacy friendly way of processing personal data, and enforce this.

Privacy should not only be guaranteed through technical means, but also through organisational means. It should be part of the organisational culture and be propagated by higher management. Otherwise nobody will feel responsible. A clear privacy policy will provide scope and guidance. The enforce strategy is internally oriented, towards the organisation itself. The strategy ensures that the externally communicated privacy statement (see the inform strategy) is also enforced internally through a privacy policy.

The creation of a privacy policy, that is subsequently maintained and upheld, is central. One particular approach is to implement a privacy management system similar to the plan-do-check-act cycle from the information security management standard (ISO 27001). This could be integrated with the data protection impact assessments (DPIA) that has to be performed in any case (and that is further discussed under the demonstrate strategy).

Associated tactics:

> *Create*: acknowledging the value of privacy and deciding upon policies which enable it, and processes which respect personal data.
>
> *Maintain*: considering privacy when designing or modifying features, and updating policies and processes to better protect personal data.
>
> *Uphold*: ensuring that policies are adhered to by treating personal data as an asset, and privacy as a goal to incentivize as a critical feature.

Example / impact on DECODE:

Services provided over the DECODE infrastructure should specify a clear privacy policy. The DECODE infrastructure itself should provide tools for easy enforcement of such privacy policies. This is partially achieved through the concept of entitlements, matching attributes, and the use of smart rules that govern the access to personal data.

4.8. Demonstrate

Definition: provide evidence that you process personal data in a privacy friendly way.

This strategy addresses the new requirement in the GDPR that organisations need to demonstrate compliance to privacy regulations. The demonstrate strategy is externally oriented, towards the data protection authorities (possibly through the internal data protection officer).

Associated tactics:

Log: tracking all processing of data, without revealing personal data, securing and reviewing the information gathered for any risks.

Audit: examining all day to day activities for any risks to personal data, and responding seriously to any discrepancies.

Report: analyzing collected information on tests, audits, and logs periodically to review improvements to the protection of personal data.

Example / impact on DECODE:

So called Data Protection Impact Assessments (DPIA), also for the cases in which is not mandatory, can be a good tool for assuring certainty, transparency of the processing and fostering trust in the architecture (mainly for public sector bodies and companies who are committed to adopt a DPIA). It is recommended to each service provider offering services over the DECODE architecture to perform such a DPIA.

The DECODE architecture itself facilitates this through the use of open blockchain technology for the mediation of all transactions involving personal data, thus providing accountability as a built in feature.

5. Privacy strategies for the Barcelona pilots

5.1. DECODE IoT ecosystem

The emergence of Internet of things (IoT) has led to large-scale analyses of data generated from heterogeneous devices in various scientific and governance domains. In this sense, the SmartCitizen project was the baseline of the DECODE IoT ecosystem. This project allowed the implementation of a crowd sensing initiative for environmental monitoring, deploying methodologies for community engagement and co-creation. Smart Citizen developed tools for citizen actions, namely: to create public local maps of noise, temperature and air quality and to use them to raise awareness and find solutions for issues that matter to the respective community.

In the scope of DECODE, the interactive dashboard "BarcelonaNow" (Marras et al, 2018) allows several data processing tasks, including data acquisition (Open Data and Private/Sensitive data), fusion, aggregation and integration. In the case of CitizenSensing (IoT) pilot it should be emphasized that individual and or community data streams that potentially lead to a privacy threat are identified, encrypted and anonymized before data processing in BarcelonaNow, to prevent any privacy breaches.

Therefore, we will consider several concerns regarding identification, location tracking and profiling, as well as threats raised by the integration of data about users of IoT networks and open data released by the government.

On the one hand, IoT information integration tasks in BarcelonaNow are considered crucial to gain new insights from environmental sensor data storage, analysis and visualization. On the other hand, the autonomous nature of IoT exposes some privacy threats, in which the DECODE architecture needs to ensure that the privacy of individuals and/or community are not compromised.

5.1.1. Information linkage in DECODE IoT ecosystem

In case of DECODE IoT ecosystem, in the phase of data integration the attributes and policies provided by the Smart Citizen infrastructure & IoT wallet, and the respective standardized encrypted data belonging to different sensors (e.g. temperature and noise) of an individual or a community will be stored in an encrypted datastore, and therefore decrypted and consumed in BarcelonaNow. We will not go into the details of all system architecture for IoT Pilot which is described in deliverable D5.4 "Prototype data visualization tool". Here we will only focus on the personal privacy of the data aggregation and standardization phases considered in BarcelonaNow.

However, it should be noted that in this process sensitive and personal information is revealed, such as demographic location and indoor or outdoor sensor measurements, which lead to privacy concerns. Therefore, user profiling, localization and tracking, and information linkage are some of the critical challenges that were addressed for a

secure data standardization, data encryption, and data visualization in the DECODE IoT ecosystem.

