BPDIMS:A Blockchain-based Personal Data and Identity Management System

*Benedict Faber[1], Georg Michelet[1], Niklas Weidmann[1], Raghava Rao Mukkamala[1,2], Ravi Vatrapu[1,2]
[1]Centre for Business Data Analytics, Copenhagen Business School, Denmark
[2]Department of Technology, Kristiania University College, Oslo, Norway
{rrm.digi, rv.digi}@cbs.dk

Abstract

Recent scandals on the abuse of personal information from social media platforms and numerous user identity data breaches raise concerns about technical, commercial, and ethical aspects of privacy and security of user data. European Union's new General Data Protection Regulation (GDPR) is one of the largest changes in data privacy regulation and entails several key regulatory measures for both data controllers and data processors to empower and protect EU citizens' privacy. In this research work, we propose a conceptual design and high-level architecture for a Blockchain-based Personal Data and Identity Management System (BPDIMS), a human-centric and GDPR-compliant personal data and identity management system based on the blockchain technology. We describe how BPDIMS's architecture utilizes blockchain technology to provide a high-level of security, trust and transparency. We discuss how BPDIM's human-centric approach with GDPR compliance shifts the control over personal data to the end users and empowers them better.

1. Introduction

Our lives have become increasingly digital and so has the vast amount of personal data traces that we leave behind. The current situation is that a few large multinational corporations make the majority of profits through offering services users pay for with their data. While data analytics can provide users with better services, the users' overview and control of their personal data has decreased. Moreover, the recent Cambridge Analytica scandal of misusing people's personal information from Facebook to influence voters in the US Elections 2016[1] has raised serious concerns about the technical, commercial, political and ethical aspects of personal data collection and analysis by platform owners such as Facebook and other third parties.

In May 2018 the European Union's new GDPR [1] came into effect. While aiming to protect the users, the new regulation can potentially be a burden for companies [2]. While the GDPR aims to give control of personal online data to European users through new regulation, several further initiatives have been launched both from private and public spheres, to argue for a human-centric approach to personal information [3, 4]. In 2014 the Finnish government published a study on the concept of *MyData* [4]. MyData facilitates the idea that users should have a better overview of where their data is stored, who uses it, and be able to change this. It is a human-centric approach to people's data and aimed at giving control of personal data back to the users. On a different note, blockchain technology generated significant research interest and industry attention in recent years mostly due to the hype and success created by the cryptocurrencies. For example, Bitcoin was first described in 2008 [5] and ever since has attracted the attention of the research community from diverse academic fields [6, 7] and gained mainstream popularity due to its disruptive characteristics, such as the absence of centralised control and high degree of anonymity. Applications which were previously run through a trusted intermediary, can now - using blockchain technology - operate more transparently in a decentralised mode without the need of having a central authority and in a much more transparent way [8]. We addresses the problem of personal data identity and management by adopting a human-centric approach that ensures a GDPR compliance by employing blockchain-based technologies.

Currently users lack transparency over which service is processing their personal data for which purpose and possibly handing over personal data to third party providers without the user's knowledge. This is partly due to extensive and complicated terms and conditions of a service and the user requirement to agree to these, if they wish to use the service. Moreover, there are no suitable mechanisms that enable users to opt-out from

*The first three authors contributed equally for first authorship.
[1]https://www.theverge.com/2018/3/16/17132172/facebook-cambridge-analytica-suspended-donald-trump-strategic-communication-laboratories

URI: https://hdl.handle.net/10125/60121
ISBN: 978-0-9981331-2-6

a service gracefully, e.g. deleting all the history of using the service from the service provider. And lastly, currently there is a lack of systems that enable users in an effective and user-friendly way to obtain an overview of the usage of their personal data and to exercise fine-grained control over the usage of their personal data. While the GDPR addresses the aspects of transparency and consent and puts the legislation in place to enforce appropriate mechanisms, the latter issue of user control has not sufficiently been solved yet. Furthermore, after users have gained full transparency, they need adequate means to control the consent that is connected to the usage of their personal data. The GDPR will put the regulation in place to empower the user to request deletion of or revoke consent to use their personal data. However, there is a need to research and develop a system that facilitates this request or revocation of personal data. The main focus of this research work is to come up with a conceptual design for such a system called Blockchain-based Personal Data and Identity Management System (BPDIMS) that empowers users to get full transparency and control over the usage of their personal data. Consequently, the overarching research question is:

> *How blockchain can be utilised to develop a system for personal data and identity management which is human-centric and GDPR compliant?*

The rest of the paper is organised as follows: First, we provide a concise description of related work. In the next section (sec. 3) we describe the theoretical foundations of relevant concepts. Sec. 4, introduces and describes our proposed conceptual BPDIMS design, while sec. 5 describes different use case scenarios and the system functionality. In the last two sections (sec. 6 & 7), we discuss technical and usage aspects of the BPDIMS and conclude with future work.

