Blockchain Topic Group Report

Blockchain Technology and FAIR Digital Objects - what is needed in science?

V2.0, Status of November, 2019

Editor:

Peter Wittenburg (GEDE)

Contributors:

We like to thank in particular Wolfgang Kuchinke (ECRIN), Christophe Blanchi (DONA), Petr Holub (BBMRI), Rudolf Witner (BBMRI), Larry Lannom (CNRI) and Keith Jeffery (EPOS) for their many highly valuable contributions to this document.

The Research Data Alliance is supported by the European Commission, the National Science Foundation and other U.S. agencies, and the Australian Government.

Document Revision History

2019/ May	V 1.0 Initiating the document
2019/05/24	V 1.0 Opened for comments to GEDE - Blockchain members
2019/June	V 2.0 Summarising the discussion by Peter
2019/July	V.2.0 Opened for Comments to GEDE – Blockchain members
2019/August	V2.0 Summarising the discussion by Peter
2019/08	V3.0 Current state of the document
2019/November	Finishing the document

Abstract

There is a lot of attention on Blockchain Technology based on much PR and commercial interests. It is often presented as panacea for various sorts of problems in the data domain and people seem to repeat an error which is often made in IT: People start from a "hot" technology and try to make their problem fit into the Blockchain paradigm. Blockchain technology is still new and some of its solutions are still subject of discussions. In particular, the enormous electricity and cooling requirements that are needed to run the full nodes in blockchains such as being used for Bitcoin form a huge problem. This waste of energy occurs in a time where it is known that the energy consumption should be reduced.

About GEDE

The aim of the Group of European Data Experts in RDA (GEDE-RDA) is to promote, foster and drive the discussions and consensus relating to the creation of guidelines, core components and concrete data fabric configurations, based on a bottom-up process. To achieve these goals GEDE-RDA is composed of a group European data professionals appointed by invitation from various research and e-Infrastructures and European co-chairs of Research Data Alliance (RDA) Groups. GEDE-RDA will operate within the global RDA framework, thereby guaranteeing that discussions are openly communicated and publicly accessible to the global community of experts – RDA members. For more information, see the group's web pages at https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda.

Contents

1. Executive Summary

There is a lot of attention on Blockchain Technology based on much PR and commercial interests. It is often presented as panacea for various sorts of problems in the data domain and people seem to repeat an error which is often made in IT: People start from a "hot" technology and try to make their problem fit into the Blockchain paradigm. Blockchain technology is still new and some of its solutions are still subject of discussions. In particular, the enormous electricity and cooling requirements that are needed to run the full nodes in blockchains such as being used for Bitcoin form a huge problem. This waste of energy occurs in a time where it is known that the energy consumption should be reduced.

The question addressed is what blockchain technology does offer for science in addition to traditional technologies such as database technology or Digital Object based approaches as illustrated in this paper. Current trends in research communication are determined by the term "Open" and the FAIR principles. "Openness" implies a high degree of flexibility required by data scientists knowing that too high bureaucratic hurdles would hamper scientific progress. However, there are exceptions in science where, for example, sensitive data about persons are being managed and where special measures need to be taken to adhere to rules, to raise the security level, to document all steps and to be able to trace usage. Meeting such criteria easily leads to closed systems which would be the opposite of what the term "openness" is requesting.

The masterminds behind blockchain technology are addressing the problem of high energy consumption in what is called permission-less solutions. Permission-less solutions are characterised by 1) maintaining a transparent, distributed ledger, 2) a race between full nodes about which of them is allowed to create a new block and 3) the chaining of blocks which includes the checksums of earlier blocks to make stored information tamper-proof. It is the underlying trust model that motivates serious full nodes to win the race for creating new blocks which is consuming so much energy. Various models for the race problem have been suggested, but others than the computationally intensive one used in Bitcoin-type implementations are not broadly accepted yet. The other solution to reduce energy is to use so-called "permissioned" blockchains where the second key element that characterises blockchains is omitted. Therefore, these solutions resemble very much traditional database solutions.

Summarising this paper comes to the conclusions that

- permission-less blockchain technology is not of broad use in science applications
- permissioned blockchain technology is very close to what has already been realised with other technologies
- there are special areas as for example in the health sector where special requirements for data protection ask for special technical solutions - in such applications the promises of a tamper-free and ready-made solution such as blockchains needs to be compared in detail with existing technology

This paper is mainly based on three serious documents, an elaboration from NIST experts on blockchain technology, an ERCIM special interaction on blockchain technology and an overview on blockchain usage areas. We also refer to serious statements about blockchain technology made at recent meetings.

2. Introduction

Just the other day, a large German newspaper devoted 3 pages on blockchain technology (BCT) in a special add-on section about modern digital economy addressing the question whether BCT is a hype or offers new chances. Why is this technology becoming so interesting whenever talking about "values" to be transferred or traded? We see two basic motivations to look for new ways:

- It is the urgent need to identify trustworthy mechanisms for value transfer in distributed scenarios, given that there is hardly any trust anymore in traditional institutions such as banks.
- There is the urgent need for more efficient and direct means of transferring values, given, for example, the many micro-payments that we will need to process in future or the evolving machine-to-machine value transfers.

In this paper we are focusing on the domain of science where the principles of Open Science and Open Data find broad support, i.e., digital data or other products from publicly funded research should be made available for re-use. In this context the FAIR principles [1] are also widely accepted which state that "(meta) data to be Findable, Accessible, Interoperable and Re-usable". Open Science does not imply that all research data needs to be open and also the FAIR principles do not make statements about openness, however, the "open-by-default" principle is widely supported in science. In the domain of open or fairly open data, the security aspect of trust does not play such an important role. Open Science will soon be as common as it happened with communications about research results in the 17th century as Strawn points out [2]. But in all cases where sensitive data are managed such as, for example, in the medical sector, the exchange of data can only be done based on strong trust relationships and on the application of privacy enhancing technologies [3]. There are other examples for sensitive data in the social sciences, humanities, biomedical field, etc. We should also point to data which is being protected by researchers who are collaboratively working on finding new evidences to be published which is an area of competition. These researchers will also establish special trust relationships with their collaborators before sharing data. Repositories managing large data collections are part of federations to exchange blindly their collections to create redundant copies for preservation or access optimisation purposes. Also in these cases it must be ensured that trust mechanisms have been implemented between the partners in the federation.

