Chapter 10
Sharing of Research Data by Blockchain

Chris Dai, Tadaaki Chigusa, and Makoto Yano

Abstract This study investigates the new method of sharing research data by a blockchain. We focus on the sharing not only among researchers but also between a group of researchers and test subjects from whom researchers collect data. To explain the use of blockchain technology, we explain the use of blockchain in securely handling research data in comparison with a simple real-world use case.

Keywords Research data · Blockchain · Data security

1 Introduction

Date sharing is one important issue faced by researchers building data covering both human health and behaviour. By data sharing, we do not simply mean the sharing of data among researchers, which is obviously highly important for us researchers. However, more important is the sharing of data between researchers and survey subjects. The latter issue would become sensitive when we deal with serious illness from which a person wants to hide the fact that he suffers or a person who can spread to others. In Japan, there remains serious social stigma attached to certain diseases. At the beginning of the COVID-19 outbreak, in Japan, those contracted Sars-Cov-2 were ostracized in certain traditional regions. Even if such a stigma is not attached to contraction of a disease, losing private data and misinforming subjects of wrong test results could be intolerable for test subjects. In these circumstances, there can be a strong hesitancy on the side of a data builder to inform test subjects a result of the test.

C. Dai
Recika Co., Ltd, Tokyo, Japan

T. Chigusa
Two Plus Associates Co. Ltd., Tokyo, Japan

M. Yano (✉)
RIETI, Tokyo, Japan
e-mail: yano-makoto@rieti.go.jp

M. Yano et al. (eds.), *Socio-Life Science and the COVID-19 Outbreak*, Economics, Law, and Institutions in Asia Pacific, https://doi.org/10.1007/978-981-16-5727-6_10

It is our view that blockchain technology may drastically ease the possibility of malicious attacks, mistakes, and errors that may occur while researchers share data among researchers and with test subjects. Blockchain is a technology on decentralized computer networks that allows end producers of data to share their digital data among themselves in a secure manner (see Omote and Yano 2020, for an explanation on the technology). It is therefore natural that the standard blockchain technology lets researchers securely share their research data among themselves. Data sharing between researchers and test subjects is, however, different because of its nature and of possible test subjects' unfamiliarity with blockchain. To deal with the unique feature of researchers/test-subjects data sharing, we need a new type of blockchain application.

In this chapter, we introduce a design of such an application, with which data can be shared securely between researchers and test subjects as well as among researchers. In doing so, we explain how we should handle unique features in this type of data sharing through blockchain.

Blockchain is a new technology, with which many readers may be unfamiliar. Strong resistance exists among ordinary people who do not understand exactly what blockchain does, because it is a technology that supports crypto currency, and because, unfortunately, crypto currency is currently used as a device for speculation and shady transactions; if speculation and shady transactions subsided, crypto currency might replace the existing hard currencies (Yano 2020). For this reason, we start this study with explaining basic features of blockchain technology and present a use case.

2 Blockchain and Decentralized Data Ownership

Blockchain is a technology, or a computer algorithm, that makes it possible to assign a unique owner to each piece of data in a data file on the internet. For this purpose, many copies of a digital file are maintained independently at separate network computers. Supported by modern cryptography, blockchain algorithms ensure those copies are identical and secure in such a way that external attacks on the file are almost, if not totally, impossible (see Omote and Yano 2020, for more explanation). The mere fact that many internet users regard crypto currencies as a way to hold their assets attests that blockchain can establish the secure ownership of data, or records of past transactions.

2.1 Data Security and Data Privacy

"Files containing personal information on some 9500 people infected with the novel coronavirus in this western Japan prefecture were accessible to a third party online, the Fukuoka Prefectural Government revealed on Jan. 6" according to

Mainichi Japan, a Japanese national daily newspaper (Mainichi Japan 2021). The files contained names, addresses, and symptoms, and other information on COVID-19 infected residents in the prefecture were accessible to unintended third parties. Needless to say, this kind of information if leaked severely damages personal privacy and even causes possible harm to the people who submitted their personal information to the government.

This kind of information leakage is becoming more frequent as institutions such as government and companies push forward to digitize all data and allow online data access with related parties. At first glance, the problem is caused by a random operational error performed by careless staff. Some may even claim that the risk of such mistake is inherent in sharing data online through servers. But the real issue is that data is stored in one place with a single point of failure, in this case the Fukuoka prefectural government.[1] If someone within the government makes a mistake or the system has a small bug then it becomes a honey pot for outside attacks by hackers. How to solve this issue? Can we put multiple check processes in between to prevent such human error? As you remember in your high school physics class, the second law of thermodynamics states that the level of disorder in the universe is steadily increasing and hence systems most likely to move from order to disorder. You can think of a file management system as a centralized and highly ordered system. As time passes by, more users using the file management system will cause disorder to creep in and incidents will happen. Therefore, the real issue is how we can make a system that can counter disorder in the long run. The answer lies in a decentralized system where individuals, not a central data management entity, manages and controls their own data.

