D6.3: Comparison of existing blockchain technologies to safeguard responsible OS

Authors: Arild Johan Jansen & Svein Ølnes

Editor: Olivier Le Gall

Project title: Responsible Open science in Europe

Project acronym: ROSiE

Grant Agreement no.: 101006430

Lead contractor for this deliverable: University of Oslo

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

Deliverable factsheet:

Project Number:	Horizon 2020 GA #101006430
Project Acronym:	ROSiE
Project Title:	Responsible Open Science in Europe
Title of Deliverable:	Comparison of existing blockchain technologies to safeguard responsible OS
Work Package:	WP6 "The ROSiE Knowledge Hub for Open Science"
Due date according to contract:	M18 – 31 August 2022
Editor(s):	Olivier Le Gall
Contributor(s):	Arild Johan Jansen & Svein Ølnes
Reviewer(s):	Olivier Le Gall
Approved by	Rosemarie de la Cruz Bernabe

ABSTRACT:	This report examines the roles Blockchain technology (BCT)–based solutions may play to promote and support European and national responsible OS data infrastructures. The main argument is that blockchains can help implement legal and ethical requirements as well as the FAIR principles (Findable, Accessible, Interoperable and Reusable), and in particular security functions such as integrity, confidentiality and authentication of data as well as prevent falsification or misuse. Using encryption techniques, timestamping and hash functions (digital "fingerprint"), BCT offers various ways to protect and secure source data as well as program code that result from research. The use of BCT can also be supported by different EU initiatives that aim at stimulating innovation in and diffusion of BCTs, e.g. the EU Blockchain Partnership and EU Blockchain Observatory and Forum. However, the use of BCT also implies a number of legal issues, not least to comply with the EU General Data Protection Regulation (GDPR). Accordingly, these issues need sufficient attention in the design as well as the use of BCT-based systems.
Keyword List:	

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

Consortium:

	ROLE	NAME	Short Name	Country
1.	Coordinator	University of Oslo	UiO	Norway
2.	Partner	Austrian Agency for Research Integrity	OeAWI	Austria
3.	Partner	European Citizen Science Association	ECSA	Germany
4.	Partner	European Network of Research Ethics Committees	EUREC	Germany
5.	Partner	Federation of Finnish Learned Societies	TSV	Finland
6.	Partner	High Council for the Evaluation of Research and Higher Education	Hcéres	France
7.	Partner	National Research Institute for Agriculture, Food and the Environment	INRAe	France
8.	Partner	National Technical University of Athens	NTUA	Greece
9.	Partner	Universidade Católica Portuguesa	UCP	Portugal
10.	Partner	University of Latvia	UL	Latvia
11.	Partner	University of Tartu	UT	Estonia
12.	Partner	University of South-East Norway	USN	Norway

Revision history:

VERSION	DATE	Revised by	Reason
v1 (draft)	17 Feb 22	Arild J. Jansen & Svein Ølnes	v1 produced
	21 Feb 22	Olivier Le Gall	Reviewing of v1
	10 Mar 22	WP6 partners	Discussion of v1
v2	06 May 22	Arild J. Jansen & Svein Ølnes	v2 produced
	09 May 22	WP6 partners	Discussion of v2
	29 Aug 22	Olivier Le Gall	Reviewing of v2
v3	30 Aug 22	Arild J. Jansen & Svein Ølnes	v3 produced
v4 (final)	31 Aug 22	Olivier Le Gall	v3 edited to produce v4

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

The aim of this report is to examine what potential roles Blockchain technology (BCT)–based solutions may play to promote and support European and national responsible Open Science data infrastructures. Thus, we limit our discussion mainly to the capabilities and functions of BCT that may be most relevant in this context.

1 Blockchain technology in OS Data infrastructures - introduction

We believe that Blockchain technologies (BCTs) can play an important role in developing open science (OS) data infrastructures. The main argument is that blockchains can help implementing legal and ethical requirements, among them the FAIR principles of OS (Findable, Accessible, Interoperable, and Reusable data), and in particular security functions such as integrity, confidentiality and authentication of data as well as prevent falsification or misuse. Using encryption techniques, timestamping and hash functions (the hash value as result of a hash function described as a digital "fingerprint"), BCT offers various ways to protect and secure source data as well as program code that result from the research. To protect program code or other digital representation of methods and procedures may be difficult by traditional database techniques, while BCT provides appropriate tools as e.g. digital fingerprints, smart contracts, tokens etc.

In a short-term perspective, we do not see that OS data themselves may be stored on a blockchain. Rather, OS data will be kept in traditional databases, as e.g. on an OS data cloud, or stored in a distributed systems such as Interplanetary File System (IPFS) and other similar solutions. Following this we argue that selected metadata may be stored on-chain, such as data descriptors (title, keywords, etc.), author identification credentials, possible licenses and conditions for use, etc. In addition, one may include a digital fingerprint for later verification and, if necessary, also authentication. In this way, researchers can make their data accessible by an access key stored on-chain, creating a quasi-immutable record of initial ownership, and even encode 'smart' contracts or tokens to license the use of data. In the case of program code; by storing an access code and a checksum of the code on-chain, it will be possible to prevent, or at least hamper, misuse or forgery (Smith & Sandbrink, 2022).

Smart contracts are simple programs stored on a blockchain that run when predetermined conditions are met. They are typically used to automate the execution of an agreement so that all participants can be immediately certain of the outcome, without any intermediary's involvement.

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

1.1 Mapping metadata linked to Open Science data

When discussing what types of metadata may be relevant to store on-chain, we need a systematic mapping and classification of relevant metadata along with the research data itself. In addition to the necessary descriptive and identifying metadata, various metadata related to organizational and legal matters are required. The context will be the GDPR requirements, the FAIR principles, as well as other relevant legal and RE/RI requirements, e.g. the IPR (Intellectual property) legislation.