There may be more examples in the research domain where trust mechanisms need to be implemented before data can be exchanged. However, these aspects are not new to the research domain. Specific data has always been protected and in cases such as medical records of individuals strict governmental rules needed to be followed. Since there is no "free lunch", the question to be addressed is whether the costs of applying certain methodologies and technologies to increase security are in proper relation to the possible risks and justifiable. In this context, it is important to note that the trust establishing algorithms in permission-less blockchains[1] [4] (such as in use in Bitcoin [5]) consume an enormous amount of electrical power. The power consumption of the current Bitcoin blockchain network can be compared to the consumption of Switzerland [6]. Executing the corresponding blockchain algorithms require large clusters of high performance CPU/GPU/ASIC chips and this in a distributed fashion, i.e., trust is based on the assumption that quite a number of full nodes are participating in the exchange of the crucial transaction information to prevent fraud. Actually, the extremely high costs of maintaining such huge compute installations necessary to execute a certain class of algorithms is the guarantee that only "rich" institutions can participate as serious actors in establishing overall trust[2]. The Bitcoin network for example is in the hands of slightly more than a handful of super-nodes that win the races for being allowed to add (mining) new blocks and thus get the corresponding rewards. This waste of energy is the reason that continuously new types of mechanisms are suggested to change this strong correlation between CPU power and getting rewarded[3] [8]. These and other phenomena described elsewhere cannot give the impression that current BC technology is mature and trustworthy.

[1] When discussing blockchain technology we will often refer to the NIST report.

[2] P. Nikander [7] described in detail the requirements to run a full BC node and that current BC technology is not a panacea for IoT security.

[3] Another source which we often refer to in this paper is the special ERCIM theme on blockchains.

The aspect that research is performed in a relatively flexible domain is also important to be mentioned in the context of this paper. The group of collaborating actors and the re-use intentions in the research domain will remain highly dynamic which implies the need for a high degree of flexibility. Let's take the example of medical research where hospital A has an agreement with hospital B allowing certain medical data to be exchanged between 2 researchers for a specific purpose. It will be likely that a colleague of one of the researchers will also see the potential of the data for his/her research or that while using the data new ideas for improved re-use are evolving. The question then will be, given that repositories are managing increasing amounts of data being of interest to many different researchers worldwide, how much effort will be needed to make it possible that also the new researcher will be allowed to access the data or to use it in beforehand not intended ways. Any attempt to not only store information about explicit transactions of data but to precisely document any calculation on the data is currently far away from daily practice. Any measure which increases barriers for the reuse of other researchers' data will hamper scientific progress. There are clear practical limitations to trace all actions on data in research. Trying to establish/manage a closed system is very costly and leads to bureaucracy which can only be accepted in exceptional cases.

It needs to be noted as well that improved data sharing in science will depend on a change in culture where the use of a specific technology is only a small factor. Trust has many different dimensions[4] and only a few are addressed by technologies such as BlockChain Technology (BCT). Turning the FAIR principles into practice and certifying the processes in repositories [9] address other eventually more important dimensions in science. Often we only look at static events, but not on processes including chains of operations when we speak about trust. BCT is claiming to have especially a relevance in process scenarios which we therefore need to analyse.

Finally, we need to accept that in industry much more than in science people look for ready-made solutions that bundle a set of useful technologies which have already shown their value. This is probably one of the reasons why in industry BCT gains much more interest than in science. BCT is bundling a set of well-known technologies which are sold as a package driven by a broad developer community making it attractive. Science seems to be protected against myths that such packages offer per se a higher level of security. According to Lyons and Kahn [10] security depends on the implementation of trust raising mechanisms at various ends.

In this paper we will not elaborate on all the hypothetic questions raised in blogs such as whether blockchain technology will make "states" obsolete since all regulatory power as an intermediary agent would then be replaced by a distributed system of trust. We will limit ourselves to describe the core ideas behind FAIR Digital Objects (DO) and BCT, ways to link these technologies if necessary and discuss the benefits and costs for some application scenarios in the research domain.

3. FAIR Digital Objects

3.1 Basic Information

The first paper about Digital Objects was presented by Kahn & Wilensky in 1995 and an updated version in 2006 [11,12]. A broad analysis based on about 30 use cases from different research disciplines by the RDA Data Foundation & Terminology group resulted in the RDA DFT Core Model which put the term "Digital Object" in its focus [13]. The DO concept as developed by this group from a scientific point of view can best be described by two diagrams. The left diagram in Figure 1 states

[4] Long-term existence of data, metadata and PIDs, understanding of the context of data, relying on the correctness of the production chain, relying on the authenticity of data, trusting the procedures in repositories, etc.

that any digital object independent of its content and type (data, metadata, software, semantic assertions, etc.) can be represented by a bit sequence stored in some repositories, referenced by a persistent, unique and globally resolvable ID (PID) and described by different kinds of metadata (descriptive, deep scientific, rights, transactions, etc.). This construct is an **abstraction** of the actual scientific / semantic content encoded in DOs' bit sequences. It should be noted that metadata descriptions are themselves DOs and that of course DOs can be aggregated to collections which are also, recursively, DOs.

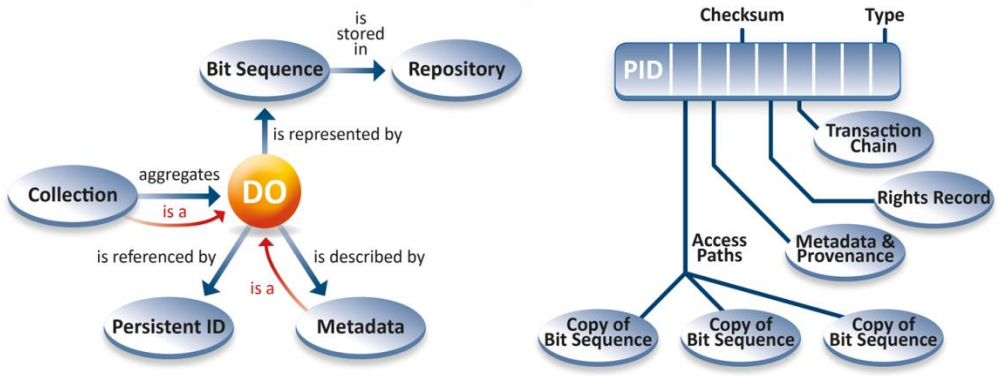


Figure 1 indicates the abstraction and binding capabilities of the Digital Object concept. Independent of the DO's content (type) it has a bit sequence stored in a repository, is referenced by a PID and described by various kinds of metadata. Metadata descriptions themselves are DOs as well as aggregated collections of DOs. The DO's PID is resolved to typical properties of the content and to references where essential information of the DO to be FAIR can be found.

The right diagram indicates the **binding** role of the DO concept. A DO's PID will be resolved to a set of standardized attributes (also called state information) which can describe properties such as checksum and type, or point to all the information relevant to access, interpret and reuse a DO's content. Both aspects, abstraction and binding, are indispensable twins. In addition, the DO's "type" enables the path towards **encapsulation** which we will not describe in detail here[5]. A recent paper from Schultes & Wittenburg [14] explains why this flavour of a DO can be called a FAIR-DO: it presents a way to implement the FAIR principles.