2. Related Work

We limit our discussion to the systems and architectures that proposed personal data management using blockchains. Blockchain technology is still evolving and the number of applications using blockchain are slowly increasing. However the applicability of blockchain technology for personal data management is not well-explored yet. One of the first contributions in this direction is [9], where a protocol was developed which turns a blockchain into an automated access-control manager for a decentralised personal data management system. Use of auditable contracts deployed on blockchain infrastructures for a transparent data access, sharing/processing of personal data by data owners and a privacy-preserving architecture was proposed in [10]. Similarly, a framework for aggregating online identity and reputation information based on social dependency network to provide online behavioural ratings is proposed in [11].

In the healthcare domain several studies explored blockchain technology for the medical data access. A seminal and highly relevant contribution is [12], where authors proposed an architecture based on artificial intelligence and blockchain technology to enable control of their personal data including medical records to the users. In the similar lines a decentralised record management system to handle electronic patient records using blockchain technology was proposed in [13]. The research work in [14] proposed a mobile app architecture based on blockchain to enable patients to own, control and share their own data easily and securely without violating privacy. When compared to the existing research, our research proposes a new conceptual design and system architecture for human-centric personal data and identity management based on the MyData initiative, by using blockchain and smart contracts technology that is in compliance with the forthcoming GDPR [1].

3. Theoretical Foundations

GDPR: The GDPR [1] is one of the largest changes in data privacy regulation in recent history and came into effect in May 2018 in place of Data Protection Directive from 1995. The key aim is to harmonise data privacy laws across Europe and particularly to empower and protect EU citizens' privacy. One of the most central issues is the question of user's consent. The regulation states that the service provider must show what the user's consent is for and it should be easy for an user to withdraw his consent. If the user withdraws his consent or if there are changes in data usage other than what the consent is for, then the service provider required to delete the data related to the specific user. Furthermore, it is the user's right to access, meaning that on the user's request the service provider must provide an overview of whether the user's personal data is being processed and the purpose of processing. The service provider must also provide all data to the user in a machine-readable format. Similar to the right to access is the right to data portability; the user should be able to get an extract of his personal data from the controller in a machine readable format and has the right to transfer his data to another controller. Voilations of GDPR can result in large fines for companies of up to 20 million Euros or 4% of global turnover, whichever is larger [1].

MyData Human-Centric Personal Data Management: MyData [4] is a concept that refers to a paradigm shift from current organisation-centric focus to human-centric focus in personal data management. The pri-

mary idea behind MyData is that users should have a better overview of where their data is stored, who uses it, and be able to influence/decide who can use it and what it is being used for. In other words, it's a concept aimed at giving the control over their personal data back to the users. This is achieved through a human-centric approach that empowers the users by placing them in the centre of the *data ecosystem*. MyData intends to change the infrastructural approach so it ensures data portability and interoperability through open infrastructures. Furthermore, the concept is consent-based, so the user can control the flow of data without storing the data on centralised repositories. Lastly, the MyData approach facilitates data sharing across sectors with the goal of advancing the benefits of data sharing and usage which would profit the users, businesses, and society as a whole. Main objectives from the user perspective are: 1) right to know what personal information exists, 2) right to see the content of personal information, 3)right to rectify false personal information, 4) right to audit who accesses personal information and why, 5) right to obtain personal information and use it freely, 6) right to share/sell personal information to others., and 7) right to remove or delete personal information.

3.1. Blockchain

Blockchain is the decentralised distributed database technology that is combined with guarantees against tamper-resistance of transactions/records using cryptographic methods. By using time-stamping of its transactions and messages, blockchain provides universally verifiable proofs for existence or absence of a transaction in the distributed database and the underlying cryptographic primitives using hash functions and digital signatures provide guarantee that these proofs are computationally secure and verifiable at any point in time. Blockchain is decentralised, jointly maintained by a plurality of independent parties/nodes and achieves consistency of transactions among distributed nodes by using distributed consensus protocols (such as Byzantine fault tolerance algorithm [15]) without the need of having a central authority. Blockchain transactions are transparent and visible to all users of the system and at the same time blockchain provides anonymity to its users by allowing them create pseudo-anonymous transactions without the need for disclosing their personal information. The disruptive and innovative nature of blockchain technology resulted in the evolution of many decentralised applications such as cryptocurrencies and smart contracts. Bitcoin, a decentralised cryptocurrency based on blockchain technology was introduced in 2009 [5] and as of now, Bitcoin is the largest cryptocurrency with a market capital of approximately more than 100 billion USD[2]. Simply put, blockchain technology is built on three main concepts: a distributed database, a trust protocol and cryptography. In the following subsections we will explain them briefly.

Distributed database: Built on the concept of peer-to-peer networks and distributed storage [16], blockchain technology can be considered as a distributed data store with state machine replication using peer-to-peer protocol, where the transactions are the atomic changes to the data store which are grouped into blocks [12].