One fruitful way to categorise metadata may be to follow the structure used in the EU EOSC Interoperability Framework[1], which distinguishes four layers:

1. Technical: Metadata describing security and privacy requirements, formats, syntax, software details, etc.
2. Semantic: Description of concepts, metadata, data schemes in standardized ways such as the W3C recommendation Linked Data expressed in RDF, OWL, SKOS[2] and other standards
3. Organisational: Descriptive metadata: Title, authors, research field/discipline, source, publisher, license information, managerial issues…
4. Legal: GDPR compliance and license requirement in machine-readable format, restriction data access

Standards for metadata of types 1 and 2 are e.g. Dublin Core, while metadata of types 3 and 4 will be dependent on the type of research area. As stated above, we will also need metadata to describe relevant requirements related to the GDPR regulation as well as how to comply with the FAIR principles.

[1] https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/european-interoperability-framework-detail/

[2] RDF = Resource Description Framework, OWL = Web Ontology Language, SKOS = Simple Knowledge Organisation System

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

Departing from these categories, the following classification scheme can be used:

Type (layer) of metadata	Examples of data elements	Requirements
Legal metadata	Owner	Globally unique and persistent ID e.g. SSI/DID[3]
	Intellectual property rights (IPR) licenses	IPR management plan
	Data protection information	GDPR requirements, DPIA (Impact Assessment)
	Consent description	Consent requirement /coverage/duration
Organizational metadata	Title, authors, owner	
	Dates, versions, source, publisher	Must use an accepted standard, e.g. Dublin Core
	Metrics	Data from one /more metrics
Semantic meta	Description of variables	Use of semantic standards
Technical metadata	Formats, syntax, security standards,	Technical & syntactic description, security & privacy, (pseudo) anonymization information

Table 1: Categories of metadata relevant for OS Data

1.2 Blockchain basic

Blockchain and the broader concept Distributed ledger technology (DLT) acts as an umbrella for different distributed and decentralised systems. The National Institute of Standards and Technology, NIST, defines blockchains this way[4]:

> *Blockchains are distributed digital ledgers of cryptographically signed transactions that are grouped into blocks. Each block is cryptographically linked to the previous one (making it tamper evident) after validation and undergoing a consensus decision. As new blocks are added, older blocks become more difficult to modify (creating tamper resistance). New blocks are replicated across copies of the ledger within the network, and any conflicts are resolved automatically using established rules.*

1.2.1 Bitcoin's innovation

Blockchain technology has gotten the most attention since it first appeared as an architectural foundation for Bitcoin (Nakamoto, 2008). However, the idea of hash-linked and time-stamped documents dates back to work by the researchers Haber and Stornetta (1990). The true innovation in Bitcoin was the novel combination of well-known technologies in a way that opened for transferring value without the need for a third-party authorisation. Especially the way the consensus method Proof of Work (PoW) was paired with the currency distribution, the difficulty adjustment, and the halving mechanism of the bitcoin supply.

[3] SSI = Self-Sovereign ID, DID = Decentralised Identity
[4] NISTIR 8202: Blockchain Technology Overview (2018)

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

Bitcoin was built on well-established research and standards in cryptography, including earlier attempts to create virtual currencies (Back, 2002; Chaum, 1983; Dai, 1998; Szabo, 2008). The core technological principles of Bitcoin are (1) the peer-to-peer architecture where full nodes keep copies of the complete database, (2) the use of an open, append-only chain of block as storage, including hash linking and time stamping, and (3) the consensus mechanisms framing the rules and the security model (Valkenburgh, 2016). The name 'blockchain' comes from the time-stamped ordering of transactions into batches called blocks. The blocks are connected by hash-pointers; the cryptographic hash of the previous block is included in the next block and thus forms a chain (see Figure 1). This arrangement creates a tamper-evident structure where attempts to manipulate transactions will be easily discovered. However, it is the consensus method that ensures the security and resistance to manipulation: while hash-pointers make a blockchain *tamper-evident*, the PoW consensus method (or other consensus methods) makes a blockchain *tamper-resistant* (Ølnes, 2021).

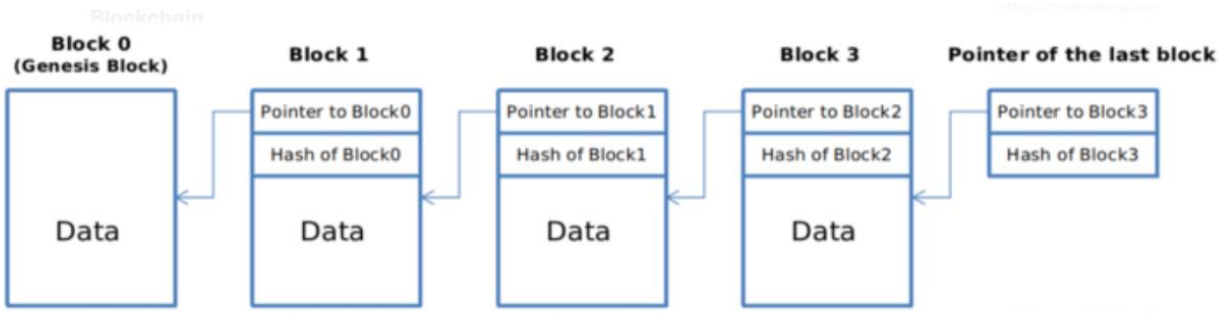

Figure 1: Illustration of hash-linked blocks forming a blockchain (zhaohuabing.com)

1.2.2 Decentralisation and distribution

BCT does, to a large extent, build on and extend the Internet architecture. Internet was originally designed as a distributed network to withstand attacks from the outside. One of its founders, Paul Baran, distinguished between centralised, decentralised, and distributed communications networks (Baran 1964), as illustrated in the figure below. Unfortunately, in blockchain literature the terms "distributed" and "decentralised" are often used interchangeably. We hold, however, that these terms should be distinguished to highlight their special characteristics and to avoid misunderstandings.