5.1.2. Considerations for privacy concerns

DECODE IoT ecosystem comprises a number of heterogeneous sensors in SmartCitizen devices, performing a variety of tasks that gather and distribute environmental data about surroundings (indoor and outdoor) of the involved individuals and communities. Thus, in Smart Citizen Onboarding application, the context for which the devices and services are authorized to collect data may vary from the context of smart contracts and policies of data-use (public or private).

In addition, the information linkage can only occur in SmartCitizen devices and at metadata levels, but not at data-stream in Stream Encoder and visualization level. BarcelonaNow was designed with privacy protections related to integration, standardization and data aggregation to avoid the information linkage, and possible implications on data privacy. During data integration phase, attributes of IoT devices and data belonging to different services (such as listings of Airbnb) can be correlated. At times, this can reveal information or insights about subjects, demographic location and activities of individuals, which lead to severe privacy concerns. Therefore, user profiling, localization and tracking, and information linkage are some of the critical challenges that need to be addressed for data processing in DECODE IoT ecosystems.

Currently, legal issues under the General Data Protection Regulation (GDPR) are being discussed, to reduce the risk of information linkage and protection of user's rights over data-use and sharing.

5.2. Privacy in encrypted data integration: CitizenSensing (IoT) technical solutions

As described in deliverable D5.4 "Prototype data visualization tool", there will be three data integration steps between BarcelonaNow and other components in the DECODE system, namely:

1. *Create Data source*: consists in a manual process during which a private/public key pair is generated in IoT wallet and a new data source is registered in BarcelonaNow with its own code for the IoT data collector/decryptor.
2. *Collect data*: This is the secure data reading from encrypted data store and storing it in the BarcelonaNow server. The technical details of this flow, and of how the data will be consumed from this new secure data source is given in section 5.1.3 of the abovementioned deliverable.
3. *User Login*: the user management for CitizenSensing (IoT) pilot is done outside BarcelonaNow, where the user credentials and their affiliation to the community and right to access its data will be validated by the IoT Wallet. The technical details of this integration are given in section 5.1.2 of the deliverable D5.4.

5.3. Technical caution to legal protection

5.3.1. Data model for CitizenSensing (IoT)

The encrypted data models for different sensor measurements, as discussed and analyzed with the partners involved in the CitizenSensing (IoT) pilot, will be implemented according to the DECODE standards. Namely the JSON structure considered in other data collectors and explored in BarcelonaNow (see an example in Figure 3).

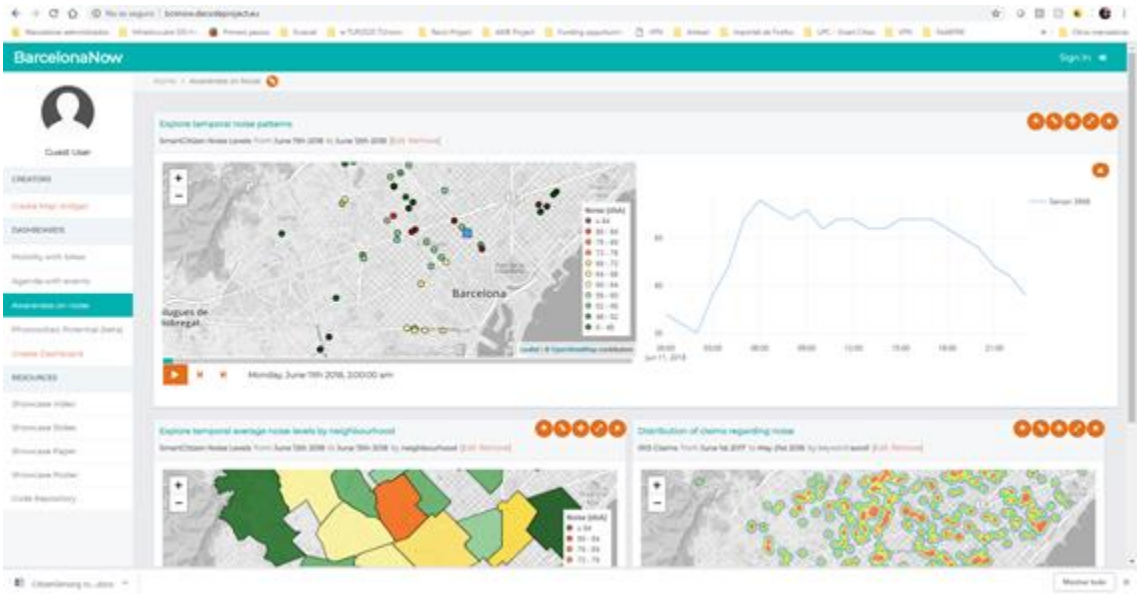


Figure 3 Temporal noise patters based on Sentilo data collector

Example output for sharing a single CitizenSensing (IoT) sensor in full resolution:

```
{
  "location": {
    "longitude": 2.234,
    "latitude": 54.213,
    "exposure": "INDOOR"
  },
  "recordedAt": "2018-06-03T22:26:02Z",
  "sensors": [
   {
      "id": 29,
      "name": "MEMS Mic",
      "description":
        "MEMS microphone with envelope follower sound pressure sensor (noise)",
      "unit": "dBC",
      "type": "SHARE",
      "interval": 900,
      "value": 64.252
   }
  ]
}
```