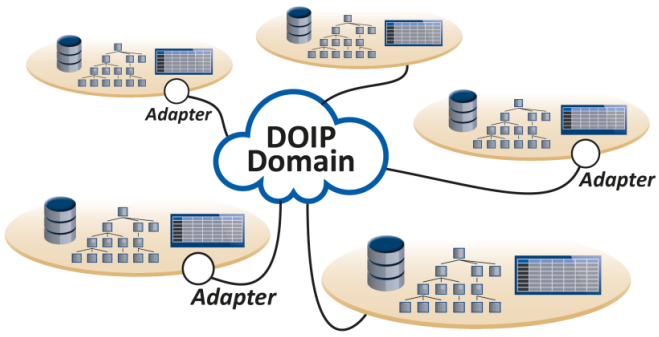


Figure 2 indicates the interoperable domain of Digital Objects independent of how the various repositories store, model and organize their data. Experience from many projects shows that they all do it differently, i.e., some use file systems, cloud objects or databases of different types. In addition, the way they store metadata types is also very different and there are often no links. DOIP helps overcoming this heterogeneity.

The recently published DOIP V2.0 protocol [15] separates the two steps of interacting a) with a PID resolution system and b) with a DO service (repository, registry) using two separate protocols.

[5] Clients can invoke operations on a DO by using standardized interfaces without knowing in detail how the DOs' bit sequences (content) are implemented.

Therefore, the DOIP protocol defines a Digital Object as having a structured bit sequence, a PID and a Type. It does not include the resolution step which can resolve a PID to other relevant information as indicated in the above RDA DFT Core Model. To avoid confusions we distinguish these two different terms, Digital Objects and FAIR Digital Objects, although practically they are complementary. The term FAIR Digital Object has also been used by the report of EC's expert group on implementing FAIR [16]. The DOIP protocol helps to create an interoperable domain of DOs independent of the way how the various repositories organise and store their data as is shown in figure 2.

In this discussion about appropriate trust mechanisms at affordable costs it is important to discuss the available components, what kind of mechanisms can be built with them, the security and the cost aspects.

3.2 Available Components

Crucial for the Digital Object Architecture is the availability of a secure, persistent and globally available resolution system for identifiers, since it is accepted that every digital entity should have a globally resolvable PID. The Handle System [17] is being used for two decades now globally and has shown its functionality and secure operation. The Global Handle Resolution System (GHR) is in the hands of the non-profit and independent Swiss DONA Foundation taking care that it will evolve smoothly, remain stable and open for participation. Currently 10 root nodes distributed globally share the burden of resolving Handles doing this in a redundant fashion guaranteeing 24/7 performance. Local Handle Resolvers (currently more than 4000 worldwide) can be operated by domains like publishers (DOI) [18], film industry (EIDR) [19], PID service providers such as ePIC [20] or simply individual institutions managing large amounts of DOs. Since the Handle System is compliant with the ITU X.1255 identifier interoperability framework [21] and the GHR System is acting on a self-sustained business model, long-term service perspectives are given. Security is given, since at the GHR level only the resolution at prefix level is being carried out by a set of selected and trustworthy national centres. The eventually sensitive local PIDs with all the information in the PID records maintained by the Local Handle Service Providers are protected by using PKI, i.e., only the PID owners have access to the records, the database is protected and the content can be encrypted.

> A Handle has the following syntax: <prefix>/<suffix>, DOIs have the prefix 10. The prefixes are specified and resolved at global resolution level, the suffixes at local level.

Additional components being available are

- The DO Interface Protocol defining a unified access to Digital Objects in Repositories.
- The Data Type Registry[6] [22] allowing specifying types of for example PID record attributes, DOs and relating DO types with operations to support machine actionability.
- The CORDA software [23] being a DOIP reference implementation for a prototypical DO repository.

3.3 Security Levels for FAIR DOs

In this chapter we will introduce a few security levels when working with FAIR DOs, which may indicate how far one can go with non-BCT technology[7]. This will help identifying the unique points of Blockchain Technology. In the following we assume that "blocks" of data are simply FAIR DOs as defined by RDA DFT, i.e., they have a structured bit sequence, a PID and metadata. For all subsequently described levels we assume that there is a secure resolver for Persistent IDs that resolves to state information which is stored as attributes in the PID records. For the purposes of this discussion we assume that the records have at least the following attributes amongst others: 1) links

[6] The Data Type Registry was specified in an RDA Working Group.

[7] We could also have referred to the extensive work in database technology including all kinds of technologies to protect data.

to the blocks (the bit sequences), 2) a hash value calculated on the bit sequence, 3) links to the metadata DO. This is indicated in figure 3.

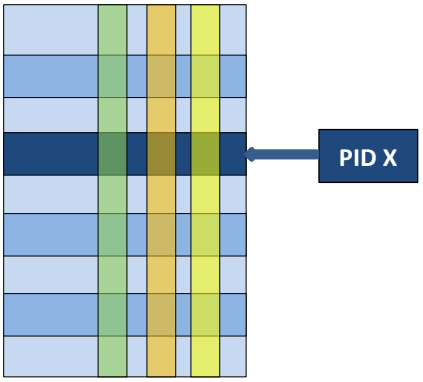


Figure 3 shows schematically the content of the database associated with the PIDs. A specific PID X addresses a certain record. The resolver exports the attributes of this record on request. For the purposes of this paper, it is assumed that each record has at least attributes that point to the bit-sequence of the data block (green), point to the metadata (orange) and stores some form of checksum calculated over the block's content (yellow). For specific security layers we will see that more information can be added.

For the discussion about trust and security mechanisms that follows we assume that the Handle System is used applying its built in features and that Handle Service providers are regularly certified. For the Global Handle System this means that

- only prefixes are managed by the network of root nodes (MPA) which do not contain private information
- only the MPAs have the rights to create and maintain prefix entries in the Handle database
- all actions are based on signed certificates (PKI) of the MPAs and the mandates of the MPAs are also based on signed certificates from the DONA foundation
- the consistency of the information exchanged between the MPAs is continuously checked automatically using a fairly straightforward consensus algorithm
- the DONA foundation evaluates MPA behaviour regularly to take measures in case of misbehaviour

For the Local Handle Systems responsible for resolving Handles with a specific prefix the following is valid:

- If people adhere to the Handle specifications defined by RFCs[8] they can build their own LHS software and thus can implement any trust mechanisms they are interested in.
- For the default Local Handle Software that can be down-loaded the following principles hold:
  - Each Handle Record has an administrator and PKI is used to sign records guaranteeing that only the administrator can change record attributes.
  - A variety of mechanisms allow administrators to make the content of Handle Records invisible to others than a selected group.
  - A LHS can be setup using different servers each of them resolving a specific group of Handles and being associated with different protection mechanisms.
  - The information in the Handle Record can be encrypted, sealed by a public key and different encryption algorithms can be used.
  - We also assume that the LHS is operated by a unit that is independent from the data creators/depositors.