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

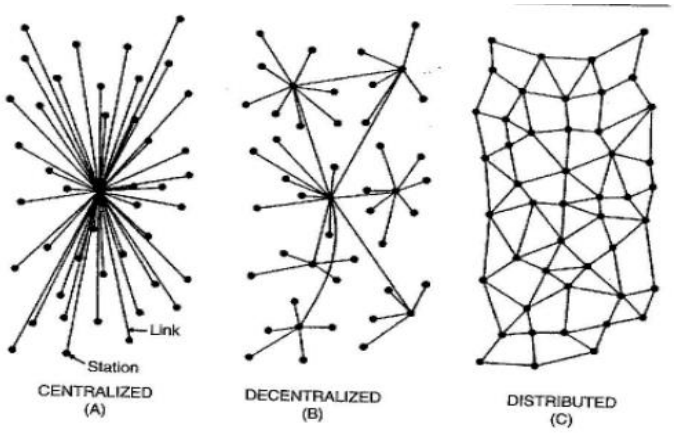

Figure 2: Paul Baran's illustration of a centralized, decentralized, and distributed system (Baran, 1964)

To our understanding, in a *decentralized system* the lower-level components operate on local information, based on delegated decision power from a central component (authority), whereas a *distributed* "system" (more precisely a *network)* consists of a collection of autonomous, cooperating units.

On the other hand, a distributed architecture consists of a collection of autonomous nodes linked by a network and operating according to a set of common rules. Distributed autonomous nodes can function without a central unit in operation (as the case of Internet). However, in distributed systems as well, some type of "body" must define and maintain the necessary common rules, (as e.g. the Internet Engineering Task Force, IETF, in the case of the Internet, i.e. the open, global community concerned with the evolution of the Internet architecture and protocols). For an open blockchain such as Bitcoin, there must be a "responsible" community mandated to make necessary decisions. Following from this, a decentralized system is still bounded, while a distributed architecture is as an open network: it may be extended continuously without changing its way of functioning.

BCT is based on a peer-to-peer network that resembles the distributed architecture part of Figure 2, where the full nodes[5] can act both as client and server for the users in that they keep a copy of the database. This is in contrast to the more common client-server model, e.g. the WWW architecture.

1.2.3 Permissionless and permissioned BCT

A key challenge for peer-to-peer systems is to agree on the state of the system. Without a middle-man or a trusted third-party, there is no authority to tell what transactions should be accepted or be rejected. The nodes need to have a method

[5] A Bitcoin full node is a node (typically a pc) that has downloaded the complete blockchain and the Bitcoin Core software and thus are able to receive and forward transactions after verifying them.

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

for agreeing on the current state of the network; that is the consensus method. In many blockchains, including Bitcoin, the consensus method is called Proof of Work and is based on proving that a certain amount of work (translated to computing power) has been spent to be able to add transactions to the blockchain (Narayanan et al., 2016)[6].

			READ	WRITE	COMMIT	EXAMPLE
BLOCKCHAIN TYPES	OPEN	Public permissionless	Open to anyone	Anyone	Anyone	Bitcoin, Ethereum
		Public permissioned	Open to anyone	Authorised participants	All or subset of authorised participants	Supply chain ledger for retail brand viewable by public
	CLOSED	Consortium	Restricted to an authorised set of participants	Authorised participants	All or subset of authorised participants	Multiple banks operating a shared ledger
		Private permissioned "enterprise"	Fully private or restricted to a limited set of authorised nodes	Network operator only	Network operator only	External bank ledger shared between parent company and subsidiaries

Figure 3: Main types of blockchains segmented by permission model (Hileman and Rauchs (2017), from OECD (2018))

Blockchains can be completely open (public and permissionless) or completely closed (private and permissioned) as Figure 3 shows. Bitcoin and Ethereum[7] are examples of completely open blockchains where anyone can participate in any of the operations of the blockchain (read, transact ["write"], and add new transactions/blocks to the blockchain ["commit"])[8].

Permissionless blockchains are, however, implicitly public since the term "permissionless" indicates that all parts of the blockchain is open for anyone to participate in. The concept "permissioned", however, needs to be further explained; a permissioned blockchain can be open for anyone to read and/or transact, but not for the commitment part (control functions). An example would be a central bank issuing a digital currency on a permissioned blockchain. It would be public in the sense that anyone could transact with it and possibly also read from it. Still, the

[6] Bitcoin's reliance on PoW ensures a high level of security but also a high demand in energy in computational power; other BCTs do not share this characteristic, for instance when their consensus method relies on a Proof of Stake (Schmidt & Powel, 2021).

[7] Ethereum is a blockchain system launched in 2015 as a result of some software developers that wanted a blockchain system with more programming functionality and flexibility.

[8] The terminology is a bit ambiguous here, we often see the term "public blockchain" used of a public, permissioned blockchain, and correspondingly "private blockchain" used of a private, permissioned blockchain. One should, nevertheless, distinguish between the public/private and the permissionless/permissioned part.

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

consensus part would be carried out by a group of trusted actors. Blockchains used in an OS data context must be open (publicly available). It can, however, (initially) be feasible to use permissioned BCT to comply with what we understand is the current EU recommendations in this field.

1.2.4 The new architecture of trust

The users of ordinary digital services must trust the service provider or the authority controlling the service. In blockchains, at least permissionless ones, there are no authority or trusted third-party. However, the need for trust does not disappear. Bitcoin advocate Andreas Antonopoulos (2014) argues that BCT enables a shift from trusting people to trusting mathematics. Kevin Werbach, on the other hand, argues that BCT represents a new architecture for trust (2018)[9], reflecting the shift from trusting a third-party or a group of pre-qualified validators to trusting the dynamic distribution of powers between key stakeholders and actors. In Bitcoin, the most important of these are the full-node users, the miners (validators), and the developers.

1.2.5 Smart contracts and tokens

The term smart contract was first used and described by Nick Szabo (1997) and a definition of the term is "an automatable and enforceable agreement" (Clack et al., 2016). Automatable refers to the execution by computers and enforceable refers to legal enforcement of rights and obligations (ibid.). Although smart contracts are created by agents outside the blockchain, they are self-contained and not controlled by any private keys, and cannot be modified. They should be accessible for all, and have to be initiated by a transaction to the blockchain in the first place as they cannot self-execute (Caldarelli, 2020).