The Trust Protocol: In order to avoid having a central authority for enabling the trust in the system, there needs to be some mechanism that establishes trust between the involved parties, which is achievable by distributed consensus of the involving parties. In blockchain trsust is ensured through a distributed consensus protocol. Although the protocol can vary slightly from system to system, the idea of achieving trust with the consensus of involving parties remains the same. The two most widespread concepts of this protocol are proof-of-work and proof-of-stake which follow a Byzantine fault tolerance scheme [15].

Proof-of-work (PoW) refers to the idea that a service requester is required to solve a cryptographic puzzle (*computational work*) to participate in a network and it was initially proposed in hashcash [17] as a counter measure for denial of service attack using CPU cost-functions. In blockchain and especially in Bitcoin [5], it is used as a verification techniques for finding the suitable appropriate header for new blocks of data and to append them to the chain of blocks. To add a block, a node has to solve a cost-function (find the right *nonce*), that results in a pre-defined hash format with certain restrictions. At the same time, blocks can only be added to the longest chain (with the most proof-of work invested), to avoid 'dishonest' attempts of altering the ledger.

Proof-of-Stake (PoS) is another method for verifying and adding blocks to the blockchain, where the node that creates the next block is chosen [18]. Therefore, a node adds and verifies blocks according to how much stake they have in the system. Thereby, ownership will lead to actors behaving honestly, otherwise they would lose their stake, if they behave dishonestly. Even though there are other anchoring schemes similar to the above, we skip their description due to space limitations.

3.2. Cryptographic Primitives

Hash Functions: Hashing is used to ensure integrity of data and a hash function is an input independent average linear time algorithm that takes set of variables or data

[2]https://charts.bitcoin.com/

and transforms it into a fixed size hash digest [19]. A successful hash function has the following characteristics: *deterministic* - the same input always creates the same output, *efficient* - output is computed in a timely manner, *distributed* - evenly spread across the output range, meaning that similar data should not correlate to similar hashes, *preimage-resistance* - it needs to be infeasible to find the input x, based on the hash value ($h(x)$) and *collision resistance* - no two different inputs x and y, create the same hash $h(x) = h(y) \implies x \equiv y$. Furthermore, hash functions are used for organising and linking data together in blockchains. Another key concept of hash functions in the blockchain is that of organising and linking data together. This is done through the hashing of various elements in the block header containing hash of previous block, merkle root of transactions, time, and nonce. The concept of Merkle Tree [20] is that each transaction is hashed, then the resulting hash of each transaction is hashed to build a tree structure until top node known as the merkle root is obtained. This type of organising of data allows secure and efficient verification of contents of a block and summarise all the transactions in a block [21].

Digital Signatures: One of the main goals of blockchain technology is to be able to verify authenticity and non-repudiation of data/transactions. Digital signature is a cryptographic scheme that guarantees two properties: *authenticity*, that the data/message created or owned by the known sender and the *non-repudiation* property guarantees that the data is not altered, using a pair of keys with an asymmetric cryptographic algorithm like RSA [22]. Over the years, more secure versions of digital signatures have been developed. For instance, Bitcoin, uses the Elliptic Curve Digital Signature Algorithm (ECDSA) for key generation [23].

The orchestration of the above-described technologies lead to the following characteristics (Tab. 1).

Immutability	Data written to database cannot be changed or deleted without consensus leading to data integrity
Decentralization	No single point of failure/control achieved by decentralized architecture and a distributed database
Transparency	All data sent through the blockchain is visible to all network participants
Pseudonymity	The identity of data senders and receivers is unknown
Chronology	Every transaction is time-stamped and can be traced back

Table 1. Characteristics of the Blockchain

Using blockchain as a tamper-proof ledger would record the transfer and prove ownership of assets beyond any doubt. This enables smart contracts, an idea conceptualized already 20 years ago [24]: the creation of computer programs that can securely enforce previously closed contracts. Concluding, the idea of smart contracts is to take contractual clauses, translate them into code and thereby making them self-enforceable. Hence, intermediaries who are responsible for enforcing the contract are not needed, but instead a trusted computer program is relied upon. Complex contractual and payment agreements can be included in standardised contracts and then be monitored and executed at low transactional costs, as they are managed digitally and immutably [25].

4. Conceptual Design

Methodology: We use design science as the methodology for building the conceptual design of the proposed system. The proposed conceptual design in this research work serves as an artefact and we want use the conceptual design as a basis for building a prototype implementation later. Furthermore, we want to do several iterations of the design artefact and prototype, validating and evaluating them according to design science guidelines for meeting the specifications of the proposed system. We also want to integrate feedback on the conceptual design and subsequent prototype from different stakeholders of the system systematically according to the design science guidelines.