You may feel the centralized data storage used by the Fukuoka prefecture government has nothing to do with you. Consider the vast amount of personal data you share with internet platforms like Google, Amazon, Facebook, Apple, Microsoft, etc. Unlike the Fukuoka prefectural government that collected and shared data of COVID-19 infected patients for the social good, those internet platforms collect user data for just a single purpose of making a profit, as they should. And those data are also stored in centrally managed databases.

The issue of internet platforms storing your personal data goes beyond whether they are securely storing your data or not. Most people are not even aware what kind of data they have collected, how are those data being used, and for what purpose? You can get a glimpse of how important personal data can be for these platforms from a recent BBC report titled "Facebook pours fuel on Apple privacy row". Earlier in 2020, Apple announced it would implement a new policy for its iPhone users to ask for permission to have part of the user data shared to enable targeted, personalized advertising. As its main revenue come from such advertising paid by third party advertisers, Facebook fear it will make targeted advertising difficult following Apple's new policy because a great proportion of users will not give such permission. Facebook has put on a public offensive, including putting a full-page adverts

[1] See Yano et al. (2020) and Pu (2020) for further discussions on leaving a single point of failure unprotected in a network system.

in printed newspapers, to claim that Apple's new policy is hurting small businesses that utilize personalized advertising. It also alleged Apple's move is about forcing people to use Apple's own advertising platform.[2]

Watching two internet giants fight over whether to share users' personal data raises the question who really own the users' personal data? The official answer is, of course, the user owns his or her data. But, in reality, the data is stored in the internet platformers' database, and it is not easy to find out how that data is being used or even leaked. To counter this problem, a decentralized data ownership model should be devised. A decentralized data ownership model is inherently more secure in its architecture, as data is distributed between many individuals and has no single point of failure. It is also more resistant to hacking because, instead of breaking into one system, hackers need to break into all the individual systems to collect all the data, which makes it much less desirable as an attack target. To make individual ownership of data in the true sense, we need a new system model different from the centralized client server database model.

This is true for research data as well. Currently, a lot of research data are underutilized in fear of data tampering and data abuse. The more sensitive data a database contains, the more difficult for researchers to share. In extreme cases, researchers are required to be physically present in the site in which a database server is located. In less extreme cases, data are provided only by CD roms, which many PCs are no longer equipped to read. Although all these are necessary precautions if there is no other ways to share research data, they severely limits the sharing of valuable data. Blockchain is expected to alleviate this problem.

2.2 Centralized or Decentralized Data Ownership

We need to first define what centralized data ownership or decentralized data ownership mean. Most data are centrally managed. Even before the advent of the internet, governments were responsible to issue birth certificates, register families, record ownership of real estates and so on. Based on these government records or issued documents, you can open a bank account, purchase an airline ticket, rent an apartment. In due courses, those data are stored and managed by companies offering services to you. It is easy to see the necessity of centrally management data because, as a citizen, you need to have someone to record your place in the social machine. With the internet coming to play, this centrally managed data became more interconnected but the overall structure stayed the same. The current internet service model is mainly the client server model where one server or a cloud server will manage data for many client devices, and the client is dependent on the server to store and manage data. In such a model, a client is provided an ID and a password, so the client can access the server and exchange data with it. All the interaction data is stored together

[2] For an analysis on the peril of such data monopoly in a broader context, see Yano (2020), Pu and Yano (2020) and Yano et al. (2020).

with other data provided by the client in the server. Normally the client cannot see what operations are done to the data it stores in the server. Hence, in this model, the server that is centralized has the ownership of the data. If data needed to be shared with another client or another server, perhaps in a different system, the server that stores the data will have to do the work, not the client.

In a decentralized data ownership model, data is not being controlled by the server; rather, it is stored and managed by the client. How is this possible? There are different ways to achieve this. Data can be stored locally in the client's device rather than the server, and when data sharing is needed the owner of the data and the device will give permission to share that data, similar to Apple's proposed method. On the other hand, the data can still be stored in the server, but encrypted with a key only the owner of the data has. When others request that data, the owner of the data can re-encrypt the data with the key the requesting party provides. An individual or the owner of the data can have control of with whom to share the data in a secure way.

We must note that there is no pure centralized data ownership or pure decentralized data ownership. Any data ownership model rest somewhere in between the centralized and decentralized. However, the proliferation of client server models in the internet platform architecture has ensured that most systems manage data closer to the centralized model rather than decentralized model. The growing needs of personal data collected by internet giants also feeds and favours the centralized model as platforms can easily use data to scale its business, which brings in more users/client. Then it can collect more data and grow bigger and investors will invest more money to build an even bigger platform to store and collect data. The advent of blockchain technology has shown a different option for data management, with the decentralized data ownership model. This new model may even bring change to the business models of internet business.

2.3 Bitcoin, the First Blockchain Application

Blockchain technology was first introduced to the public by the bitcoin whitepaper.[3] Although in the paper there was no mention of the word "blockchain", only references to "blocks" and "chain", the decentralized data ownership model was very clear. Bitcoin's purpose was to build a global ledger that is not stored in a single server but maintained with the consensus of tens of thousands of servers. Without a centralized data management entity such as a bank, ownership of data can be managed individually and that ownership can be transferred between users. Bitcoin, the first successful use case of blockchain, demonstrates the promise of the decentralized data ownership model. Bitcoin became a digital currency that does not need banks to manage the storing and transfer of money. In fact, bitcoin does not need any intermediary for one to send or receive money from one another.