The Ethereum blockchain made smart contracts its focal point. The founders of Ethereum found Bitcoin's programming language too limited and developed a blockchain with a Turing-complete programming language interpreted by an Ethereum Virtual Machine (EVM) (Buterin, 2013). A Turing-complete programming language means that all possible algorithms can be expressed in it. This way Ethereum has paved the way for a wider use of BCT than Bitcoin originally did.

A (digital) token is an object (in software or in hardware) that represents the right to perform some operation, or represent any kind of asset or utility. Blockchain technology can store transactions of all kinds of tokens, including domain names, identity records (e.g. driving licenses), ownership deeds, public records (such as land

[9] Bitcoin can serve as an illustration of the new architecture for trust where anyone who understands the Bitcoin protocol can be confident that the network will generate a particular quantity of new bitcoin (6.25 bitcoin at the moment) under specific conditions (whenever a miner finds a correct answer to the hash challenge and produces a new block) and at a particular pace (within an average of 10 min), without the need to rely on any financial institution or other centralized authority (De Filippi et al., 2020).

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

titles) etc. Crypto tokens are a type of cryptocurrency that represents an asset or specific use and reside on their own blockchain. Other types are security tokens or cryptographic tokens, physical objects for computer authentication, and access tokens, representing the subject of access control operations[10]. Consequentially, blockchain configurations are also being explored in the copyright domain. For example: if tokens represent rights, and digital account ('wallet') holders represent rights holders, blockchains may host public copyright registries that record, in a transparent manner, the ownership, distribution, and use and remuneration of works[11].

1.2.6 Decentralized Autonomous Organizations (DAO)

Smart contracts are (usually) key elements in a DAO, understood as "an organization constructed by rules encoded as a computer program that is often transparent, controlled by the organization members and not influenced by a central government. In general terms DAOs are member-owned communities without centralized leadership. A DAO's financial transaction record and program rules are maintained on a blockchain"[12].

1.2.7 Immutability, oracles and data quality

The BCT does not guarantee that user data stored on-chain (as e.g. OS data or metadata) is correct; all such "external" data must be verified by a third party. BCT can only guarantee that data once stored has not been tampered with[13]. In the case of OS data, both the science data itself and the metadata have to be verified and authenticated before being registered on a blockchain. However, use of BCT techniques as encryption and hash algorithms may also be used outside the blockchains.

It is also worth mentioning that blockchains typically do not store large amounts of data. They are designed to only handle small parts of data and the actual data is therefore most often stored off-chain with only a "fingerprint" stored on the blockchain itself. By using a hash function one can guarantee that the data outside a blockchain has not been changed or manipulated as that would result in a different hash value. A hash function combined with an immutable blockchain thus

[10] For more, see e.g. https://en.wikipedia.org/wiki/Token

[11] This brings the attention to NFTs (non-fungible tokens). These tokens are based on the description above as they are unique in appearance and can be assigned to different digital or physical objects, the latter being represented by a digital object (a digital twin). NFTs are the latest hype in the blockchain universe, the technology offers interesting applications in the future.

[12] For more, see e.g. https://en.wikipedia.org/wiki/Decentralized_autonomous_organization [Accessed May 05, 2022

[13] Strictly speaking, permissionless BC are tamper-resistant, while permissioned are only tamper-evident.

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

serves to verify the authenticity of data. Immutability, however, cannot be 100 % guaranteed. Blockchains are more or less immutable.

1.2.8 The blockchain trilemma

Blockchain technology has three core properties that it aims to optimise; *decentralisation, security and scalability (capacity)*. These properties are somewhat internally conflicted and incompatible, hence the blockchain trilemma (Conti et al., 2019).

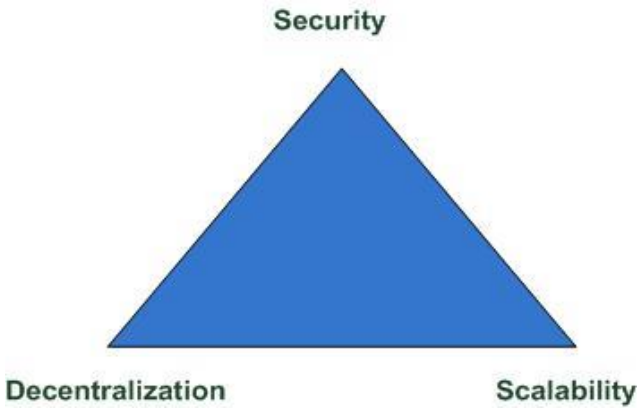

Figure 4: The Blockchain Trilemma

Only two of the three core properties can be optimized at the same time. Bitcoin sacrifices scalability (capacity) to achieve the highest possible security and decentralisation. Other blockchains have sacrificed either decentralisation (e.g. the complexity of the consensus mechanisms) or security, or both, to achieve better scalability (higher throughput, more transactions processed per unit of time). Permissioned blockchains typically sacrifice decentralisation and achieve higher throughput of transactions (more capacity, better scalability). In the context of OS data, security, understood primarily as immutability, along with decentralisation will be important aspects of a solution, while capacity may not initially be a critical factor, which may favour permissionless blockchains.

1.3 How can Blockchain technology support OS data infrastructures

The Open Science initiative is an effort to develop science from Open Access to open up the whole research process. Today we only start to share information at the point of publication and track everything from then onwards. We do not track important stages in the research process, some of which are relevant for research ethics or research integrity such as:

- the research design development
- the data collection
- the peer review process
- and more

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

In all we don't track many processes very important to reproducibility, a crucial part of science. The failing ability to reproduce scientific results is also called the "reproducibility crisis" or the "replication crisis" (Pashler & Harris, 2012). Open Science is often considered as an effort to mitigate the replication gap and BCT can be one of several an important tool in this endeavour.

Below, we will therefore address how important requirements such as confidentiality, authenticity, accessibility, and integrity of OS data, as well as auditability and authorization that can help preventing falsification or misuse of data of OS data and in other ways can safeguard responsible OS.