In this research, our motivation is to develop a concept that maximizes the transparency as well as the control over personal data for users. MyData has proposed a human-centric approach that empowers the user by placing him in the center of his *data ecosystem*. The main focus here is not *owning* the data (i.e. storing the data on the user's own server), but to control the data flow from data to service provider by controlling the associated consents from the user to the respective service. While the approach of MyData, requires a significant shift in the ecosystem and that service providers agree on this way of handling data, we sought to develop an approach that can enable a fair balance in the ecosystem without support from the service provider, but only through technology and legislative means. It needs to be mentioned, that users also have little transparency over the value of their data, which is currently used as a type of *digital currency* to pay for the use of a *free* service. Service providers most often use personal data to tailor and improve their services as well as sell their user's data to third party providers for money. Our system design does not aim at avoiding the collection and usage of data for service improvement, research, etc., but it aims at enabling the user to gain transparent insights and receive a monetary return for offering his personal data directly to other service providers. We deem that blockchain as a vehicle of decentralisation, shifting the power from cen-

tralised service providers to the end users. Through the characteristics mentioned in section 3, blockchain has the potential to put control in the personal data ecosystem in the user's hands with an increased trust that is distributed among all parties. Furthermore, the immutable and chronologic storage of consents and personal data transactions increases trust from the service providers' perspective. Lastly, capturing the conditions of data and monetary exchange between service providers and user through smart contracts omits the need for a third party while ensuring a reliable way of storing and enforcing the agreements. Smart contracts increase the level of trust from both sides at reduced costs, as the conditions are stored in and executed through immutable code. However, it is important to mention, that blockchain is not a suitable mean to store personal data on it as it is replicated across many nodes, which will leads to lot of redundancy and at the same time the immutability aspect of blockchain conflicts with the GDPR right to be forgotten in case of personal data as the data once stored on blockchain can not be deleted. Therefore, in order to address this challenge, we propose to use off-chain repository to store the personal data of users and let the blockchain store a hash data pointer to the storage location of personal data on off-chain repository. In this way, if some wants to use the GDPR right to be forgotten, then the personal data on off-chain repository is deleted in order to comply with GDPR and the immutable hash data pointer stored on blockchain will be become null and void and thereby becomes GDPR compliant. Next, we will introduce a human-centric approach and highlight the advantages of blockchain technology.

4.1. System Design Guidelines

The overarching goal of the system is to provide a holistic, personal data management tool to the user, meaning that the user of the system can expect full transparency and control over his personal data. We believe that this can be achieved by creating a system that embodies the following system design guidelines:

1. User-centric: empowering the user
2. Transparency: user knows at any time how and by whom his data is utilised
3. New rights: GDPR-compliant give and revoke consent for data processing, deletion and portability.
4. Data economy: provide a financial value to the data and facilitate the trading of it
5. Validated data: a repository with validated data that is of high value to service providers.
6. Security: user data stored in an encrypted form with the secure storage of encryption keys

As mentioned above, the MyData principles are incorporated into the design guidelines. In the proposed conceptual design, we ignore scenarios of data transfer from one provider to another. We do not want to facilitate a data transfer between service providers as it is not in line with the full control of the user data by the user. The proposed system would be built on a private or permissioned blockchain with public visibility, which means that anyone can view the transactions / blocks on the blockchains and verify certain permitted validity checks (as a public user), but the permissions of various stakeholders for example who can make a transaction or who can be data validator etc. are regulated by the governing body of the proposed system which typically includes major stakeholders.

4.2. System Overview

As illustrated in figure 1, the BPDIMS incorporates several roles and components as further explained below.
System Roles: The following are the key stakeholders in the proposed system.

1. User: end users utilizing the system
2. Service provider: company providing a service to user, either paid or free.
3. Data purchaser: an entity (company or person) purchasing the user data for a specific stated purpose
4. Data validators: entities who validate the user data to make sure that it contains what the user claims to be.

System Components: As illustrated in figure 1, the system incorporates several components, namely three blockchain layers: a smart contract blockchain, an access blockchain and an identity blockchain, the off-chain data repository, and the user interface.
Data Types: The system distinguishes between two different types of data collection. On the one hand, we have a static identification data type, which is already in *control* of the user. This can be the user's name, age, address, and personal information etc. The user can verify his data by an institution (e.g. municipality) to use the identification functionality (sec. 5.4) or to increase the value of his data. On the other hand, we have a dynamic data type, which is generated while using a service and is in the control of the service provider. This includes shopping history, performance data in a fitness app, and social media data, among others. The distinction is important, because each data type is captured differently.
Blockchain Layers: The proposed system contain blockchain layers (fig. 1) as further described below.
1) Smart contract layer (Smart Contract Blockchain): The first layer is a Smart Contract Blockchain, which