[3] See Omote and Yano (2020) for technical details of blockchain-bitcoin technology.

You might ask what money has to do with data. If you think about your money saved in a bank, when you transfer your money to another person in the same bank, the actual cash does not move; rather the data on your account is updated by the bank. In essence, money saved in a bank is just a piece of data managed by the bank, and is a form of centralized data ownership. In the case of the bitcoin, because there is no bank to manage your data, individuals have ownership of their data.[4] To achieve decentralized data ownership and at the same time ensure secure transfer of money, the bitcoin system has to ensure several key functions. First, the system has to know you own your account every time you send an instruction to send money. With a centralized system the bank will issue you an ID and password to make sure when you are sending online money transfer instructions, you are the account owner. However, in a decentralized system like bitcoin that utilizes blockchain, you generate your own ID and password using an encryption technology called public private key encryption. When you send money to another account, money or balance will be deducted from your account and added to another account and this process is not reversible. To make the transfer non reversible or immutable, blockchain protocol has all the servers on the network to reach consensus on the data of the transfer of money and keep the same copy in all the servers. To reverse the transaction, hackers have to hack over half of the servers in the blockchain network, which is nearly impossible considering there are more than 10,000 servers running bitcoin full node. Bitcoin has demonstrated blockchain technology can make a decentralized or individual data ownership model possible. It is important to mention that since the start of bitcoin on January 3, 2009, the system has never been successfully hacked or been taken down for maintenance. A rather incredible feat for any online system.

3 UniCask: A Use Case of Blockchain Technology

The use of blockchain technology is not limited to support crypto currencies like bitcoin. Blockchain can set the ownership of even a physical material and support a digital marketplace in which the physical material can be traded among participants. The existence of such a technology implies that research can be shared among researcher on a blockchain.

This is made possible because a blockchain can provide a secure internet platform on which any software program can run and execute predetermined arrangements. Ethereum is the first blockchain that offers such a capability.[5] Simply put, Ethereum is like an operating software (like Windows and iOS) on the internet of which the security is endowed in a decentralized manner. Just like some computer games on the internet, Ethereum lets an algorism designer write a program on Ethereum blockchain (which we call a "use case" here) and incorporate virtual money into that use case. By using this function, a blockchain engineer can write a specific use case for a particular

[4] See Yano (2020) for an explanation why decentralized data may have monetary value.

[5] For more explanation on Ethereum and its applications, see Metcalfe (2020) and Dai (2020).

group of researchers who can securely store and share with other researchers, their valuable research data on the use case and even trade their research data for a price among themselves.

In order for researchers to share research data, a blockchain must have several crucial features. Before studying those features, it may be useful to examine a use case that makes it possible to digitalize a physical commodity or, more specifically, whiskey casks and put its ownership on a blockchain marketplace. From the economic viewpoint, a whiskey cask is a very simple product. Because of this simplicity, a comparison of its economic features with research data lets us highlight what might be necessary to design a blockchain to sharing research data, which we will discuss at the end of this section.

3.1 Proving Ownership on Physical Good Over Internet

Many of the use cases for blockchain are indeed arising from financial industry where data itself is money and hence easier for a decentralized data ownership to show its value. Crypto currency, decentralized finance are the thriving examples. On the other hand, decentralized data ownership model is not limited to finance and many industry sectors are utilizing blockchain to build new business models to utilize individual data ownership. One such interesting use case is UniCask. UniCask links physical casks of whiskies to data on the blockchain to register the ownership of the casks. This model is not only applicable to whiskies but to most physical assets if applied correctly.

Most whiskies we know are packaged in glass bottle form. They are sold to bars, supermarkets, and liquor stores etc. However, what most people don't know is that for whisky collectors, bottled whiskies are not good enough because once the whisky is bottled, it stops aging, a term used to describe whisky stored in a wooden cask and in contact with the wood. Aging is important for whisky because the longer it ages, the taste becomes smoother and it is also a good indicator of how rare and expensive the whisky is. However, whisky in the cask form is not sold in stores. It is not easy to buy whisky in the cask form unless you have good connections in the whisky industry and are in the inner circle of the trade. To buy a cask of whisky usually means storing the aging whisky in the distiller's warehouse for years until the time is ready for you to drink it or sell it to someone.

While the cask is in the warehouse of the distiller, only you and the distiller know about your ownership of that particular whisky cask. Your proof of ownership of the cask is stored and managed by the distiller. If you need to prove to someone you own the whisky cask, you have to have to take that person together with you to go to the distiller to show your cask and get the distiller to confirm your ownership. Of course, the distiller can print a proof or ownership for you or you can even make a photocopy of your contract with the distiller. However, it is difficult to prove the authenticity of such documents and even if the other party can trust that any document you produce

can only prove your ownership up and until the date of the document. It is difficult to prove your "current" ownership if such data is managed centrally.