Level 1: Independent Blocks

This is the usual case which many repositories in the research domain are already applying where no particular security features are required. Every block of data is treated independently although at the metadata level relationships are being specified.

[8] RFCs 3651/3652

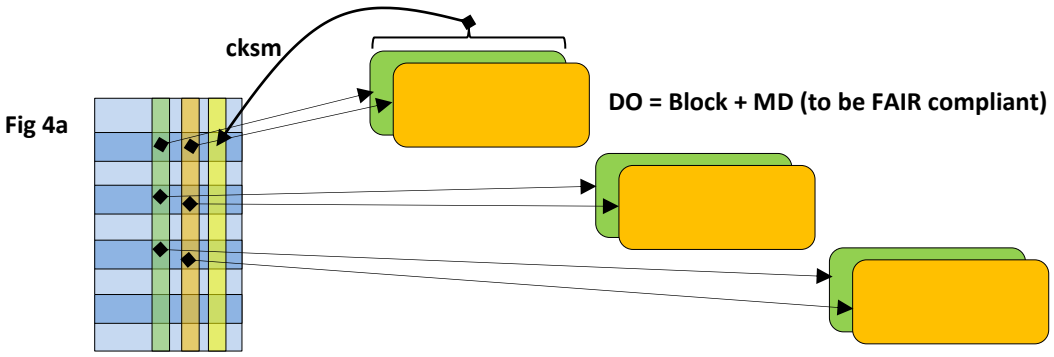


Fig 4a

Features:

- the checksum (only indicated once in the diagram) is stored in the Handle Record and can be used to check authenticity
- the security of the data in the blocks and MD is regulated via the repository AA mechanisms

Level 2: Sequence of Blocks

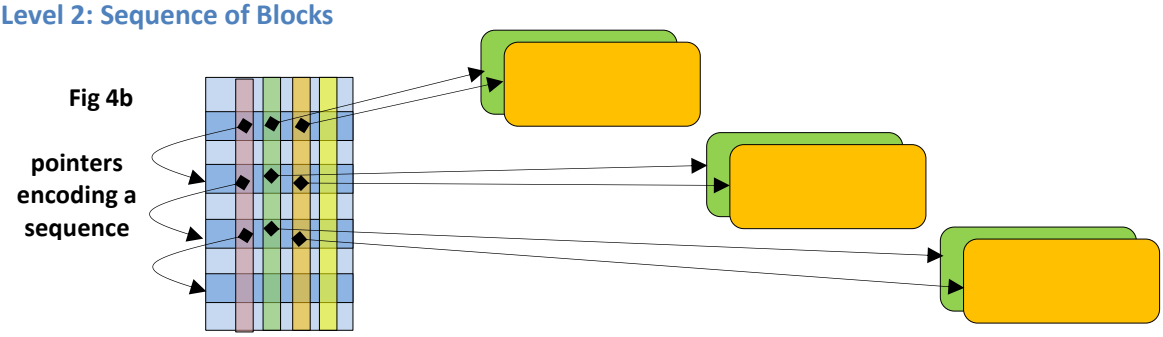


Fig 4b

An additional attribute (pink) could be used that points to the subsequent block. Some repositories are using this opportunity to point to new (and old) versions for example.

Features:

- the PID records contain information about a chain of blocks that cannot be manipulated, i.e., the PID database documents securely the chain

Level 3: Sequence of Trusty Blocks

An additional feature can be built in by using the hash values encoding the content as suffixes in the

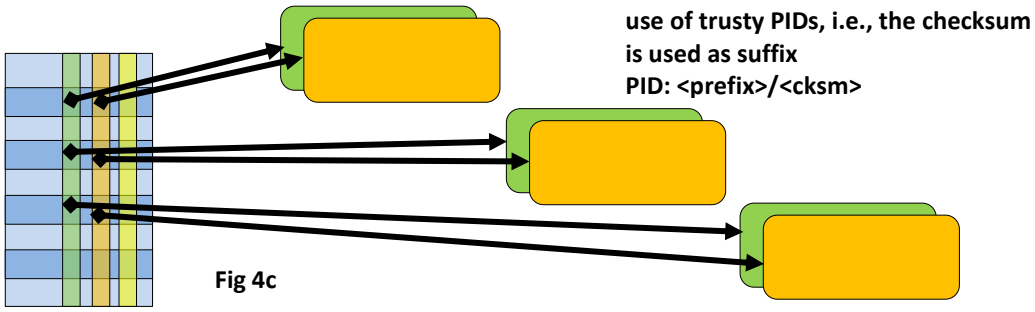


Fig 4c

PID. This type of PIDs is called Trusty IDs: <prefix>/<checksum> and used by some communities.

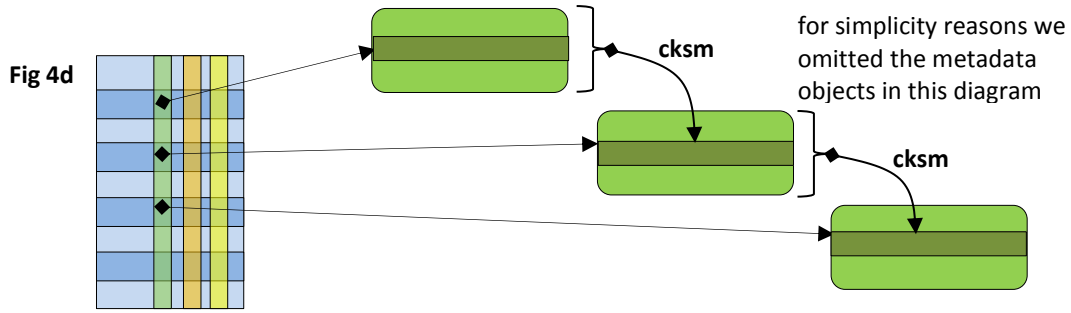
Features:

- the PIDs cannot be changed and thus manipulations can be prevented in so far as the PID itself points to the content it encodes

Level 4: Chain of Trusty Blocks

Yet another mechanism can easily be implemented when the subsequent blocks will contain the checksum of the previous block. This requires that the blocks have a header structure such as it is

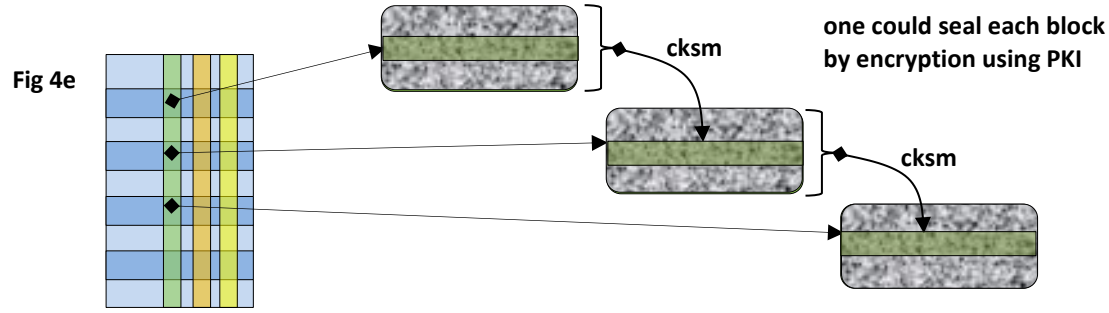
used in many scientific data formats (netCDF, DICOM, etc.). This technique is actually used in blockchains.