Example output for a binned CitizenSensing (IoT) sensor:


```
{
  "location": {
    "longitude": 2.234,
    "latitude": 54.213,
    "exposure": "INDOOR"
  },
  "recordedAt": "2018-06-03T22:26:02Z",
  "sensors": [
   {
      "id": 29
      "name": "MEMS Mic",
      "description":
        "MEMS microphone with envelope follower sound pressure sensor (noise)",
      "unit": "dBC",
      "type": "BIN",
      "bins": [40, 80],
      "values": [0, 1, 0]
   }
  ]
}
```


Example output for a moving average CitizenSensing (IoT) sensor:


```
{
  "location": {
    "longitude": 2.234,
    "latitude": 54.213,
    "exposure": "INDOOR"
  },
  "recordedAt": "2018-06-03T22:26:02Z",
  "sensors": [
   {
      "id": 29,
      "name": "MEMS Mic",
      "description":
        "MEMS microphone with envelope follower sound pressure sensor (noise)",
      "unit": "dBC",
      "type": "MOVING_AVG",
      "interval": 900,
      "value": 64.252
   }
  ]
}
```


Example of a stream which combines sharing, binning and moving averages:


```
{
  "location": {
    "longitude": 2.234,
    "latitude": 54.213,
    "exposure": "INDOOR"
  },
  "recordedAt": "2018-06-03T22:26:02Z",
  "sensors": [
   {
      "id": 14,
      "name": "BH1730FVC",
      "description": "Digital Ambient Light Sensor",
      "unit": "lux",
      "type": "MOVING_AVG",
      "interval": 900,
```

```
      "value": 6.34
    },
    {
      "id": 29
      "name": "MEMS Mic",
      "description":
        "MEMS microphone with envelope follower sound pressure sensor (noise)",
      "unit": "dBC",
      "type": "BIN",
      "bins": [40],
      "values": [0, 1]
    },
    {
      "id": 12,
      "name": "HPP828E031",
      "description": "Temperature",
      "unit": "ºC",
      "type": "SHARE",
      "value": 22.41268
    }
  ]
}
```


As previously discussed we are in presence of sensitive data, as defined in Art. 9 (1), GDPR, and the IoT decrypted data to be stored and visualized in BarcelonaNow cannot identify a natural person and reveal racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership of a subject. In this sense, some of the attributes of the collected IoT data will be anonymized (e.g. "id", "name", "description", "longitude", "latitude" and "exposure").

5.3.2. Controller-Processor Contract or other legal act

There is a growing trend where customers are increasingly sharing their location data with map or navigation and IoT environmental services. To address privacy concerns and bring current privacy rights in accordance with digital age, EU is implementing the new General Data Protection Regulation (GDPR). Based on Art. 28 (3) GDPR, that governs the processing by the processor under a contract or other legal act, a contract or other legal act will be defined between IAAC (SmartCitizen) and Eurecat. Therefore, Eurecat will be responsible for processing personal data in BarcelonaNow on behalf of the controller (IAAC).

5.4. DECODE's Digital Democracy and Data Commons

In the Digital Democracy and Data Commons (DDDC) pilot, citizens are invited to share voluntarily their demographic data through two different strategies: (1) an initial survey to have a clear picture of the citizens in the participatory process and, (2) the petition signing process.

5.4.1. Survey

The initial survey is hosted at the Decidim instance with the aim of increasing the inclusiveness of the participatory process (Figure 4). The questions of the survey are as follows:

- What is your gender?
- What is your age?
- Where are you from?
- What is the highest educational level you have completed?
- What is your job situation?
- If you belong to a collective, NGO or an organization that deals with issues related to online privacy, data governance and / or technological sovereignty, put the name of your organization here.
- Where do you live?
- What is your district? [for those who live in Barcelona].
- What device do you use mostly to connect to the Internet?
- In a scale from 0 to 5, where 0 is "no at all" and 5 is "very much", how worried are you about the management of your data by internet companies?
- What are the issues that worry you the most about the current ways in which data is managed?