One way to solve this issue is to use the blockchain technology and create a decentralized global registration of casks. UniCask did just that, and created a standard for distillers to issue tokens or ownership certificates for its casks to the owner. These tokens are issued on Ethereum blockchain. Unlike the bitcoin tokens which we call fungible tokens where every token is virtually the same and inter changeable, the UniCask ownership certificate tokens are non-fungible tokens that each token carries a unique cask id to avoid duplicate tokens. The ownership certificate token is stored in the blockchain account created by the owner. If you are the owner and want to sell your cask, then before you negotiating with your potential buyer, the buy will ask you for the proof of ownership of the cask. First you can send the link to show this cask ownership certificate is in the account. Then using your account private key to make a digital signature, you can digitally show that you have control of the account and hence own the cask. The high security level of decentralized data ownership mode also ensures your transaction much more resistant to scamming and hacking. UniCask allows whisky collectors easily and more securely trade casks over the internet. It also takes away the hassles of the distiller to centrally record and manage cask ownership.

3.2 Blockchain for Researchers to Share Their Data

A whiskey cask is a very simple product from the economic viewpoint. Until it is opened, its value for consumers does not materialize.

Once it is opened, it will become a totally different object, i.e., a large empty container and a large quantity of whiskey usually in many bottles. So long as a cask is not opened, its physical quality/quantity will not change, no matter how many times a cask is traded and changes its owner. The value of a cask will be anchored by the value of whiskey at the time at which the cask is opened. Thus, the digital ownership of a cask cannot become a target for speculation, unlike a blockchain currency. Digitalization may be a perfect way to own a cask because it permits a change of ownership without physically moving it; moving casks physically might harm the stored whiskey. The transaction of digital ownership for cask makes it possible for potential owners of casks to share storing costs. The transaction of a cask does not reduce the physical quantity/quality of whiskey.

These features of whiskey casks are shared by research data. Once data becomes publically known, it loses its proprietary value. So long as data is kept secret, its physical quality/quantity will not change. The digital ownership of data cannot become a target for speculation. Digitalization may be a perfect way to own research data because it permits a change of ownership without physically moving it. Physically moving data from one server to another involves a large risk by an accident that may occur in transportation. The transaction of digital ownership for research data makes

it possible for potential owners of data to share storing costs. The transaction of data does not reduce the physical quantity/quality of research data.

An important difference between research data and whiskey casks is that sharing of data involves handing over data from the hands of an original owner to a user. For these reasons, a blockchain for research data sharing needs to be capable of delivering data securely to the data user. This implies that a blockchain has to provide a virtual environment on the side of users to securely hold data. If data would spill out of users' hands, the blockchain would lose its value.

For the owner of a whiskey cask, it does not matter who buys his cask. In some cases of data transactions, it may be desirable to limit users of research data to a certain group. This is particularly so if data involves national security so that the authority desires to keep the data within its country. It is also possible that for certain industrial data, a certain group of companies desire to share data only within themselves.

A related issue is that once researchers acquire data, they will use it for their research. It is possible that data users wish to keep it secret that they plan to use particular data, which might reveal their research ideas to competitors; researchers rarely desire to reveal their research agenda before research is completed.

Whiskey casks have large asset values. Thus, data security should be of the fore most importance for potential participants of a blockchain marketplace for casks. In contrast, many types of data do not have such an asset value. Instead, researchers may simply seek for convenience of secure data sharing, which a blockchain makes possible. Sometimes, research data may lose immediate research value after a certain period, but continues to be potentially useful for future research, which is unknown. In many cases, the production of research data is funded by a government, which may desire to make data output open for public purposes; it is often the case that government funded research data will be made open to the public after a certain period.

In designing a blockchain for research data sharing, it is important to take these points into account. They are however very partial, and a blockchain designer should design a blockchain that is tailored for specific necessities and requirements for researchers.

4 Individual Taking Control of Personal Data: Research Data Access (RDA)

As is explained in the previous section, physical products like whisky can be owned and traded as data on blockchain, so long as a proper contractual arrangement is made. If so, there is no question that research data can be owned and traded on blockchain. Additional difficulties will arise, however, if researchers are to share their research data with test subjects. If data involves a serious disease like COVID-19, as is discussed in the introduction, extreme care is necessary in data handling. Anecdotal evidence indicate that governmental officials are highly reluctant to handle

research data involving sensitive test results, for they are afraid of possible errors associated with data handling; this is not at all a surprise if we think of the Fukuoka incidence discussed in Sect. 1.

In this chapter we are discussing the characteristics of this new system model as opposed to the current centralized model, and the implication of such change. The *Nagahama Study*, a social scientific survey on the life and views of participants to the Nagahama Prospective Genome Cohort for comprehensive human bioscience provided an excellent use case for demonstrating how individuals' personal information can be securely managed by the individual test subject. The system architecture and design for implementing the decentralized data ownership model for this project is analyzed.