Our discussion departs from the assumption that OS data itself will not be stored on the blockchain. Rather, as discussed above, the digital "fingerprints" (hashes) of science data along with selected metadata will be stored on-chain. Such metadata may also include documentation of the research process. Thus, we need to assess what OS data may be published on the blockchain in each individual case. However, as the BC technology can only guarantee the integrity of on-chain data, we also have to define procedures that can secure the data quality, preferably controlled by those parties that create or own the data.

1.4 Use Case: Open Science metadata on a blockchain

To provide an illustration, we will refer to Blockcerts, a system developed for self-management of educational credentials[14], already in commercial exploitation[15]. Blockcerts stores a fingerprint of an academic credential (= a hash of the underlying credential) on public blockchain. The solution relieves the issuing body, e.g. a university, from having to verify the correctness of credentials each time it is accessed. Using this technology, individuals can take control of their own credentials through the possession of verified records, which they can use as needed. The potential of such an approach has been widely recognized (EU Blockchain Observatory and Forum, 2018; Grech & Camilleri, 2017) and several projects have already been initiated and have provided working solutions.

Accordingly, we will suggest a similar use case describing how selected metadata associated to an OS data set can be securely stored and shared on a blockchain, and in this way offer a mechanism for all relevant parties to be able to verify the correctness of the metadata. Relevant metadata can include data descriptors, author identification credentials, and possible licenses or other conditions for use. These metadata may be produced by the responsible researchers themselves, their institution or their publisher, and/or other relevant (authorised) bodies. In addition, a digital fingerprint of the dataset itself may be stored on-chain.

[14] For more details see e.g. Ølnes & Jansen (2021)
[15] https://www.blockcerts.org

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

Metadata can e.g. be organised according to the categories described in Table 1 and formatted using the Dublin Core metadata standard. A "fingerprint" of the metadata is calculated and stored on the blockchain together with the hash of a certificate identifying the responsible research organisation. The science data itself is not stored on the blockchain; only the address pointing to the data is part of the metadata. It is, however, important to note there needs to be a qualified assessment of the authenticity of the certificate when the fingerprint is first uploaded to the blockchain, most likely by the certificate issuer.

For a more robust and immutable storage of OS data a distributed solution such as the Interplanetary File System (IPFS) may be used, which implies that identical data is spread over a number of individual nodes (PCs) to prevent blocking or tampering of the data, but also to offer better accessibility. Such nodes can be part of an EU OS cloud infrastructure. This is, however, optional and the described process here will work just as fine with centrally stored data.

To summarize, the use of this system would involve the following steps:

- A fingerprint of the OS data is stored as part of other relevant metadata related to the OS data
- Selected metadata is hashed and stored on the blockchain, along with a digital certificate of the owner e.g. a research performing organisation (RPO)
- The OS data together with the metadata is stored on a distributed system such as IPFS (or it could be stored centrally, e.g. components of the EOSC – European open science cloud)
- The correctness of the metadata, and thus the research data, can be assured by comparing a "fingerprint" of the metadata obtained with the "fingerprint" stored on the blockchain
- The correctness of the research data itself can be assured by comparing a fingerprint of the data with the fingerprint stored as part of the metadata

Important issues that are relevant when designing an OS data infrastructure including the use of BCT will be:

- The degree of decentralization and distribution of data
- The capability of the participating RPOs to deliver digital certificates covering authenticity of the research data, compliance of the research protocols and other processes having yielded the data to the best RE/RI standards, etc.
- The capability of OS research and data infrastructures to manage certificates and more generally the use of the BCT
- The choice between permissionless versus permissioned blockchains, and public versus private blockchains
- Transparency versus confidentiality
- Authentication/authorisation, access control and logging of its use
- Legal and ethical questions linked to auditability, accountability and responsibility etc.
- Processes and procedures needed when parts of the research data has been updated

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

2 References

Antonopoulos, A. M. (2014, February 20). Bitcoin security model: Trust by computation. *O'Reilly Radar*. http://radar.oreilly.com/2014/02/bitcoin-security-model-trust-by-computation.html

Back, A. (2002). *Hashcash—A Senial of Service Counter-Measure*. http://c65mcoididjlt3zo.onion.city/pdf/hashcash.pdf

Baran, P. (1964). On distributed communications networks. *IEEE Transactions on Communications Systems*, *12*(1), 1–9. doi 10.1109/TCOM.1964.1088883

Buterin, V. (2013). Ethereum White Pepar—A next generation smart contract and decentralized application platform. *White Paper*. https://www.weusecoins.com/assets/pdf/library/Ethereum_white_paper-a_next_generation_smart_contract_and_decentralized_application_platform-vitalik-buterin.pdf

Caldarelli, G. (2020). Understanding the Blockchain Oracle Problem: A Call for Action. *Information*, *11*(11), 509. doi: 10.3390/info11110509

Chaum, D. (1983). Blind signatures for untraceable payments. *Advances in Cryptology*, 199–203. doi: 10.1007/978-1-4757-0602-4_18

Clack, C. D., Bakshi, V. A., & Braine, L. (2016). Smart contract templates: Foundations, design landscape and research directions. *ArXiv Preprint ArXiv:1608.00771*

Conti, M., Gangwal, A., & Todero, M. (2019). Blockchain trilemma solver algorand has dilemma over undecidable messages. *Proceedings of the 14th International Conference on Availability, Reliability and Security*, 1–8. doi 10.1145/3339252.3339255

Dai, W. (1998). B-money. *Consulted*, *1*. http://www.weidai.com/bmoney.txt

De Filippi, P., Mannan, M., & Reijers, W. (2020). Blockchain as a confidence machine: The problem of trust & challenges of governance. *Technology in Society*, *62*, 101284. doi 10.1016/j.techsoc.2020.101284

EU Blockchain Observatory and Forum. (2018). *Blockchain for Government and Public Services*. European Commission. https://www.eublockchainforum.eu/sites/default/files/reports/eu_observatory_blockchain_in_government_services_v1_2018-12-07.pdf

Grech, A., & Camilleri, A. F. (2017). *Blockchain in education*. Luxembourg: Publications Office of the European Union.