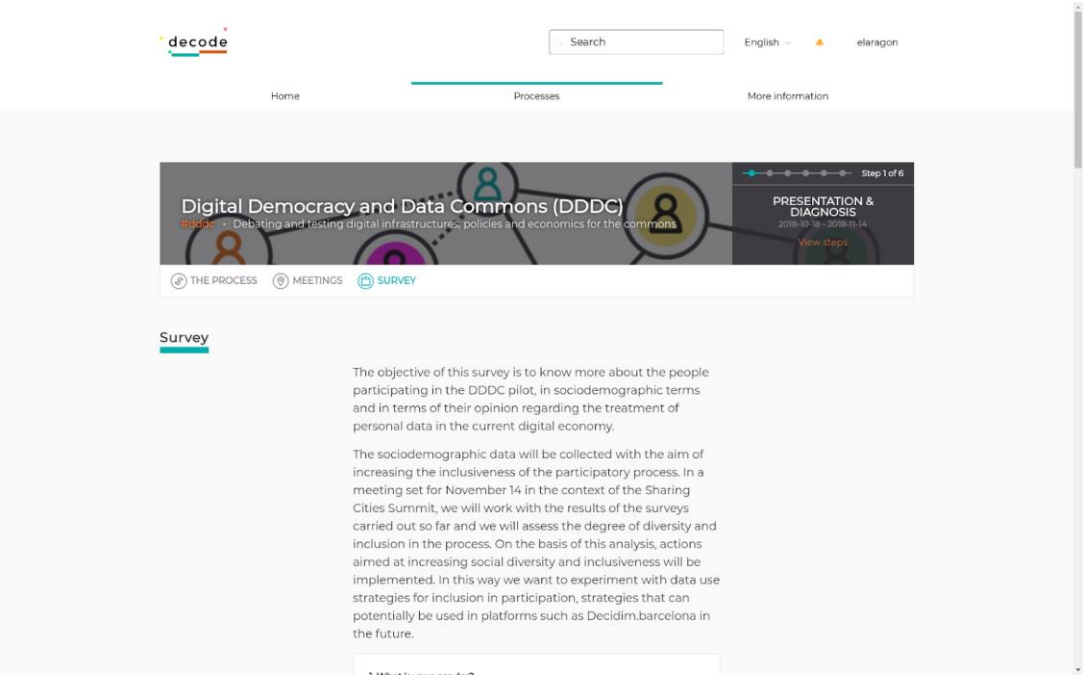


Figure 4 Decidim survey

Data from decidim surveys () can be retrieved from an admin account in three formats: CSV, JSON and XLS. Example of a survey response in JSON:


```
  {
    "1. Cuál es tu género?": [
      "Mi género se ve mejor representado por el término: "
    ],
    "2. Cuál es tu edad?": [
      "25-34 "
    ],
    "3. De dónde eres?": [
      "España"
    ],
    "4. De dónde eres?": [
      "Europa"
    ],
    "5. Cuál es el nivel educativo más alto que has completado?": [
      "Estudios universitarios (diploma, licenciatura, máster, doctorado…)"
    ],
    "6. Cuál es tu situación laboral?": [
      "Empleado/a "
    ],
    "7. Si perteneces a algún colectivo, ONG u organización que trata temas relacionados a la
privacidad online, la gobernanza de datos y/o la soberanía tecnológica, pon aquí el nombre de
tu organización": "Tecnopolítica",
    "8. Dónde vives?": [
      "En Barcelona"
    ],
    "9. ¿Cuál es tu distrito? [Para quienes viven en Barcelona]": [
      "Sant Martí"
    ],
    "10. ¿Qué dispositivo utilizas mayoritariamente para conectarte a Internet?": [
      "ordenador"
    ],
    "11.  En una escala de 0 a 5, donde 0 es “nada” y 5 es “mucho”, ¿cómo de preocupado/a
estás sobre cómo se gestionan tus datos por parte de las compañías de internet?": [
      "3"
    ],
    "12. ¿Cuáles son los temas que te preocupan más acerca de la forma en la que los datos
están gestionados actualmente?": [
      "Persuasión masiva"
    ]
  },
```


These data will be analyzed collaboratively in offline workshops along the pilot to validate social diversity and inclusiveness.

5.4.2. Petition signing

Citizens will be also invited to provide voluntarily their demographic data (gender, age range, district) during the signing phase of the participatory process. The data will be stored in a secure and transparent manner through the DECODE technology for secure and transparent signature and then retrieved in an aggregated manner to produce statistics useful for the community.

6. Privacy-preserving and non-discriminatory data mining

Living in the information society facilitates the automatic collection of huge amounts of data on individuals, organizations, etc. Publishing, mining and personalising of high data quality data for secondary analysis (e.g. learning models, finding patterns, personalise services) may be extremely useful to policy makers, planners, marketing analysts, researchers and others. Yet, data publishing and mining do not come without dangers, namely privacy invasion and also potential discrimination of the individuals whose data are published and analyzed.

Note that, some of the data mining and recommender systems may run on end-user's personal device and sharing personal data with a company/third party is not required and thus, a secure analysis is achievable. However, in the most of the cases the data mining and recommender system approaches are different because a big collection of end-users data has to be considered to create one single recommendation or decision. As a consequence, companies maintain huge databases for end-user data. The problem is how to design privacy-preserving solutions for data mining and recommender system to ensure the privacy issues:

1. Data ownership. Ensuring that data subjects/users own and control their personal data. As such, the system is designed by default to recognize end-users as the owners of data and the mining and recommendation services as guests with delegated permissions. In some cases (for instance when personal data are collected by companies who are the data owners) there will be needed to design different solutions to distinguish between the data subject (the person the data are referring to) from the data owner (i.e. public bodies or private companies who collected the data and who hold rights on data, for instance IPRs) and to recognize rights to end-users.
2. Data transparency and auditability. Each end-user has complete transparency over what data is being collected about her and how they are accessed.
3. Fine-grained access control. At any given time the end-user may alter the set of permissions and revoke access to previously collected data.