4.1 Personal Data

In the previous sections, individual ownership of financial data (money) and physical asset data is discussed. Now let us analyze how you can own your personal data. The definition of personal data varies between different countries and jurisdictions. Because the EU is leading the world in protection of personal data, we will take its interpretation. According to the EU's General Data Protection Regulation[6]:

> "personal data" means any information relating to an identified or identifiable natural person ("data subject"); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

Refereeing back to Sect. 1, the data that the Fukuoka prefectural government managed and the data Apple collected from its user and managed are all categorized as personal data. As pointed out, personal data collected on individuals are mostly stored in the central server of the institution that manages the data without the individual's knowledge of how that data is being used. It is doubtful that people on the infected resident list would have given their personal information if they knew staff in the prefectural government would just email a link with access to their personal data to a third party. In the same way, Facebook is worried that if Apple asks its users to give Facebook permission to track them, users will not give such permission. Although centrally managed personal data seem to have many issues, but people did not have much of a choice but to use them. With the development of a decentralized model such as blockchain, we see new options.

[6] Regulation (EU) (2016/679) of the European Parliament and of the Council 2016, p.33.

4.2 Research Personal Data Management

With any new technology, the actual breakthrough needs to have a robust, easy to understand use case to demonstrate its superiority. Past cases such as the steam engine used for locomotive and gunpowder used for guns demonstrate this well. What is a good use case for decentralized management of personal data that can benefit both society and the individual? In the previous chapters of this book, a wide-scale collaboration of natural science and social science is proposed. Centering on the same cohort/panel participant group and connecting data between the two distinct area of science will provide great insight to understand the impact of the COVID-19 outbreak on the social, human, and health aspects of individuals and the community. As the research extends its scope, more research can be expected to build on top of the data collected for this research. The data collected is personal data, which should be kept private for privacy reason. On the other hand, we want to encourage the individual or test subject to also give more research teams permission to use his or her data for social good. To create such a sustainable system for managing research data it needs to satisfy the following requirements.

(1) Individuals or test subjects will be given the right to view their personal data and a history of what research it is being used for.
(2) Personal data of the individuals will be accumulated across research projects but research projects do not store personal data that can identify the individuals.
(3) Data is securely stored and only individuals or institutions with permission can access the data.
(4) Minimizing manual operation and cost of running the system.

With the above requirements in mind, we have designed a system utilizing blockchain technology to be the first of its kind to be used for managing personal information collected for research purpose. The design, building, and operating of this new system itself is a research to validate the implementation of decentralized personal data management in a real-world use case.

4.3 Why Blockchain?

4.3.1 Decentralized ID and Use of Private Key Public Key Encryption

When the centralized data ownership model is compared with the decentralized data ownership model, the first difference is user authentication. In a centralized data ownership model, the central server takes the responsibility of linking data with the ID. When the user inputs ID and password correctly, the server recognizes that the user is logged in and the user account is given access to read the stored data or is able to instruct the server to execute actions such as giving access to another user. When operated properly, the entire process is very smooth and nothing to worry about.

However, what we do know is that this kind of centralized system is always under threat from hackers to exploit its weakness. How is a decentralized system different in user authentication?

A decentralized system does not use the centralized system to authenticate the user; rather the user creates their own ID (public key) and password (private key) using the public key encryption algorithm. Then the user registers the public key on the blockchain. The actual personal data can be stored in a server, but encrypted using the public key of the user. As a result, even if the server is hacked and controlled by a hacker or the data somehow leaked, data decryption needs the user's private key, which is not managed centrally but managed by each user.

In a decentralized system, the whole system is divided into smaller autonomous sub system unit with simple logic. Individuals directly interact with those sub system units; some of them can run directly on the individuals' device instead of a server. Because the sub system unit does not have to take into account of all the states of the user, it is far less complicated and easier to manage. A centralized system, on the other hand, has to take into account all the states of the system, with millions of variables all orchestrated centrally. Thus, a centralized system is a lot more complex than the decentralized system, where the sub systems individually are simpler, and therefore can contain less opportunity for hack to attack.

To put it in a more intuitive perspective, normally commercial software has 20 to 30 bugs or error in the computer program per 1000 line of code. And these bugs become security risks. For example, Windows XP contains about 40 million lines of code; suppose it is about 20 bugs per 1000 lines for Window XP then there should be about 800,000 bugs contained in the software (WIRED 2004). All of Google's services together take two billion lines of code. With that many possible security risks contained in the software, you may now understand why systems gets hacked; it is almost unavoidable. Bitcoin in comparison only took 14,000 lines of code, which means a lot fewer possible bugs.

Because creating a secure centrally managed research data management system is complex and difficult, manual processes and organizations are put in place to protect and anonymize personal data. However, as shown in the example of the Fukuoka government case, human errors are inevitable and such manual operation cost adds heavier burden to the continuing operation of the full system.

4.3.2 Storing Hash Values to Public Chain to Prevent Tempering of Data

You may ask what if the hacker gains access to the server and replaces the data with fake data? How is a decentralized system going to prevent it? After all, anyone can find the user's public key on blockchain and use that to encrypt arbitrary data to replace the original encrypted personal data. To counter that, a hash value or digital fingerprint of the original unencrypted personal data is written into blockchain. Data written into blockchain is very difficult, if not impossible, to hack, as attested by the track record of bitcoin. Data written into blockchain is very temper-resistant. It is a

lot easier to hack a central server to replace a file compared to replacing data in the blockchain. Anyone given the right to decrypt personal data can calculate the hash value of the data and compare it against the hash value stored in the blockchain. If the two hash values are the same, the authenticity of the file is proved; otherwise, we suspect the stored encrypted personal data file is not the original and must have been tampered with.