Features:

- the same level of tamper proof documentation as in BCT can be achieved.

Level 5: Chain of Encrypted Trusty Blocks



Of course, the content of such blocks can be encrypted using PKI technology in addition to other methods described above.

Features:

- decryption requires that the user has access to public/private keys of the creator

3.4 Summary

In this chapter we have presented FAIR Digital Objects and the components implementing them as a way to stepwise improve data security, here in particular by the smart use of the global Handle (PID) resolution system. The Handle system is implementing a trustworthy and protected domain of persistent identifiers at comparatively little costs. The Global Handle Resolution system is based on a distributed network of trustworthy nodes committed to the principles established by the DONA Foundation, i.e., all activities are non-profit and based on open protocols and specifications. This global network guarantees that the GHR is operating self-sustained and thus can be maintained over long periods.

The nodes of the Global Handle Resolution network have to define their business models which can include costs for requesting a prefix to run a Local Handle Service or for the registration of Handles. Given the strict dedication to non-profit operation the costs will be at a minimal level and various nodes could act in competition preventing that non-justified costs will be claimed. Running Handle Resolution Servers is at the low cost end since common middle class servers will be sufficient. The costs will actually be defined by the level of support that is offered.

As has been shown, various levels of security can be realised without adding costs to the system. The costs are in developing the intelligent software to implement the various features which is in the hands of the user and not defined by the Handle System. Some complex systems have already been

implemented using the Handle System. As an example, we refer to the implementation of a food supply chain control system developed in China where every sold bin with baby milk powder has a QR code that can immediately be turned into complete information about all stages of the production chain using Handles for identifying all entities and using security mechanisms to protect the system against manipulations. From the 3 key elements of blockchains this system allows an implementation of the distributed ledger and a tamper-resistant chain of digital objects. It does not require a costly mechanism to create new blocks. However, implementing these features is shifted to the application developers while BCT has it all included.

4. Blockchain Technology

There is much literature about Blockchain Technology, but most of it is authored by companies who want to create business cases, i.e., this information is often of little use. Therefore, we will here cite the executive summary of the excellent and factual report from NIST about "Blockchain Technology Overview" [4] which contains valuable statements summarizing BCT: "*Blockchains are tamper evident and tamper resistant digital ledgers implemented in a distributed fashion (i.e., without a central repository) and usually without a central authority (i.e., a bank, company, or government). At their basic level, they enable a community of users to record transactions in a shared ledger within that community, such that under normal operation of the blockchain network no transaction can be changed once published. In 2008, the blockchain idea was combined with several other technologies and computing concepts to create modern cryptocurrencies: electronic cash protected through cryptographic mechanisms instead of a central repository or authority. The first such blockchain based cryptocurrency was Bitcoin.*

Within the Bitcoin blockchain, information representing electronic cash is attached to a digital address. Bitcoin users can digitally sign and transfer rights to that information to another user and the Bitcoin blockchain records this transfer publicly, allowing all participants of the network to independently verify the validity of the transactions. The Bitcoin blockchain is stored, maintained, and collaboratively managed by a distributed group of participants. This, along with certain cryptographic mechanisms, makes the blockchain resilient to attempts to alter the ledger later (modifying blocks or forging transactions).

Because there are countless news articles and videos describing the "magic" of blockchain technology, this paper aims to describe the method behind the magic (i.e., how blockchain technology works). Arthur C. Clarke once wrote, "Any sufficiently advanced technology is indistinguishable from magic". Clarke's statement is a perfect representation for the emerging applications of blockchain technology. There is hype around the use of blockchain technology, yet the technology is not well understood. It is not magical; it will not solve all problems. As with all new technology, there is a tendency to want to apply it to every sector in every way imaginable. To help promote correct application, this document provides information necessary to develop a high-level understanding of the technology.

Blockchain technology is the foundation of modern cryptocurrencies, so named because of the heavy usage of cryptographic functions. Users utilize public and private keys to digitally sign and securely transact within the system. For cryptocurrency based blockchain networks which utilize mining, users may solve puzzles using cryptographic hash functions in hopes of being rewarded with a fixed amount of the cryptocurrency. However, blockchain technology may be more broadly applicable than cryptocurrencies. In this work, we focus on the cryptocurrency use case, since that is the primary use of the technology today; however, there is a growing interest in other sectors.

Organizations considering implementing blockchain technology need to understand fundamental aspects of the technology. For example, what happens when an organization implements a blockchain network and then decides they need to make modifications to the data stored? When using a database, modifying the actual data can be accomplished through a database query and update. Organizations must understand that while changes to the actual blockchain data may be difficult, applications using the blockchain as a data layer work around this by treating later blocks

and transactions as updates or modifications to earlier blocks and transactions. This software abstraction allows for modifications to working data, while providing a full history of changes. Another critical aspect of blockchain technology is how the participants agree that a transaction is valid. This is called "reaching consensus", and there are many models for doing so, each with positives and negatives for particular business cases. It is important to understand that a blockchain is just one part of a solution.

Blockchain implementations are often designed with a specific purpose or function. Example functions include cryptocurrencies, smart contracts (software deployed on the blockchain and executed by computers running that blockchain), and distributed ledger systems between businesses. There has been a constant stream of developments in the field of blockchain technology, with new platforms being announced constantly – the landscape is continuously changing.

There are two general high-level categories for blockchain approaches that have been identified: permissionless, and permissioned. In a permissionless blockchain network anyone can read and write to the blockchain without authorization. Permissioned blockchain networks limit participation to specific people or organizations and allow finer-grained controls. Knowing the differences between these two categories allows an organization to understand which subset of blockchain technologies may be applicable to its needs.

Despite the many variations of blockchain networks and the rapid development of new blockchain related technologies, most blockchain networks use common core concepts. Blockchains are a distributed ledger comprised of blocks. Each block is comprised of a block header containing metadata about the block, and block data containing a set of transactions and other related data. Every block header (except for the very first block of the blockchain) contains a cryptographic link to the previous block's header. Each transaction involves one or more blockchain network users and a recording of what happened, and it is digitally signed by the user who submitted the transaction.