6.1. Current solutions for achieving privacy preserving data mining and recommendation systems

There have been various attempts to address the above privacy risks, not only from a legislative perspective, but also from a technological perspective:

1. Data anonymization methods attempt to protect personally identifiable information. k-anonymity, a common property of anonymized datasets requires that sensitive information of each record is indistinguishable from at least k−1 other records [Sweeney, 2002]. Related extensions to k-anonymity include l-diversity, which ensures the sensitive data is represented by a diverse enough set of possible values [Machanavajjhala et al., 2007]; and t-closeness, which looks at the distribution of sensitive data [Li et al., 2007]. Recent research has demonstrated how anonymized datasets employing these techniques can be de-anonymized [Narayanan et al., 2006], [Montjoye et al., 2013], given even a small amount of data points or high dimensionality data.
2. Differential privacy, a technique that perturbs data or adds noise to the computational process prior to sharing the data [Dwork, 2006]. The main drawback of these technique is that the utility of the perturbs data is low.
3. Encryption schemes that allow running computations and queries over encrypted data. Specifically, fully homomorphic encryption (FHE) [Gentry, 2009] schemes allow any computation to run over encrypted data, but are currently too inefficient to be widely used in practice.
4. Secure multiparty computation also known as distributed privacy-preserving data mining [Lindell et al., 2009]. In this scheme data are split into pieces, and shared among a distributed network of nodes or data owners. Computations are performed interactively and collaboratively on between the data owners. Thereby, individual data owner doesn't not get access to meaningful raw data, but only on encrypted shards or the final results of the computation. The approaches also have high communication and computation complexity.
5. Block-chain supported secure multiparty computation. In recent years, a new class of accountable systems emerged. The first such system was Bitcoin, which allows users to transfer currency (bitcoins) securely without a centralized regulator, using a publicly verifiable open ledger (or blockchain). Since then, other projects (collectively referred to as Bitcoin 2.0 [Evans, 2014]) demonstrated how these blockchains can serve other functions requiring trusted computing and auditability. A novel blockchain-based approach [Zyskind et al., 2015a] is able to cryptographically guarantee the proper usage of personal data. The core component is a decentralized peer-to-peer network that allows storing encrypted data in a tamper-proof way and runs secure computations while no one but the data owner has access to the raw data.

The main features of a blockchain, an immutable public log, and a programmable token of value, have been used to advance secure multiparty computation systems in terms of fairness and operational efficiency. Enigma [Zyskind et al., 2015b] implements those advancements in order to provide an open decentralized network for encrypted data storage and secure multiparty computations. Identity, access and contract management is facilitated by the underlying protocol. Private contracts provide the programming interface to access private and public data and to specify the computations. Thereby end users can permit and audit the usage of their data in fine granularity. Moreover they can revoke the permission at any time.

6.2. Our proposal for decentralised privacy-preserving mining and non-discriminatory recommendation[13]

Private data from the pilots are not available through the DECODE infrastructure at the moment in which this deliverable is written. However, we have already defined an approach for their privacy-preserving treatment, building on the state of the art methods, findings and known issues detailed in the above section.

6.3. Privacy-preserving aggregation and mining of demographic data

In the case of the DDDC pilot we will have to deal with demographic data in the sensitive context of politics and online petitions. In order to avoid the possibility of de-anonymization of the data, several measures will be taken:

- the possible values of the attributes will be split into classes general enough to avoid individual citizens to be identified as clearly fitting a very specific attribute value (e.g.: treating age with age ranges, origin by continents and not by countries…).
- the different attributes will be received and stored in decoupled way, so to avoid the possibility of de-anonymization by combining different attributes to get a specific class in order to fall in the case above.
- to avoid de-anonymization through temporal correlation, data will be accumulated in an encrypted way, and will be readable only at the end of the process.

In this way, we will ensure privacy-preserving data aggregation and mining, robust to possible attempts to mine sensible data and de-anonymize it through the values of specific attributes or temporal correlations, and thus breaking privacy of any individual citizen.

6.4. Privacy-preserving aggregation of geographic data

In the context of the IoT pilot, data that comes with geographic location is very sensitive for user privacy as discussed extensively in the previous section. Indeed, the patterns of accumulated geospatial information can give away crucial information, even if it is successfully anonymised. These IoT patterns are highly valuable for end-users,

[13] WP3 Blockchain for decentralised data and digital identity management - task 3.3

organisations and companies, but they can be misused. We hereby describe the strategies for robust privacy-preserving aggregation of this data.