Haber, S., & Stornetta, W. S. (1990). How to time-stamp a digital document. *Conference on the Theory and Application of Cryptography*, 437–455. doi 10.1007/BF00196791

Hileman, G., & Rauchs, M. (2017). 2017 Global Blockchain Benchmarking Study.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. http://www.bitcoin.org/bitcoin.pdf [Accessed August 31st, 2022]

Narayanan, A., Bonneau, J., Felten, E., Miller, A., & Goldfeder, S. (2016). Bitcoin and cryptocurrency technologies: A comprehensive introduction. Princeton University Press.

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

OECD. (2018). *OECD and distributed ledger technology*. https://www.oecd.org/finance/blockchain

Ølnes, S. (2021). Bitcoin and Blockchain Security A Study in Misconceptions. *Norsk IKT-Konferanse for Forskning Og Utdanning*, 2.

Ølnes & Jansen (2021) In "Blockchain and the Public Sector: Theories, Reforms, and Case Studies"; Reddick, C.G., Rodríguez-Bolívar, M.P., Scholl, H.J., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 1–18. ISBN 978-3-030-55746-1

Pashler, H., & Harris, C. R. (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. Perspect Psychol Sci 7, 531–536. doi: 10.1177/1745691612463401

Schmidt, J., & Powel, F. (2021). Why Does Bitcoin Use So Much Energy? *Forbes Advisor*. Available at: https://www.forbes.com/advisor/investing/cryptocurrency/bitcoins-energy-usage-explained/ [Accessed August 26, 2022].

Smith, J. A., & Sandbrink, J. B. (2022). Biosecurity in an age of open science. PLoS Biology 20, e3001600. doi: 10.1371/journal.pbio.3001600

Szabo, N. (2008). *Bit gold*. Website/Blog. http://unenumerated.blogspot.no/2005/12/bit-gold.html [Accessed August 31st, 2022]

van Valkenburgh, P. (2016). *Open Matters—Why Permissionless Blockchains are Essential to the Future of the Internet* (p. 62). Coin Center. https://www.coincenter.org/open-matters-why-permissionless-blockchains-are-essential-to-the-future-of-the-internet/ [Accessed August 31st, 2022]

Werbach, K. (2018). The Blockchain and the New Architecture of Trust. MIT Press.

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

3 Appendix 1: Relevant EU blockchain activities

The European Commission has a number of initiatives to stimulate innovation in and diffusion of BCTs[16]. The EC, notably through Horizon 2020, has been funding several EU projects where blockchain and DLT contribute to bringing up new trust paradigms as well as societal, technical and infrastructural solutions. CEF, the Connecting Europe Facility, is already funding the EU Blockchain Services Infrastructure deployment, including use cases agreed under the EU Blockchain Partnership by the EU Commission. The Commission is also managing EU Parliament Pilot Projects, which resulted, for instance, in the creation of the EU Blockchain Observatory and Forum.

3.1.1 The European Union Blockchain Observatory and Forum

The European Union Blockchain Observatory and Forum[17] aims to accelerate blockchain innovation and the development of the blockchain ecosystem within the EU, and so help cement Europe's position as a global leader in this transformative new technology. It **facilitates dialogue between decision makers, thought leaders, and the blockchain community.** This EU Blockchain Observatory and Forum has set as one of its objectives to perform the analysis of and reporting on a wide range of important blockchain themes. Such examples are setting up Proof of Concepts and Pilot Projects to explore, test and understand legal, regulatory, policy, research and funding needs related to BCT.

3.1.2 Blockchain Strategy[18]

The EU wants to be a leader in blockchain technology, becoming an innovator in blockchain and a home to significant platforms, applications and companies. Blockchain technology allows people and organisations who may not know or trust each other to collectively agree on and permanently record information without a third-party authority. By creating trust in data in ways that were not possible before, blockchain has the potential to revolutionise how we share information and carry out transactions online. The European Commission's strategy should include:

- **Environmental sustainability**: Blockchain technology should be sustainable and energy-efficient.
- **Data protection**: Blockchain technology should be compatible with, and where possible support, Europe's strong data protection and privacy regulations.

[16] https://digital-strategy.ec.europa.eu/en/news/eu-funded-projects-blockchain-technology/

[17] https://www.eublockchainforum.eu/ https://digital-strategy.ec.europa.eu/en/policies/eu-blockchain-observatory-and-forum/

[18] https://digital-strategy.ec.europa.eu/en/policies/blockchain-strategy/

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

- **Digital Identity**: Blockchain technology should respect and enhance Europe's evolving digital Identity framework. This includes being compatible with e-signature regulations, such as eIDAS, and supporting a sensible, pragmatic decentralised and self-sovereign identity framework.
- **Cybersecurity**: Blockchain technology should be able to provide high levels of cybersecurity.
- **Interoperability**: Blockchains should be interoperable between themselves and with legacy systems in the outside world.

The European Commission strongly supports blockchain on the policy, legal and regulatory, and funding fronts. The most significant parts in its blockchain strategy include:

- **Building a pan-European public services blockchain**: The European public sector is playing a trailblazing role in blockchain by building its own blockchain infrastructure. The European Blockchain Partnership **(EBP)** [19] is bringing this vision to life. The output is the European Blockchain Services Infrastructure (EBSI).
- **Promoting legal certainty**: The Commission recognises the importance of legal certainty and a clear regulatory regime in areas relating to blockchain-based applications. It is currently developing a pro-innovation legal framework in the areas of digital assets (tokenisation) and smart contracts that protects consumers and provides legal certainty for businesses.

One important initiative by EBP is the development of a European public sector blockchain services infrastructure, which should soon be interoperable with private sector platforms. The European blockchain services infrastructure (**EBSI)**[20] consists of a peer-to-peer network of interconnected nodes running a blockchain-based services infrastructure.