Blockchain technology takes existing, proven concepts and merges them together into a single solution. This document explores the fundamentals of how these technologies work and the differences between blockchain approaches. This includes how the participants in the network come to agree on whether a transaction is valid and what happens when changes need to be made to an existing blockchain deployment. Additionally, this document explores when to consider using a blockchain network.

The use of blockchain technology is not a silver bullet, and there are issues that must be considered such as how to deal with malicious users, how controls are applied, and the limitations of the implementations. Beyond the technology issues that need to be considered, there are operational and governance issues that affect the behaviour of the network. For example, in permissioned blockchain networks, described later in this document, there are design issues surrounding what entity or entities will operate and govern the network for the intended user base.

Blockchain technology is still new and should be investigated with the mindset of "how could blockchain technology potentially benefit us?" rather than "how can we make our problem fit into the blockchain technology paradigm?" Organizations should treat blockchain technology like they would any other technological solution at their disposal and use it in appropriate situations."

It is obvious that when discussing the trust related aspects of BCT one needs to distinguish between permissionless BCT (PL-BCT) where everyone can participate (used in Bitcoin) and permissioned BCT (P-BCT) where an authority defines participation rules.

4.1 Permissionless BCT (PL-BCT)

The functioning of PL-BCT can easily be summarized by the following two diagrams. The first diagram (Fig 5, taken from [4]) indicates where the name is originating from. The information about the previous blocks (header, data) is being hashed and included in the header of the subsequent block. This implies that manipulating block t-N would require rolling up the entire chain since the changes would have implications for the encoding of the actual block at t. If mechanisms are established that severely reduce the chance that someone can do such an operation without being noticed and rejected by others, one could speak about a secure mechanism.

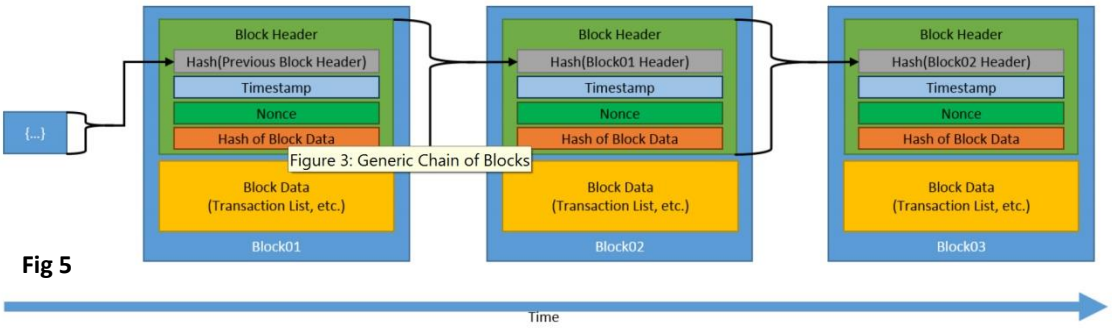


Figure 3: Generic Chain of Blocks

Fig 5

The second diagram (Fig. 6) shows the umbrella technology these chains of blocks are embedded in. New transactions are being initiated, so called miners collect some transactions to form a new block and then the miners start a race for who is allowed to add a block to the blockchain. The miner who will win the race gets rewarded in some way. The algorithms involved to solve this puzzle can be compute intensive. If there is no challenge and adding a block would be too easy the trust model would fail. There is still an ongoing debate about which class of algorithms is best[9]. The newly created blocks are exchanged between all nodes and checks are run to guarantee consistency including compute intensive consensus algorithms and algorithms to solve the problem of forking of the chain.

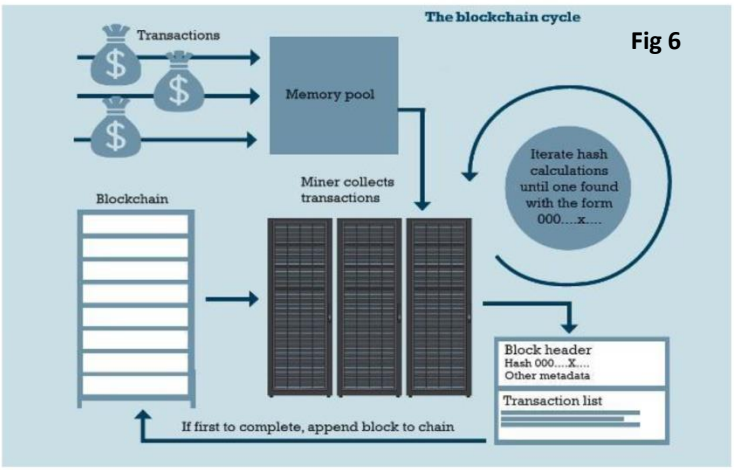


Fig 6

Pic: http://bindudeknock.com/2015/10/01/block-chain-technology-save-music-industry/

The so-called "Proof-of-Work" algorithm is based on the assumption that much CPU power is invested by the miners to do certain calculations repeatedly until one of them has reached the specified goal. These computations are the major reason why PL-BCTs such as Bitcoin are extremely unecological which today is hardly acceptable. The ambitions to win the race and thus to be rewarded with new Bitcoins are so high that specialised computing hardware is being used massively by slightly more than two handful centres who in fact "own" the Bitcoin network. Reward is strongly correlated with the CPU power installed. Other "proof of X" algorithms correlate the chances of winning the race with the amount of currencies hold, the total storage size, etc., but they are not free of weak points as well. However, the basic design of PL-BCT requires a high threshold for active participation to not open the door for too many and probably not serious enough miners. In PL-BCT trust is essentially established by

- the fact that full nodes share the whole transaction database (distributed ledger) and check consistency and correctness whenever a new block is being issued
- ensuring that sufficient trustworthy miners are active and control each other
- spending huge amounts of energy to prevent that one node is overruling all others on the one hand and not having too many dubious miners on the other
- chaining blocks and using encryption.

In cryptocurrency applications (about 600 now in addition to Bitcoin) the network of banks is replaced by a network of super users which, however, are not known and not subject of regulation

[9] See also the contributions in the special ERCIM report about BCT [8].

and control. Anonymity of end users opens the doors for doubtful businesses. BCT is still under investigation, most projects have piloting character and major weak points in addition to what has been mentioned are well known (latency due to block size, unclear scalability, renewal process in case of severe errors, difficult to implement the right to forget, merging forks and different BCs, regulation resistance).

4.2 Permissioned BC (P-BC)

In contrast to PL-BC, P-BC consists of a set of known and identifiable participants (nodes) which are given explicit permissions by a super user or a consortium. They may not fully trust each other and don't want to have others understand their business logic, but also rely on a distributed ledger to document for example transactions between them. Since the actors are well-known and will have signed contracts a completely different situation is given compared to permission-less BC. There is no need for energy-consuming races for winning the miner's game. P-BCs therefore resemble very much existing solutions of distributed databases.