6.4.1. Geomasking and GeoAggregation

According to GDPR location data is considered as "personal data" in Article 4 (1). Under this clause personal data are granted extended rights, including a right to access and a right to erasure. Location data is extremely personal and valuable. Considering its complexities, it is difficult to foresee as to how many ways location data could be used and misused in the future. Therefore, there is an immediate need to ensure secure DECODE standards and also set measures about how data is being used. To ensure the data privacy of geo-located data shared by individuals and/or communities two geoprocessing operations will be implemented.

1. Anonymization of location data with geomasking: geomasking is a class of methods to change the geographic location of an individual in an unpredictable way to protect confidentiality, while trying to preserve the relationship between geocoded locations and indoor or outdoor environmental sensor measures. This technique will be used to provide privacy protection for individual address information while maintaining spatial resolution for mapping purposes. Donut geomasking and other random perturbation geomasking algorithms will be tested (Figure 5).

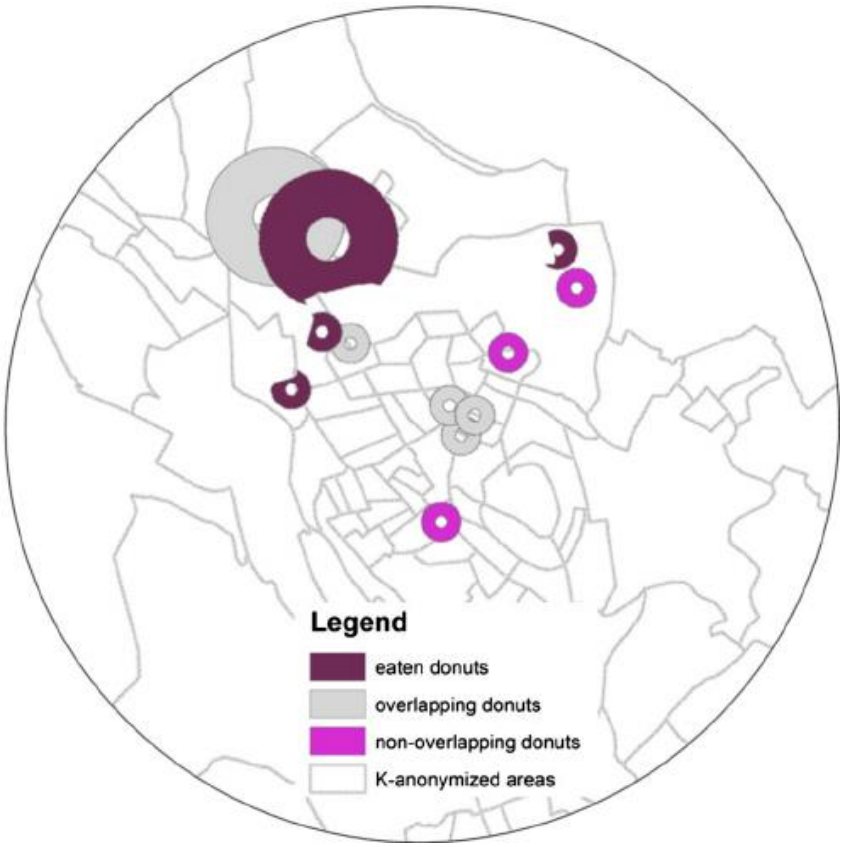


Figure 5 Example of Donut geomasking technique

2. GeoAggregation: aggregating data has often been used to protect the individual in a database or visualization system, but this method substantially reduces the resolution of the data available to the end-user. This technique will be used to group spatial data by a relational attribute (sensor measurement of an individual or community) and also at different granularity levels (district and neighborhoods of Barcelona). Thus, the Map & Feature Services of BarcelonaNow (Figure 6), that make use of location data in the spatial temporal store, will enable the IoT users to visualize on-the-fly aggregation of IoT data: inspect feature level attributes while in aggregation or masked feature views; replay via a time-slider historic observations in aggregation or masked feature views.