4.4 System Architecture

4.4.1 Overall System Architecture

The system built for this research project, called "RDA" (research data access) combines blockchain and cloud servers as shown in Fig. 1. RDA is divided into three parts, with data storage divided into two layers: the application server layer and the blockchain and recika middleware layer. By storing inspection data in both the application server (cloud) and the blockchain layer, instead of storing all antibody test data in both the application server and the blockchain layer, test result data can be stored more securely and uploaded and retrieved with higher processing speed. The test result data for this project is very similar to medical data of patients stored in hospitals. Various projects have examined mechanisms to manage medical personal data using cloud services. However, in the medical data management field in Japan,

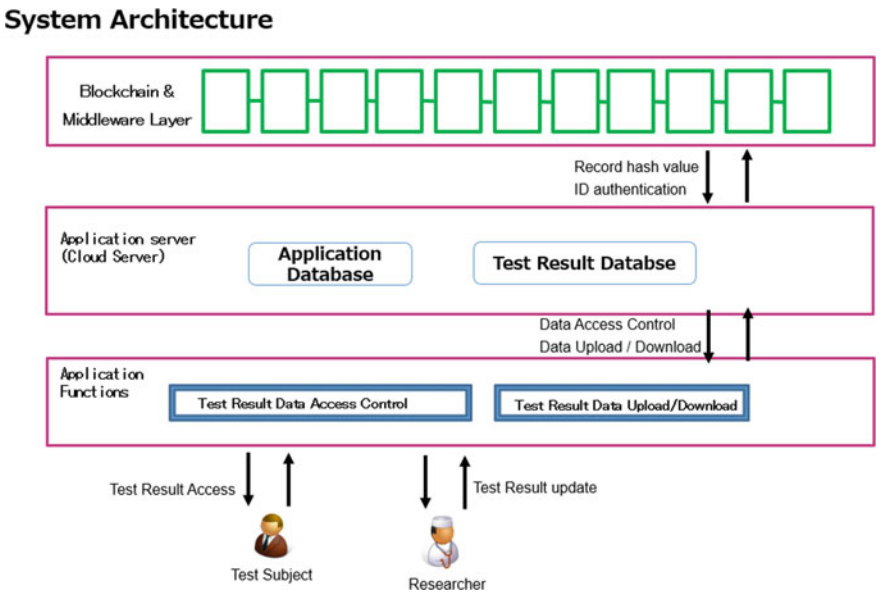

Fig. 1 Overall system architecture

people have strong safety concerns to store sensitive medical data in cloud servers. For this reason, the security of cloud services has to be strengthened through decentralized ID and encryption technology by using blockchain. By providing double security utilizing blockchain and adding a new authentication layer, we can further improve the security level of cloud servers and promote its use in the medical field. We chose to use the Ethereum chain for this project. We shall explain, with emphasis on the scalability and limitation of the blockchain, why this system architecture is most suitable for managing antibody test data.

4.4.2 Blockchain: The Challenge of Writing Data Using Ethereum

Ethereum, like bitcoin is one of the world's most widely used blockchain. Unlike Bitcoin, Ethereum can write not only transaction data such as money transfer to the blockchain, but also can write program codes called "smart contracts" to the blockchain. Ethereum public blockchain is well known to have high resistance to data tampering and hacking because of its highly decentralized structure, characterized by the absence of a centralized controlling entity. On the other hand, there is a limit to the amount of data that can be processed, and the amount of data that can be written to the Ethereum blockchain worldwide is only 20 kilobytes of data for every 15 s, compared to a normal hard disk that can write 80–160 megabytes of data per second. You can imagine Ethereum as a very slow but secure computer that every device can connect to and interact with but without the risk of being hacked.

The personal data in this project can be anything from images, videos, and text data which can add up easily to several megabytes. Theoretically it is possible to write that amount of data directly to the blockchain, but the speed mentioned above and also the associated costs that we have to pay to the blockchain for the transaction makes it difficult if not impossible to use in practice. To overcome this challenge, instead of writing the personal data itself to the blockchain, we devised a method of first writing the data to a database on a cloud server, and then writing the hash value of the data to the blockchain. The hash value is a fixed length value obtained from the original data by a certain calculation procedure (hash function), and the hash function used in this project uses the SHA-256 hash algorithm, which means that the data after being hashed is always 32 bytes. What this means is that instead of writing one megabyte of data to the blockchain, a 32-byte hash value of that one megabyte data is written to the blockchain, and this method makes it possible to increase the processing speed of the blockchain. At the same time, since the hash value will be completely different if the original data changes by even a single bit, the smallest unit of data in a computer, data tampering is easily detected. In addition, given the "transparency" of blockchain, even if the original data is written directly to the blockchain and encrypted, there is a possibility that it will be deciphered in the future. Therefore, writing only the hash value, not the original data, to the blockchain is the better choice for this project.

4.4.3 Description of the Parts of the System Architecture

In this section, we shall explain the application database, test result database, and test result data access control in more detail.