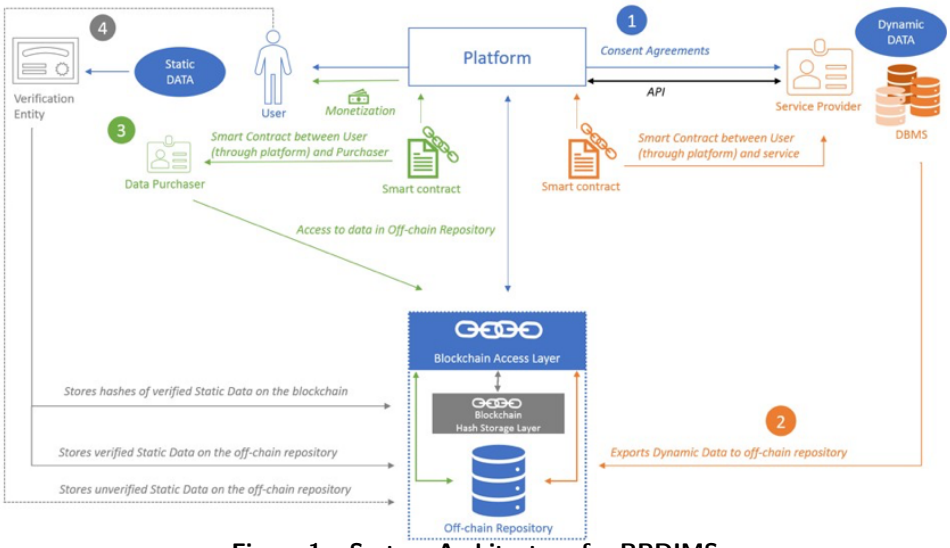

Figure 1. System Architecture for BPDIMS

is used to store conditions for data exchanges between: (1) user and service providers, which are agreements on data export on a regular basis and (2) user and data purchaser, which are agreements to access and pay for datasets of the user.

2) Access layer (Access Blockchain): The second layer will be implemented as a tool to ensure privacy, while taking advantage of the immutability and integrity characteristics of blockchain technology. The idea is based on [26] who connect an offline storage through an access layer based on the blockchain. This framework enables users to control and own their personal data, while service providers are *guests* with *delegated permissions*. Only the user can change this set of permissions and thereby access to the connected data. The basis for the access management is a smart contract-fuelled blockchain, so that the user can set automatically-enforced time limits for the access of the data. After the time limit, the consent is automatically revoked.

3) Hash storage layer (Identity Blockchain) The third layer is used for storing hashes of data. These hashes are created, when personal data of the user is verified by certain trusted authorities like governmental organisations who could verify the user's personal details. Part of this verification process is creating a hash of the verified data that is immutably stored on this Identity Blockchain of the system. Whenever a service provider gets access to this layer, he can verify personal data sent by the user and thereby verifying the *digital identity* of that user.

Off-chain repository: The user data will be stored in the external online data repositories which could be cloud storage database systems or any other online data storage repositories. For example, the underlying personal storage system could be constructed as a distributed hashtable as developed by [27], which is connected to the data pointers of the access layer. This way, data can be fragmented and is less attractive for hacking, while accessing and finding the data in the database is highly efficient. These data repositories are not part of the blockchain and therefore we can name them as off-chain repositories. Storing the personal data in the off-chain repository allows the data to be deleted from the system, should users revoke their consent, which is in line with the GDPR. Moreover, all the user data in these off-chain repositories will be stored in an encrypted form using symmetric encryption keys that are owned by the respective user who owns the data. We also propose to use threshold encryption [28] scheme to split the key and distribute them to the third party key keepers using the established key exchange algorithms such as either with Diffie-Hellman key exchange algorithm [29] or even using the public key infrastructure [20] to to securely exchange encryption keys store in a safe and distributed manner.

User Interface (UI): The user interface has two main purposes: firstly, to give an overview over all personal data of the user and secondly, to be able to manage all the data and system functionalities. The system displays all personal data that is stored at any service provider and the respective given consents (e.g. billing, targeted advertising or newsletter mailing), the data selling history and all data that is currently stored on the off-chain repository of the user. The user can manage all data in the same system, which is based on giving and revoking consents to use the data and to access the data. The data is accessed either when it was purchased by a company or when the user identifies himself through the system.

5. Functionality and Use Case Scenarios

5.1. Adding User Data

One of the most essential prerequisite workflows of the system is adding data to the off-chain repository, as this is how the user gets ownership over his data and how consents are connected to this data. Dynamic data refers to the data type that is created while using services and inevitably - also - stored on the databases of the service providers. It is important that the user gains ownership and control over this data, while forwarding or replicating this data to third parties is permitted through the consent management component. To get hold of the data, the system receives the data from the service providers. It is required by the GDPR to provide a data export of all the user's data stored at the provider, which the system will takes advantage of. As shown in fig. 1, by sending a request for a full data export, the user gets control over all data stored at the service provider. Moreover, the data received from the providers will be stored in the off-chain data repositories in the encrypted format using symmetric-key algorithms like Rijndael AES [30] and we propose that the symmetric keys should preserved using threshold cryptographic methods such as [28].
Adding dynamic data from service providers:

1. The smart contract holding the consent of the user automatically triggers a request to the service provider requesting the user's private data.
2. The service provider transfers the data in machine-readable format to the system.
3. The system transforms the data into the format needed for the repository and adds it to it.
4. The smart contract requests data exports at predefined time intervals.

Adding unverified identification data to the system:

1. The user enters information into the system, such as e-mail address.
2. The user classifies the privacy rules of this information, e.g. from open through controlled to sensitive.
3. The system stores this information as an unverified data entry with the respective privacy setting.