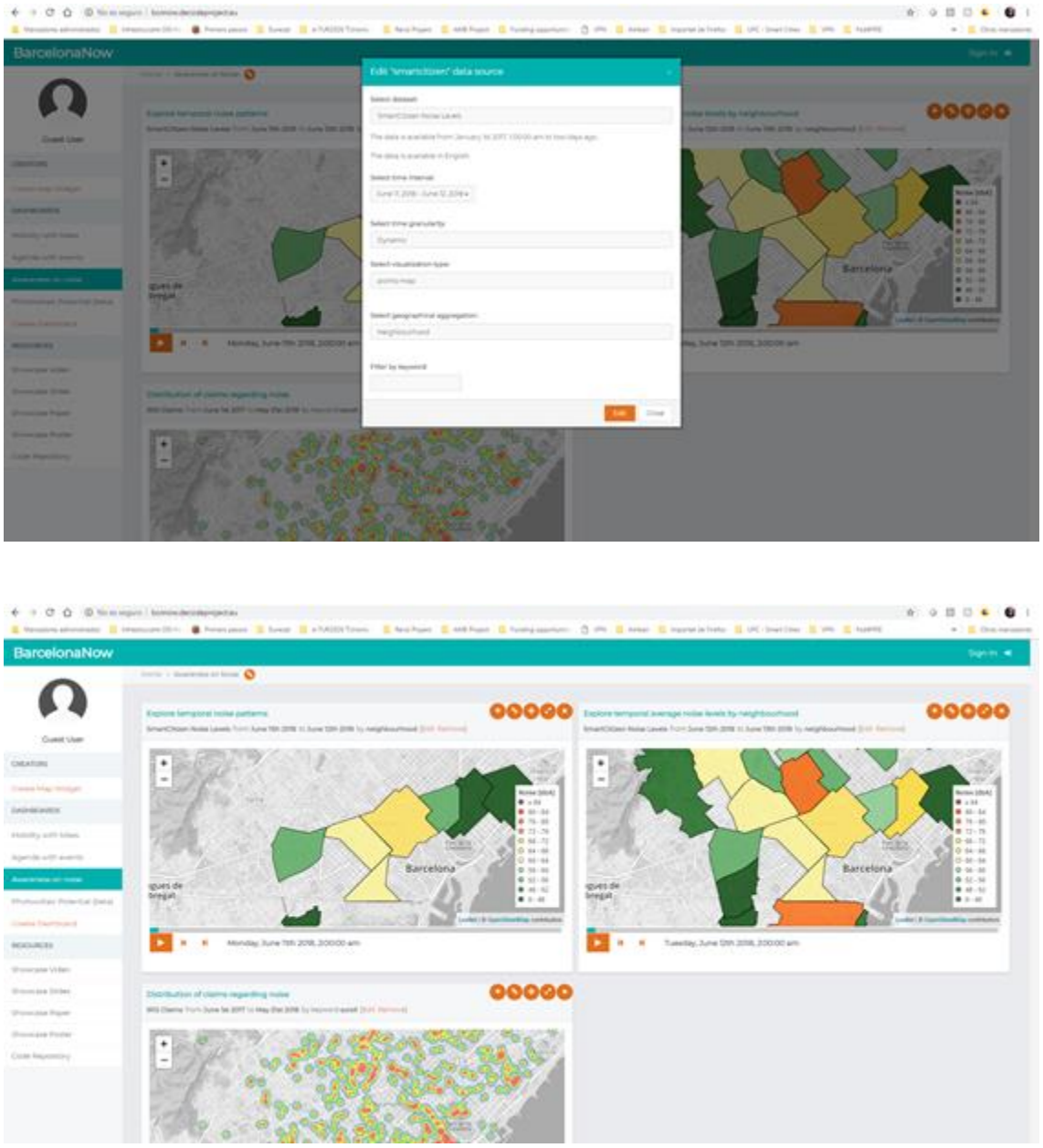


Figure 6 GeoAggregation of SmartCityizen data in BarcelonaNow

7. Conclusions

We have outlined the privacy issues relevant to DECODE, based the current architecture. Our most important findings are as follows.

1. Verifiers should be able to validate transactions without learning secrets and confidential data within the transaction.
2. Our analysis based on the privacy design strategies shows that the DECODE architecture is promising in the inherent privacy preserving properties it exhibits.

We have also outlined the privacy issues and technical solutions relevant to CitizenSensing (IoT) and Digital Democracy and Data Commons (DDDC) pilots, namely:

1. CitizenSensing (IoT) and Digital Democracy and Data Commons (DDDC) pilots will consider privacy strategies and legal issues under the GDPR (in the IoT pilot, it will be formalized as a Controller-Processor Contract or other legal act) to reduce the risk of information linkage and protection of user's rights over data-use and sharing.
2. The IoT and DDDC data will be stored in a secure and transparent manner through the DECODE technology retrieved in an aggregated manner to produce statistics and interactive visualizations useful to both pilot communities.
3. In the case of IoT pilot, BarcelonaNow will integrate technical solutions to ensure privacy-preserving aggregation of geographic data (two geoprocessing operations will be implemented: Geomasking and GeoAggregation).
4. In the case of the DDDC pilot, a decentralised privacy-preserving mining and non-discriminatory recommendation is proposed, through the Implementation of several measures to ensure a privacy-preserving aggregation and mining of demographic data.

We recommend the following:

1. Use the advantages of blockchain-supported secure multiparty computation, in order to design a secure and private data mining and recommender system.
2. When refining the DECODE architecture in more detail, one needs to take the observations made in the section on privacy design strategies into account. Especially, some effort needs to be spent deciding how to address the inform, control, enforce and demonstrate strategies.
3. Implement technologies (including smart contracts) that allow the data subjects to execute some of their rights "by design" (right to data portability, right to be forgotten, withdrawal of consent, etc.).