3.1.3 The EU Blockchain Digital Identity[21]

Blockchain-based government services applications are unlikely to be able to operate without a strong underlying digital identity ecosystem. As part of the blockchain strategy, and one of several EBSI pilots, EU will build a new identity framework based on the concept of decentralized identities (DiD), known as self-sovereign identity (SSI). A decentralized ecosystem includes a wider universe of applications, deices and authorities, allowing users to store identity credentials in the repository of their choices and use them multiple blockchains and applications. Self-Sovereign Identities (SSI) are being seen as the next generation of digital identities across open networks. The eIDAS Regulation enables the use of electronic

[19] https://Digital-strategy.ec.europa.eu/en/policies/blockchain-partnership/

[20] https://ec.europa.eu/

[21] The European Union Agency for Cybersecurity, ENISA, is the Union's agency dedicated to achieving a high common level of cybersecurity across Europe.

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

identification and trust services by citizens, businesses, and public administrations to access online services or manage electronic transactions.

In the context of OS data, the use of secure digital identities is necessary to provide integrity and confidentiality and at the same time provide access to the data through authentication and authorization mechanisms. There are relevant standards (W3C and ISO) in the area of SSI and there exist currently 4 main commercial groups implementing SSI-based infrastructures (The *Sovrin Foundation*, a non-profit organisation established to administer governing the Sovrin Network (of blockchain nodes*), Hyperledger,* an open-source community (hosted by the Linux Foundation) developing blockchain frameworks, tools and libraries, The *European Self-Sovereign Identity Framework (ESSIF) is part of the European blockchain service infrastructure*.

However a 2020 study reflects how digital identity is essential to access most online services, and that digital identity is often outsourced to central digital identity providers, introducing a critical dependency. While SSI offers citizens ownership of their own identity, proposed solutions concentrate on data disclosure protocols and are unable to produce identity with legal status. [22]

Other EBSI use cases are:

- **diplomas**: Giving control back to citizens when managing their education credentials, significantly reducing verification costs and improving authenticity trust;
- **trusted data sharing**: Leveraging blockchain technology to securely share data amongst authorities in the EU, starting with the IOSS VAT identification numbers and import one-stop-shops amongst customs and tax authorities.
- **notarisation**: Leveraging the power of blockchain to create trusted digital audit trails, automate compliance checks in time-sensitive processes and prove data integrity;

[22] ENIS report: Digital identity Leveraging the Self-Sovereign Identity (SSI) Concept, see https://www.enisa.europa.eu/publications/digital-identity-leveraging-the-ssi-concept-to-build-trust

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

4 Appendix 2. Legal and regulatory framework for blockchain

The use of BCT also implies a number of legal issues. We believe that the compliance with GDPR[23] is most important. GDPR is a general European regulation that applies to all kind of personal data including scientific data (data acquired, collected, used/reused for research purposes, data produced by the research activities, etc.). We will briefly touch upon some other issues that also are relevant in an OS context, such as e.g. the use of smart contracts for managing authentication and authorization and access control and more general copyright and digital right management[24]. Other legal issues, as e.g. how the building of a BCT based platform (or even infrastructure) relates to EU competitions are not discussed.

As referred to in appendix 1, the EC recognises the importance of legal certainty. The EU strongly supports an EU-wide ruling for blockchain in order to avoid legal and regulatory fragmentation.

Note: The text below is entirely based on the repost "DIN SPEC 4997: Privacy by Blockchain design"[25]; the reader is advised to consult the entire DIN SPEC.

4.1.1 Blockchain and GDPR

There has been a high degree of legal uncertainty whether BCT-based solutions comply with data protection regulations, in particular the General Data Protection Regulation (GDPR). We will discuss below how relevant requirements in this regulation can be met when designing BCT-based solutions. The point of departure for these discussions is that in most cases, personal data should not be stored on the blockchain, primarily metadata and keys/hash values that are used to secure and control access to personal data stored off-line.

4.1.2 Controllership in a BCT-system

One uncertainty is that this legal framework applies to data processing in single server structures operated by a legally and technically tangible intermediary. So far, no uniform way has been found to attribute the role of "controller" (legal definition in art. 4(7) GDPR) to a specific entity within the framework of BC/DLT-systems. This constitutes a major source of legal uncertainty,

[23] see https://gdpr-info.eu/

[24] Bodó, B., Gervais, D. & Quintais, J.P. (2018) Blockchain and smart contracts: The missing link in copyright licensing? *Int. Journal of Law and Information Technology*. Vol 26,(4), p 311-336. DOI: 10.1093/ijlit/eay014

[25] https://www.din.de/en/wdc-beuth:din21:321277504

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

Determining the participants on a permissioned blockchain is relatively easy. However, it can still be a challenge to determine whether only the central authority providing permissions or if others are also to be considered as (joint-) controllers. However, different participants could also be considered as joint controllers according to art. 26 GDPR.

The situation is different for permissionless blockchain networks that include several participants, including the miners, validators, and other roles depending on the nature and particular implementation of the blockchain and its network. Here, the legal debate is still ongoing.

4.1.3 Personal data on the blockchain

Another challenge concerning the compliance with data protection regulation is determining under which circumstances the data processed in a BCT-based system is considered personal data. Accordingly, we need to analyse the degree to which, in BC/DLT-systems, a natural person is identifiable. If so, we need to find technical measures that increase the effort required to recreate any personal reference to a data subject or to mitigate the risk and therefore raise the data protection level in an IT system.

A BCT-system that operates with personal data falls into the scope of data protection regulation, such as the GDPR (art. 2(1) GDPR). Therefore it has to comply with several legal requirements. There is currently no jurisdiction concerning BCT specifically; the following section merely serves as a check-list for DLT, in particular blockchain applications, in order to create awareness for possible legal challenges in a development process that follows the ideal of "Privacy by Design principle" (art. 25 GDPR).

4.1.4 Can information about natural persons be identified on blockchains?

Art. 4(1) and recital 30 define personal data as information relating to a natural person that can be identified with that person. Personal data = information + identifier. However, both the information and the identifier do not need to be explicitly present in the data. Having only the identifier might be enough when information can be derived from the context or with the help of available external data sources. Persons are *identified* if there is an immediate inference, while they are *identifiable* if it is possible to link a data to them by intermediate steps. In any case, the controller does not have to actually have identified a natural person; it is enough if there is the potential to do so. *Hash values* can be used in BCT-systems in many contexts. When storing hash values of personal data on blockchains, these needs to be considered personal data in certain situations.