Application database: The application database is a database that stores the necessary data to run the application. For example, it stores data such as application configuration values, chat logs, and system access logs. Some of the data stored in the application database (for example, chat logs and system access logs) can be periodically processed to have hash values calculated and written to the blockchain to prevent data tampering. If a hacker gains control of the application database, and tampers with the stored data, such changes can be easily detected by a third party that later retrieves the altered data.

Test result database: As mentioned earlier, the test result data, which is personal data, is stored in the test result database in the cloud server because it is not possible to write all the test result data to the blockchain. The data are not stored in the original file format, but encrypted so it can only be decrypted with the test subject's private key. Since each test subject's test result data is encrypted individually, even if the database is hacked, it is very difficult for hackers to decrypt all the data. Thus, this decentralized form of data storage/encryption is more secure. It is far less cost effective to hack decentralized encrypted data than a centralized server with centralized encrypted data. In fact, most databases managed centrally have this risk. Once the walls of the security measures are broken and hackers get control of the server, data is easily stolen.

Test result data access control: One of the important features of this project is the blockchain-based access control, which was implemented to add double security by combining it with the cloud access control (Fig. 2).

First, in order for a test subject to successfully access his or her test data, the test subject must create a decentralized ID and password pair, or public key and private key pair. The test subject will have the test administered and receive an ID to identify that particular test. Then the test subject submits his or her own decentralized ID and test ID to link the two. On the researcher side, after the test is administered, if the test ID is linked with a decentralized ID, then the test data is encrypted in such a way that it can only be decrypted with the test subject's private key. This encrypted data then is uploaded to the cloud. In order for the test subject to access his or her own test data, the test subject must have the private key to unlock the encrypted data in addition to the ID and password to access the cloud database. This system scheme provides double layers of security because even if the encrypted data in the test result database is stolen, no one can read the original data or identify who owns that data.

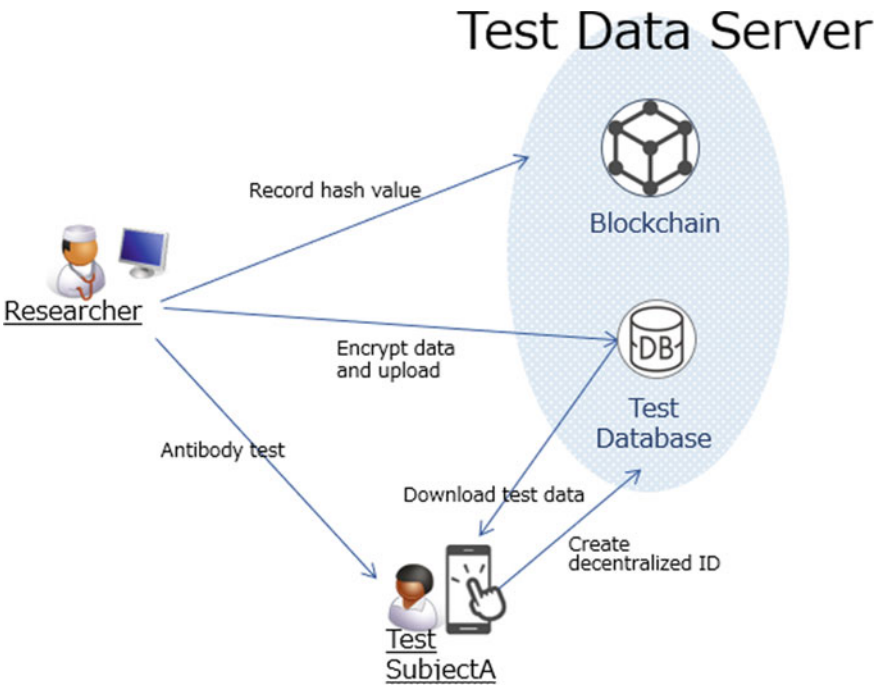

Fig. 2 Blockchain-based access control for antibody test data

5 Looking Forward

5.1 Challenges for the Decentralized Ownership Model

We have talked about the potential of using blockchain technology for a decentralized data ownership model and the many benefits it offers for storing personal data securely. There are several challenges to implement the technology in an actual use case. First is the challenge of not forgetting the password or private key. Internet users are very comfortable with a centralized server managing their ID and password and if either or both are forgotten, ID and password can be reissued. However, with a pure decentralized system, the private key is the only thing that can identify a person; thus losing the private key is equal to losing that data. Also the cost of recording all transaction data in the public blockchain is expensive. At the time of writing this chapter, the cost of writing a single transaction into Ethereum is around 30 USD, too expensive for most use cases.

The solution to both challenges lies in how to keep the balance between centralization and decentralization. A pure centralized system or a pure decentralized system only exist as concepts, but in reality most systems fall in between those idealized

concepts. For the ID and password issue, we can make only a portion of data management decentralized, but have a central entity to manage personal data offline. For individuals who forget their private key, a more analog process of reclaiming a new set of public keys and private keys can be implemented. The solution may not be elegant but can still minimize the risk of leaking data by utilizing blockchain technology. The approach to lower the cost of using Ethereum public blockchain is similar too. There are different types of blockchain with varying degrees of decentralization. A public blockchain is considered the most decentralized but also most costly to use. Therefore, by sorting out the types of transactions and use consortium blockchain and private blockchain for recording of most transactions and only record hash value of the private blockchain or consortium blockchain chain periodically to ensure immutability can be one solution that gives up some decentralization for cost. These solutions are being tested in this research project and we shall analyze the results in the future.