Adding verified identification data to the system:

1. The user finds the institution that is responsible for issuing the identity data.
2. The user enters the information into the system.
3. The institution gets access to this information through the access layer, as a consent transaction is created that stores the shared identity.
4. Through the access layer the institution gets access to the *identity blockchain*, and then stores a hash of that information immutably.

5.2. Consent Management

The GDPR states that it shall be as easy to revoke consent as to give consent for the user regarding processing and storing of private data. Consent appears in our system in three ways: 1) Consent for processing personal data in return for services 2) Consent for storing personal data, and 3) Consent for selling/access to personal data. All user's consents are stored on the Access Blockchain of the system. The second and third type of consent regarding monetization and storage, however, also has a link to the Smart Contract Blockchain. While the consent is stored on the Access Blockchain, it is used to access data in the creation of smart contracts. The creation of the smart contract entails a different type of consent, that is binding regardless of the initial consent, due to the new contractual agreement between the parties. It must be noted that the consents regarding monetisation and storage of personal data is given directly in the system UI. This process involves a request sent from the system on the user's behalf with a valid signature to the service provider in question. Obtaining information regarding a consent can however be a more cumbersome process, depending on the technical implementation method. Especially if the service provider in question is unwilling to partake in the ecosystem our system creates. With this in mind, we identified email and a API integration as the two most feasible options for communication of consent between the system and service providers, where one option does not rule out the other. Meaning the system could feasibly operate with both, depending on the service provider's willingness to participate.
Give consent to service provider:

1. User agrees to terms and conditions of service provider (gives consent)
2. System sends request to service provider for all the user's personal data.
3. Service Provider sends data to the system in a commonly-used and machine-readable format.
4. System is updated with the information from the service provider and displayed in the user's UI.
5. If purpose for data processing or handling changes, service provider must ask for new consent, which is updated in the system in the same manner.

Revoke consent from service provider:

1. User removes consent through UI.

2. System sends request to service provider to stop processing and delete personal data regarding the user.
3. System receives confirmation of deletion from service provider.
4. System deletes the information from interface and/or repository, based on user's demands.

5.3. Data Monetization

Data has become a trade-able asset, which we most often trade for the usage of free services, often without explicitly knowing for what purposes our data is used or to what third party provider the data is sold. By facilitating the trade of datasets between user and data purchaser for a monetary compensation, the marketplace functionality of our system attempts to enable the user and service providers to participate in the data economy in a more direct and transparent way. Receiving a monetary reward for sharing personal data is not an unknown concept. There are various web services out there, that offer to sell different parts of user data to third parties for which users receive a recurring or one-time payment. For instance, users can share their mobile behavioural data, or their browsing activity on particular websites. In fact, the data brokerage market is estimated to be 156$ billion in 2016. Seen from a data purchaser point of view, the aggregation of a large pool of diverse datasets bears the opportunity to access data profiles that would have been normally out of reach. Furthermore, datasets would be verified by data validators, for instance participating ecosystem service providers guaranteeing for the quality of the data. In return for validation of data they would be rewarded with a portion of the money from the sale of said data. AS the proposed system is designed as a permissioned blockchain, the key stakeholders of the blockchain (who act as the governing body) will decide who can join the blockchain as a data validator. The authenticity of the data validators can be validated/monitored by using the feedback from the data purchasers. In case, if there are any discrepancies noticed by the data purchaser in the data validated by the data validator, then that particular data validator may be warned or even block-listed in case if the validation failures are repeated. The consents given or revoked by the user are stored on the blockchain and data purchasers can browse through the marketplace to find relevant datasets. Finally, the data purchaser and user enter a smart contract, that enforces compensation and the access to the dataset.
Listing a dataset for sale:

1. User gives consent to what data, if any, can be sold in the user interface.
2. System lists this data as for sale in the marketplace.
3. Data will be validated by the data validator who will serve as auditors validating the claims of user data in terms of what the user is claiming.
4. After the validation checks, the data validators puts certification for the data, which will provide confidence to the data purchasers that they are buying the user data which is validated by the data validators.
5. Consent is put into smart contract between user and data purchaser, pointing to the data in question.

Data purchaser buys data:

1. The data purchaser can browse through the marketplace and select datasets he wants to purchase. He can retrieve all necessary information, such as price, data certification details from the overview page.
2. When the data purchaser wants to purchase a particular dataset, it is checked whether the data purchaser has sufficient means to purchase the data in question.
3. If this is the case a consent transaction is created on the access blockchain and together with the data pointer, compensation information and expiry date stored in a smart contract.
4. The compensation is transferred to the user.
5. The data purchaser gets access to the repository and can download the data files.

5.4. Identity Management

As part of a holistic data management approach, the platform also supports an identity management functionality. Both service providers as well as users can highly benefit from a blockchain-based solution. It can still take several days to onboard a customer for a new service that requires verified data (e.g. requesting a loan at a bank), while the process costs large sums of money for the service providers.
User digitally identifies himself to a service provider:

1. a user wants to access a service, which the provider requests information for
2. the user authenticates himself to the personal data storage through his private key
3. A consent transaction is created on the blockchain with a shared identity of the service provider and the user. This gives the provider access to that data point as well as the *identity blockchain*
4. the service provider can read run the information through the stored hash and verify the information
5. the user has successfully identified himself and the provider has only the information needed

6. Discussion

We will discuss the benefits of using blockchain to handle a user-centric personal data management.