Among possible challenges, these issues must be discussed

- the immutability aspects,
- the combination of transparency of data and metadata on the blockchain,

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

- the present and future potential for data analytics,
- third party data to be linked which could identify blockchain users and other natural persons.

4.1.5 Public addresses and transaction data

Public addresses (public encryption keys) and transaction data might be considered personal data when they are combined. Most often, a pure public address is a random sequence of numbers and letters. A transaction only reveals the transactional data which is by itself of technical nature. However, when combined, public addresses and transaction data often allow the identification of transactions with natural persons.

4.1.6 Encrypted personal data

Encryption transforms the original data object into a ciphertext that conceals the information contained in the original data object. Only when an approved security function is used, is it guaranteed that the entity in possession of the private key is the only one that can recover the original data object providing the information in question.

Storing encrypted personal data on immutable blockchains might also violate ICT-security standards, because encryption on blockchains works like access control with an unchangeable password.

4.1.7 Anonymization, deletion and publication

Anonymization can be considered equal to deletion. However, in case some newly available data makes it possible to reverse the (failed) anonymization, the data is neither considered to be anonymized nor deleted. When anonymous data is published, it is yet unclear who is responsible when the anonymous data can be identified through means that were neither available nor perceivable at the time of publication.

4.1.8 Requirement of an Anonymity assessment

Since the term "personal data" is not defined in a mathematical sense, the burden to identify a natural person with some information can be increased beyond the level of recital 26 and the threshold of the European Court of Justice. This can be done through a combination of technical measures and organizational measures.

4.1.9 Right to Erasure (art. 17 GDPR)

The right to erasure is technically feasible on a single computer from a technical point of view. Implementing this GDPR request in a decentralized network is relatively more difficult because deleting it on the computer of a single node does not result in deletion across the network. Although some technical measures have been proposed to delete specific transaction content from a distributed ledger (e.g. pruning), it strongly depends on the existence of an enforceable governance mechanism that allows deletion "without undue delay" (art. 17 (1) GDPR) upon request of a data subject.

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

4.1.10 Justifications for immutability

The essence of DLT, in particular BCT to store data in a temper-proof ledger, constitute a barrier for the achievement of the GDPR goal of accuracy. It poses a technical challenge to implement the right to erasure in BC/DLT-systems. However, art. 17(3) GDPR regulates exemptions from the duty to erase data upon request of a data subject. If a DLT, in particular blockchain application can persistently comply with one of these exemptions (e.g. "processing is necessary (…) for the establishment, exercise or defence of legal claims", art. 17 (3)(e) GDPR), the immutability feature of a blockchain application can be justifiable.

4.1.11 Right to rectification (art. 16 GDPR)

When data is incorrect, the data subject has the right to demand that the data is corrected (art. 16). In BC/DLT-systems, it is not entirely clear whether an additional entry that corrects incorrect information but does not delete it entirely will be sufficient to comply with the legal requirement. Although information or transactions can be invalidated with newly appended blocks, anyone can view incorrect data. Even if this technical solution was sufficient, its implementation would still be a significant hurdle especially in large public permissioned DLT, in particular blockchain networks.

4.1.12 Data Portability (art. 20)

The right to data portability enables the data subject to request all data they have submitted to a controller in a structured, commonly used and machine-readable format. Since data portability is primarily the answer to a problem that arises in closed format systems, it can be solved more easily in BC/DLT-systems if common interoperability standards are applied.

4.1.13 Processing Agreements between Controllers and Processors

Determining what constitutes a data controller and data processor as defined by the GDPR is challenging in a decentralized or distributed environment like a network of blockchain nodes). Therefore, it is already unclear who the parties of a data processing agreement (DPA, art. 28(3)) are and which content such an agreement should have in a BC/DLT-system.

4.1.14 Identification requirements for controllers

Controllers are required to inform data subjects about the processing of personal data that is referring to them (art. 13 and art. 14 GDPR). This includes the disclosure of their identity (art. 13(1)(a) and art. 14(1)(a) GDPR). In BC/DLT-systems, this comes with legal uncertainty since the role of the controller is difficult to appoint. The GDPR classification also includes exceptions, e.g. if the data subject is already informed about the processing of the data and the controller.

4.1.15 Automated decision making (art. 22 GDPR)

The goal of art. 22 GDPR is to establish high safeguards for fully automated decision-making processes that bear legal consequences for the data subject. Even though

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

DLT, in particular blockchain applications often implement highly automated processes (e.g. smart contracts), it strongly depends on the individual use case if this functionality constitutes a decision that is covered by art. 22 GDPR.

4.1.16 Documentation + record of processing activities

By art. 5 (2) GDPR, the controller is responsible for the compliance with the data processing principles according to art. 5 (1) GDPR. This obligation presupposes that the controller must fulfil his duty of compliance by submitting all documents relating to the processing operations (e.g. by keeping a protocol). This log must be regularly updated and contain all the necessary information about the work carried out and planned.

4.1.17 Right to information

The GDPR not only standardizes duties of a controller but also regulates — in a widely understood sense — its enforcement. However, this comes with difficulties since the determination of a controller is not always easy in the case of a BC/DLT-system. Another challenge is to apply a technical procedure that is suitable to extract the necessary information from the distributed network and provide it to the data subject in a comprehensive as well as timely and effective manner.

4.1.18 Data minimization (art. 5 (1) lit. c GDPR)

The principle of data minimization consists of individual requirements: appropriate, purposeful and earmarked processing. DLT, in particular blockchain technology supports the data minimization principle when the data, stored on the blockchain, is pseudo-anonymized by the use of hash functions. In the case of an immutable blockchain, data minimization must already be taken into account during the development process, since the subsequent change is technically difficult to realize.

ROSiE has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430