5.2 Data Sharing Between Research Teams

Further in our research, we would like to see how accumulated test data can be shared between different research teams to add value to scientific research in general. In order for another research team to access the test subject's data, a prior permission needs to be obtained from the test subject. This permission is normally done on paper, but the existence of that paper permission itself is personal data and can identify the test subject. A better way of digitally obtaining the permission and also analyzing the data is required. The decentralized ID created by the test subject may be a good way for the test subject to give permission to use his or her data anonymously. The test subject can use its private key in the form of a "digital signature" to give permission to access data.

For a decentralized data ownership model to truly function, we have to change the old model of passing the data to a third-party researcher to analyze the data. Instead, a better solution might be that the researcher provides the algorithm to analyze the data. Then in a safe computation environment, the data from the test subject is temporarily decrypted and analyzed with the algorithm provided by the researcher. The analysis result will be passed to the research but the input personal data is kept secret. To make this happen, technology advancement on zero knowledge proof, homomorphic encryption, secure hardware computation, and others are need. Although it may take some time before all these technologies mature, hopes are high that, in the near future, decentralized data ownership and data sharing are going to change the way we manage data.

References

Dai C (2020) DEX: A DApp for the Decentralized Marketplace. In: Yano M et al (eds) Blockchain and Crypto Currency. Springer, Tokyo, pp 95–106

Guardian (2019) Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian. 17 March 2019. www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Mainichi Japan (2021) Personal info on 9500 coronavirus patients possibly leaked in southwest Japan. The Mainichi. 8 Jan 2021 https://mainichi.jp/english/articles/20210108/p2a/00m/0na/006000c

Metcalfe W (2020) Ethereum, Smart Contracts, DApps. In: Yano et al. (ed) Blockchain and Crypto Currency, Chapter 5. Springer, 77–94

Omote K, Yano M (2020) Bitcoin and blockchain technology. In: Yano M et al (eds) Blockchain and Crypto Currency. Springer, Tokyo, pp 129–136

Pu S (2020) Industrial applications of blockchain to IoT data. In: Yano M et al (eds) Blockchain and Crypto Currency. Springer, Tokyo, pp 41–58

Pu S, Yano M (2020) Market quality approach to IoT data on blockchain big data. In: Yano M et al (eds) Blockchain and Crypto Currency. Springer, Tokyo, pp 21–40

Regulation (EU) 2016/679 of the European Parliament and of the Council 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union

WIRED (2004) Linux: Fewer Bugs Than Rivals. WIRED. 14 Dec 2004. www.wired.com/2004/12/linux-fewer-bugs-than-rivals/

Yano M (2020) Theory of Money: From Ancient Japanese Copper Coins to Virtual Currencies. In: Yano M et al (eds) Blockchain and Crypto Currency. Springer, Tokyo, pp 59–76

Yano M, Dai C, Masuda K, Kishimoto Y (2020) Creation of Blockchain and a New Ecosystem. In: Yano M et al (eds) Blockchain and Crypto Currency. Springer, Tokyo, pp 1–20

Chis Dai blockchain evangelist and founder of Recika Co. Ltd. He focuses on inventing decentralized businesses in the non-financial sector utilizing blockchain technology. An early investor of bitcoin and Ethereum and deeply attracted to the decentralized model and its business, social and philosophical aspect. Co-authored Next Blockchain: Ecosystem for Next Generation Industries (2019) and Blockchain and Crypto Currency (2020). He received BS in Management Science and Engineering from Stanford University in 2004.

Tadaaki Chigusa is the CEO of Two Plus Associates Co. Ltd. He served as a CEO of Tokyo Bay Network Co. Ltd., a cable TV company that covers the Central Tokyo. Until December 1999, he was a director at McKinsey & Co., which he joined in 1974. He started his career with Toyota Motor Corp. as an automotive design engineer. Tadaaki Chigusa holds an MS in Management from the Massachusetts Institute of Technology, an MS in Operations Research from the Illinois Institute of Technology, and a BE and ME in Mechanical Engineering from the University of Tokyo. Recently, he was also a research student at the Institute of Economic Research, Kyoto University.

Makoto Yano is Chairman of the Research Institute of Economy, Trade and Industry (RIETI); he is also a specially appointed professor at Kyoto University and Sophia University. He is an internationally known researcher who has made a number of substantial contributions in international trade, and especially on economic dynamics. In a series of recent research, Yano proposes market quality theory, addressing various problems in modern economies, including the financial market crisis since 2008 and the recent nuclear accidents in Japan, from the point of view

of market quality. Concerning quality of competition, quality of information, and quality of products, market quality is defined as an index jointly determined by the efficiency of an allocation and the fairness of the prices that are achieved in a market. An influence of his theory can be seen in Krishnendu Dastidar's book, Oligopoly, Auctions and Market Quality (2017), included in the same Springer book series as the present volume. Yano received a BA in economics from the University of Tokyo in 1971 and a PhD in economics from the University of Rochester in 1981.