Improvements Using Blockchain: Due to the concept of immutability as part of blockchain, data stored on a blockchain cannot be changed without a consensus amongst the participating nodes, which leads to a very high data integrity. The proposed blockchain system stores the hashed data pointers pointing to the user data on off-chain repositories, this provides guarantees that the user data has not been altered by the user or anyone else since the time it has been marked for sale. This kind of in-built trust provided by blockchain will be quite beneficial to the data purchasers as they can buy the data without worries about data provenance. Moreover the role of data validators and their certifications will enhance the trust in the user data that put up for sale. Blockchain provides complete transparency and verifiable proofs about various transactions related to the user data and identity management, which will enhance trust and confidence in the system to all the stakeholders such as users, service providers and data purchasers etc. Similarly, the anonymity feature of blockchain allows users to conceal identity and their personal information whenever necessary, e.g. in the case of negotiating with a data purchasers and at the same time, the system allows the users to reveal their identity in case if it is needed. Finally the decentralised and distributed consensus mechanisms of blockchain will provide guarantees against the system being taken by malicious actors easily. This means that unless a malicious actor controls more than 51% of the network, a false entry or change to the data will not be approved.

Smart Contracts: Smart contracts allow us to use fully-automated self-enacting electronic contracts which means the automation and legal certainty of consents and their management is significantly improved. Moreover, smart contracts operate as autonomous actors whose behaviour is completely predictable [8]. This is done while ensuring a very high integrity of the authenticity of the contracts in question, as well as transparency of the system. Introduction of smart contracts for creating and revoking consents will result in unambiguous legal contracts and it is easy for regulators and auditors to investigate the claims in case of disputes between the users and service providers/data purchasers.

Encrypted Data Storage: Through the implementation of storing the user data in the encrypted form using symmetric-key cryptography and with the encryption key of the user distributed over different key keepers using threshold cryptographic methods, the system avoids a single point of failure. A compromise of the off-chain data repository will not lead to a data leakage. This is due to the encryption of the data repository and the number of keys needed to decipher the data. As one key from key keepers is not enough to decipher the data, a malicious actor would have to compromise several key holders, which further increase the security of the system and decrease the likelihood of a data leakage.

The User's Perspective In the proposed system, a user would be able to grant and revoke access to personal data, but also monitor who has access to it and what it is being used for. This is a significant upgrade from today's situation where most of us us have little knowledge of where our private data is and what it is being used for. With access to personal data and insight to where the data is and what it is being used for, users are likely to become more aware of how they act. This means users will be able to see how they navigate online and where they leave data traces on a more detail level, which potentially will lead to higher awareness of users and deeper insights into their online behaviour. The potential for monetisation of the user's private data is another key change and benefit. However, a shift in transparency and access could also lead to several benefits for companies, which will be discussed later, but the broader access to data could result in a fairer market with more competitors and cross sector usage of data. In general, the proposed approach significantly empowers the user with transparency and control as the main features, with spillover effects to the services available and the reward for usage.

Business Perspective: For service providers our system can help facilitate compliance with the GDPR dealing with both consent and transparency in data handling. The incentive from a business perspective goes both through the monetization of selling data as a data validator and to buying data. The possibility of buying data of potential customers and users from competitors within the industry and also across industries provides a significant opportunity for companies to expand and improve services by getting in the intelligent insights into their customers. This incentive is particularly large for smaller companies and startups that don't have access to data. Furthermore is the possibility to buy data to discover new, potential market opportunities is another key advantage that incentivises companies to engage with the system.

7. Conclusion and Future Work

In this research work, we proposed a conceptual design and high-level architecture for a personal data and identity management system with key focus on providing transparency and control over the usage of the personal

data of users. Building on the foundations of blockchain and smart contract technologies with a human-centric focus, our proposed system provides high-level trust and security and shifts the control over personal data to the end users in a transparent manner and facilitates the functionality of creating and revoking consents for accessing and selling their data to the companies that want to buy user data.

In future, the secure data transfer from service providers to off-chain data repositories and the service provider integration will be explored. We want to work more in the direction of preparing a detailed specification for the proposed system. We would like to use a formal methods approach to derive a detailed specification by describing various interactions between different stakeholders of the system in an unambiguous manner.

References

[1] G. D. P. Regulation, "Regulation (eu) 2016/679 - directive 95/46," *Official Journal of the European Union (OJ)*, vol. 59, pp. 1–88, 2016.

[2] C. Tankard, "What the gdpr means for businesses," *Network Security*, vol. 2016, no. 6, pp. 5–8, 2016.

[3] O. K. Foundation and the Open Rights Group, "Personal data and privacy working group," 2014.