4.3 Smart Contracts

There is much talk about so-called smart contracts. From an IT point of view they can be compared with what is known in the database world as "triggers". If the content of a database is going to be changed or has been changed such triggers can be defined and software code of any sort can be executed. The concept is therefore not new. Important, however, would be the step to formalise all actions and states that have to do with contracts which is needed for all the transactions and payments which we see emerging in the IoT world, for example. But it should be clear that once proper models for a formalisation of business logic have been developed they can be implemented in many different application scenarios and systems, in particular in the area of IoT devices.

It should be noticed that a) security considerations with respect to smart contracts need to be complete, i.e., they need to include the code, the execution environment and the properties of the distribution system and b) the proof of the correctness of rules embedded in smart contracts remains a hard problem.

4.4 Summary

The unawareness about the plentiful existing solutions to increase trust levels, the ready-made BCT software and the enormous PR around BCT seems to let many people forget what the core advice from the NIST report is: *solutions should be investigated with the mindset of "how could blockchain technology potentially benefit us?" rather than "how can we make our problem fit into the blockchain technology paradigm?"* We need to understand which unique features make BCT different from existing solutions and which are the applications that require those features and justify the extreme spending of electrical energy and cooling. As indicated, we will focus on PL-BCT, since P-BCT comes very close to what one can realise with other technologies.

As is frequently stated by experts, BCT is a combination of known techniques to one ready-made, closed framework which expresses the advantage of BCT, but also its major disadvantage. When comparing BCT with other solutions we can identify three unique points:

- the transparent distributed ledger and the race between unknown peer-to-peer miners for creating new blocks.
- the chaining of blocks that includes checksums of earlier blocks although this could be realised in simple ways using existing techniques as indicated.
- the gigantic waste of energy to solve the mining race issue which we cannot find in other technologies.

Based on this, we can write a few statements to narrow down the scope of useful applications in science:

- Permissioned BCT is not that appealing in science since its trust is established by "trust on the consortium" and there are many other ways to exchange information securely and to document sequences within such domains of trust. The advantage of BCT is that one gets a ready-made application with a number of features such as in the case of Ethereum [24] without the need to waste gigantic amounts of energy, i.e., some key features of PL-BCT are omitted.
- BCT defines a closed domain of facts which only makes sense in applications where this closed approach can be maintained. In science, data will live outside of such a chain of blocks, i.e., they live outside of the controlled space. Off-BCT processes are difficult and costly to integrate.
- BCT is not meant for dealing with data and the typical metadata we are using in science and it does not address the FAIRness of data/metadata.
- BCT is meant to document sequences of events such as transfers of ownership of crucial entities each of which is characterised by only a few bytes.
- BCT's block chaining can be seen as a way to track provenance, but for general scientific applications its provenance model is too simple, does not offer the usual vocabularies, etc.
- Given the high security requirements in the medical domain one could imagine storing all information about exchange and access acts on patients' data in a blockchain. Storing all patient data, which is cumulative over time can include, for example, digitised images and time series data, in a BCT is not useful except when designing special permissioned BCT applications which, however, could also be realised with other techniques. It should be noted that storing all data in one BC can amount to privacy protection problems.

5. Linking DOs and BCT

From the FAIR-DO Model view point, transaction or process information is just one other type of

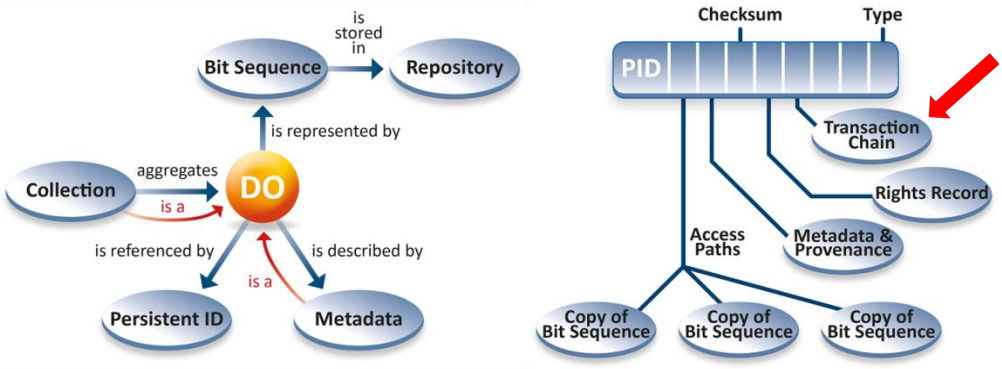


Figure 7 indicates how a FAIR-DO can easily be linked with a blockchain, since statements on transactions are also kinds of metadata assertions on the bit sequence.

metadata assertions on a bit sequence. As has been explained above the PID record does the binding of all relevant information associated with a bit sequence in a persistent and trustworthy way. Since industry is interested to use BCT for some special applications in combination with the FAIR-DO concept, an option would be simply to define and add an attribute called "transaction chain" (indicated by red arc) allowing also machines to find the link to the process information (see Fig. 7, right). This option might also be attractive for specific applications in research.

For linking there are two options since every "block" is a Digital Object and in general terms a blockchain is nothing else than a specifically linked collection which is also a DO (see Fig.7 left). Figure 8 shows a possible implementation.

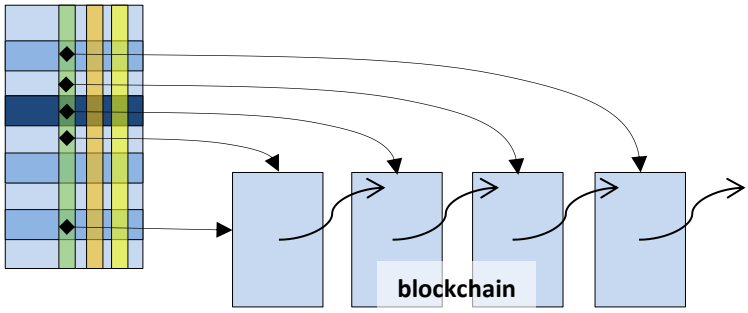


Figure 8 indicates how blocks and blockchains can be easily linked with DOs. This can include references to the individual blocks as well as the chain of blocks. The combination offers a strong combination of clear identities and where necessary the safeguarding of process information.

6. Application areas in Science

6.1 General Aspects

When reading publications such as [4], [8] and others, BCT is often suggested for application areas such as:

- protocolling all actions in autonomous vehicles in a trustworthy repository with highly accurate timestamps since one cannot assume that, for example, car builders will provide trustworthiness when accidents happen
- documenting financial transactions where valuable goods are handed over to new owners
- handling the huge amounts of micro payments in a secure and traceable way
- documenting logistic processes where every process step is clearly documented and thus delays caused by actors can exactly be indicated
- exchanging sensitive patient data between actors (hospitals, doctors, etc.)
- controlling complex supply chains to be able to trace actions
- documenting land transactions to be registered in governmental offices
- handling patent information to document with clear time stamps when which step was taken
- etc.