4. DECODE supports user-defined smart contracts that encode the business logic of specific applications. We recommend that to support privacy-friendly contracts, the design should employ mechanisms for verifiers to check validity of smart contracts without having to learn private/confidential data within the contracts.

8. References

[Berberich & Steiner, 2016] M. Berberich, M. Steiner, Blockchain Technology and the GDPR – How to Reconcile Privacy and Distributed Ledgers?, EDPL (3) 2016, pp. 422-426.

[Colesky et al., 2016] M. Colesky, J-H. Hoepman, C. Hillen, A critical analysis of Privacy Design Strategies, 2016 IEEE Security and Privacy Workshops, pp. 33-40.

[De Filippi, 2016] P. De Filippi, The Interplay between Decentralization and Privacy: The Case of Blockchain Technologies. Journal of Peer Production, 2016, Issue n.7: Alternative Internets . Available at SSRN:

[Dwork, 2006] C. Dwork. Differential privacy. In *Automata, languages and programming*, pages 1–12. Springer, 2006.

[Evans, 2014] J. Evans. Bitcoin 2.0: Sidechains and ethereum and zerocash, oh my!, 2014.

[Fink, 2017] M. Finck. Blockchain and Data Protection in the European Union. In *Max Planck Institute for Innovation & Competition, Research Paper No. 18-01*, 2017 (Available at SSRN: or).

[Gentry, 2009] C. Gentry. Fully homomorphic encryption using ideal lattices. In *STOC*, volume 9, pages 169–178, 2009.

[Hoepman, 2014] J-H. Hoepman, Privacy Design Strategies, IFIP TC11 29th Int. Conf. on Information Security (IFIP SEC 2014), pp. 446-459.

[Ibáñez, K. O'Hara and E. Simperi, 2018] L-D. Ibáñez, K. O'Hara and E. Simperi, On Blockchains and the General Data Protection Regulation, University of Southampton, June 2018.

[Koops & Leenes, 2014] B-J. Koops, R. Leenes, Privacy regulation cannot be hardcoded. A critical comment on the 'privacy by design' provision in data-protection law', 28 International Review of Law, Computers & Technology (2), 2014, pp. 159-171.

[Kurtz, Semmann, Böhmann, 2018] C. Kurtz, M. Semmann, T. Böhmann, , Twentyfourth Americas Conference on Information Systems, New Orleans, 2018, available at: https://www.researchgate.net/profile/Martin_Semmann/publication/325415927_Privacy_by_Design_to_Comply_with_GDPR_A_Review_on_ThirdParty_Data_Processors/links/5b1f77f1458515270fc4cf93/Privacy-by-Design-to-Comply-with-GDPR-A-Review-on-Third-Party-Data-Processors.pdf

[Laurie et al., 2013] Ben Laurie, Adam Langley, and Emilia Kasper. Certificate transparency. Technical report, 2013.

[Li et al., 2007] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, volume 7, pages 106–115, 2007.

[Lindell et al., 2009] Y. Lindell and B. Pinkas. "Secure multiparty computation for privacy-preserving data mining." Journal of Privacy and Confidentiality 1.1 (2009): 5.

[Machanavajjhala et al., 2007] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k- anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3, 2007.

[Marras et. Al. 2018] Marras, M., Manca, M., Boratto, L., Fenu, G., & Laniado, D. (2018, April). BarcelonaNow: Empowering Citizens with Interactive Dashboards for Urban Data Exploration. In Companion of the The Web Conference 2018 on The Web Conference 2018 (pp. 219-222). International World Wide Web Conferences Steering Committee.

[Montjoye et al., 2013] Y. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.

[Nakamoto, 2008] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. 2008.

[Narayanan et al., 2006] A. Narayanan and V. Shmatikov. How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*, 2006.

[Pagallo, 2012] U. Pagallo, Cracking down on autonomy: three challenges to design in IT Law, Ethics Information Technology (14) 2012, pp. 319-328.

[Pagallo et al., 2018 (forthcoming)] U. Pagallo, E. Bassi, M. Crepaldi, M. Durante. Chronicle of a Clash Foretold: Blockchains and the GDPR's Right to Erasure, In: Frontiers in Artificial Intelligence and Applications (FAIA), IURIX proceedings 2018

[Roio et al, 2017] Denis Roio, James Barritt, Jaap-Henk Hoepman, Mark de Villiers, Tom Demeyer (2017), DECODE Whitepaper (in preparation)

[Sweeney, 2002] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[Wood, 2014] Gavin Wood. Ethereum: A secure decentralised generalised transaction ledger. Ethereum Project Yellow Paper, 151, 2014.

[Zyskind et al., 2015a] G. Zyskind and O. Nathan. "Decentralizing privacy: Using blockchain to protect personal data." Security and Privacy Workshops (SPW), 2015 IEEE. IEEE, 2015.

[Zyskind et al., 2015b] G. Zyskind, O. Nathan, and A. Pentland. "Enigma: Decentralized computation platform with guaranteed privacy." arXiv preprint arXiv:1506.03471 (2015).