[4] A. Poikola, K. Kuikkaniemi, and H. Honko, "Mydata a nordic model for human-centered personal data management and processing," *Finnish Ministry of Transport and Communications*, 2015.

[5] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008.

[6] R. Böhme, N. Christin, B. Edelman, and T. Moore, "Bitcoin: Economics, technology, and governance," *The Journal of Economic Perspectives*, vol. 29, no. 2, pp. 213–238, 2015.

[7] S. T. Ali, D. Clarke, and P. McCorry, "Bitcoin: Perils of an unregulated global p2p currency," in *Cambridge International Workshop on Security Protocols*, Springer, 2015.

[8] K. Christidis and M. Devetsikiotis, "Blockchains and smart contracts for the internet of things," *IEEE Access*, vol. 4, pp. 2292–2303, 2016.

[9] G. Zyskind, O. Nathan, *et al.*, "Decentralizing privacy: Using blockchain to protect personal data," in *Security and Privacy Workshops (SPW), 2015 IEEE*, pp. 180–184, IEEE, 2015.

[10] N. Kaaniche and M. Laurent, "A blockchain-based data usage auditing architecture with enhanced privacy and availability," in *Network Computing and Applications (NCA), 2017 IEEE 16th International Symposium on*, pp. 1–5, IEEE, 2017.

[11] A. Yasin and L. Liu, "An online identity and smart contract management system," in *Computer Software and Applications Conference (COMPSAC), 2016 IEEE 40th Annual*, vol. 2, pp. 192–198, IEEE, 2016.

[12] P. Mamoshina, L. Ojomoko, Y. Yanovich, A. Ostrovski, A. Botezatu, P. Prikhodko, E. Izumchenko, A. Aliper, K. Romantsov, A. Zhebrak, *et al.*, "Converging blockchain and next-generation artificial intelligence technologies to decentralize and accelerate biomedical research and healthcare," *Oncotarget*, vol. 9, no. 5, p. 5665, 2018.

[13] A. Azaria, A. Ekblaw, T. Vieira, and A. Lippman, "Medrec: Using blockchain for medical data access and permission management," in *Open and Big Data (OBD), International Conference on*, pp. 25–30, IEEE, 2016.

[14] X. Yue, H. Wang, D. Jin, M. Li, and W. Jiang, "Healthcare data gateways: found healthcare intelligence on blockchain with novel privacy risk control," *Journal of medical systems*, vol. 40, no. 10, p. 218, 2016.

[15] L. Lamport, R. Shostak, and M. Pease, "The byzantine generals problem," *ACM Transactions on Programming Languages and Systems (TOPLAS)*, vol. 4, no. 3, pp. 382–401, 1982.

[16] L. Xu, *Highly available distributed storage systems*. PhD thesis, California Institute of Technology, 1999.

[17] A. Back, "Hashcash-a denial of service countermeasure." http://www.hashcash.org/papers/hashcash.pdf, 2002.

[18] BitFury Group, "Proof of stake versus proof of work." http://bitfury.com/content/5-white-papers-research/pos-vs-pow-1.0.2.pdf, 2015.

[19] J. Carter and M. N. Wegman, "Universal classes of hash functions," *Journal of Computer and System Sciences*, vol. 18, no. 2, pp. 143 – 154, 1979.

[20] R. C. Merkle, "Protocols for public key cryptosystems," in *Security and Privacy, 1980 IEEE Symposium on*, pp. 122–122, IEEE, 1980.

[21] A. M. Antonopoulos, *Mastering Bitcoin: unlocking digital cryptocurrencies*. " O'Reilly Media, Inc.", 2014.

[22] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," *Communications of the ACM*, vol. 21, no. 2, pp. 120–126, 1978.

[23] D. Johnson, A. Menezes, and S. Vanstone, "The elliptic curve digital signature algorithm (ecdsa)," *International journal of information security*, vol. 1, no. 1, pp. 36–63, 2001.

[24] N. Szabo, "Formalizing and securing relationships on public networks," *First Monday*, vol. 2, Sep 1997.

[25] M. Swan, *Blockchain: Blueprint for a new economy*. " O'Reilly Media, Inc.", 2015.

[26] G. Zyskind, O. Nathan, and A. Pentland, "Enigma: Decentralized computation platform with guaranteed privacy," *arXiv preprint arXiv:1506.03471*, 2015.

[27] P. Maymounkov and D. Mazieres, "Kademlia: A peer-to-peer information system based on the xor metric," in *International Workshop on Peer-to-Peer Systems*, pp. 53–65, Springer, 2002.

[28] A. Shamir, "How to share a secret," *Communications of the ACM*, vol. 22, no. 11, pp. 612–613, 1979.

[29] G. Al-Aali, B. Boneau, and K. Landers, "Diffie-hellman key exchange," *Proceedings of CSE 331, Data Structures Fall 2000*, vol. 67, 2000.

[30] J. Daemen and V. Rijmen, *The design of Rijndael: AES-the advanced encryption standard*. Springer Science & Business Media, 2013.