Except for the patient data example, most application examples being mentioned frequently, describe governmental processes documenting acts relevant for proper societal functioning or commercial processes documenting acts that are important for the revenues of companies and thus their survival. Research, in general, does not have comparable characteristics. Even in case of census or other sensitive sociological data we would claim that there is no sufficient justification at the moment to waste huge amounts of energy by using PL-BCT. An application that requires proper documentation of acts in science is, for example, the proper maintenance of a lab-book describing details of experiments. Currently, lab-books are being turned to become electronic and techniques using the Digital Object concept as discussed earlier become popular to document each experiment and seal the observations. The problem science is sometimes facing, is unethical behaviour, such as using counterfeit or manipulated data. A question raised is whether BCT can make sure that the lab-book is maintained in a manner transparent and trustworthy not only to the source institution, but also to any other institutions that may need to verify results. Since PL-BCT is energy inefficient and **sealing** of an entry is possible without BCT, an in-depth analysis of what can be done with sealing vs. whether BCT could have some advantages, would be rather valuable.

We think that the domain of medical research, where acts can decide about "life", seems to be important enough to take a deeper look at the special features of PL-BCT (distributed ledger, race of miners, chaining of blocks) to check their usefulness. Currently, personal health data are most often kept in hospitals or doctors' information systems, supporting only local access. There are, however, long-term efforts in sharing the data at least on the regional level with the intention to make the healthcare more effective [25]. A wide variety of scenarios could be presented, but here we refer to

another paper[10] [26]. In this paper we limit ourselves to looking into two simple scenarios. In the first one, patients keep their data themselves**.** This scenario is mentioned in literature very often but it is not common in real world. The reason why in some cases the patients are not given control of their data is that patients could remove critical parts of their medical documentation from the files, such as epilepsy or psychotic disorders, which would prevent them from applying to certain occupations. Integrity and authenticity of those files would need to be verified by any authorized medical professional. Non-repudiation of medical records is a very critical property here.

In the second situation the patient's medical records are stored in an information system used by a physician or hospital. It is typically a healthcare institution who is in charge of further sharing the data for treatment or research purposes within the scope of the purpose for which the patient provided their data. For example, a physician requests to see results of a medical examination made before by a different medical expert. One could imagine that tracking of related events such as getting permissions, accessing metadata/data, sending the key to decrypt data, adding the new medical report to the patient's database, etc. are in the interest of the patient. We believe that it makes sense for the patient or a regulatory body to perform integrity and trustworthiness checks for which purposes the data was shared and used. Another interesting type of event would be provable termination of access for the particular recipient of the data. This should happen after a specific project is over, or once the recipient does something which implies grounds for revoking access. Again, the question is whether BCT is necessary for these scenarios.

Another area where BCT could play a role in the medical sector has to do with medical trials which include several partners often bound to different national regulations and different time zones. The sequence of trials and the phases in which they are carried out are highly regulated and controlled and subject of contracts, which implies that related processes must be auditable. In many cases pharmaceutical companies are also involved which want to protect their investments. Such trial processes need to be managed with precise synchronisation between all actors and it needs to be documented precisely. BCT are designed to capture all relevant events in form of transactions in a transparent and non-repudiable manner in a distributed environment. We can easily imagine that a shared ledger is being used for auditability and synchronisation. But as stated above, a PL-BCT may be too expensive and energy inefficient for this use case and P-BCT features could be implemented using other technologies.

These scenarios in the medical area indicate that there are applications where the use of BCT could be considered and there may be other scenarios with sensitive data or with tightly controlled workflows between partners where BCT could play an excellent role, in particular, since it provides a bundle of ready-made solutions for institutions. On the other hand, it makes sense to carefully analyse the tasks and look for the most suitable solution. In general, the concept of "Open Science" gets increasing support in the research domain where research objects such as data, metadata, software, etc. are openly shared in addition to the publications, i.e. a flexible and open domain is being aimed at where it will be good research practice to refer to other people objects. BCT in contrary requires a closed approach to benefit from its promises.

6.2 Comprehensive use case from biomedical research

As we have mentioned before, there could be applications for using BCT in science, but the application of the technology should be analyzed deeper to verify that the balance between its benefits and disadvantages is appropriate. Another issue we would like to raise is that most of the current discussions related to BCT are focused on static events, but not to chains of operations performed on an object, especially in a distributed environment. In this section, we would like to discuss that there are such scenarios and that usage of BCT should be examined more it this context.

[10] We refer to the discussion paper from Kuchinke and Wittenburg for more elaborations.

Let's focus on biomedical research. Speaking of a chain of operations, development of a drug can be seen as such a chain [27]. The chain begins in preclinical research, where a drug impact is very deeply analyzed before it is tested on humans[11]. The results of preclinical research must show that the drug has desired effect on illness, that it is safe, etc. in vitro and in vivo on relevant animal species. Only after demonstration of all desired characteristics of the drug it is possible to move forward to clinical tests on humans and only after successful clinical tests it is possible to deploy the drug in healthcare. The whole process includes operations such as physical operations on biological material (acquisition of primary material, processing, transport, storage, retrieval from storage, etc.), biological material processing and digital data generation (genetic information sequencing, metabolite analyses using spectrometry or chromatography methods, various types of clinical imaging, etc.), digital data processing and new data generation (e.g., scientific workflows). We are speaking of the chain of operations because all actions and decisions made are highly dependent on previous steps. The operations are also very highly regulated by national and international laws and standards (e.g., HIPAA or GDPR). Because of involved pressures due to vested interests of individuals as well as businesses (e.g., decision on whether and how much the treatment is reimbursed from public health insurance), the whole process needs to be auditable by various third parties, which may not be known at the time when the process occurs, or even by general public in the future (while also preserving privacy of subjects donating their data and/or samples to the research). It is also important to point out that the related environment is distributed and extremely heterogeneous because it consists of clinical laboratories, hospitals, research laboratories, biobanks, regulatory bodies, etc.

Another important problem related to auditability is the lack of reproducibility in the medical research, which is a critical problem known for more than a decade now. The ability to reproduce research is naturally multi-lateral problem, where ideally anybody is able to verify study procedures and scientific results. In addition to auditability, reproducibility aims to enable independently compute results of research and thus lays down another requirement for data integrity, authenticity and reliability. There are several papers addressing the problems with reproducibility of research and show, that based on the definition of the term "reproducibility", 18 to 89 percent of research is not reproducible, which has ethical and enormous financial implications [3, 27, 28, 29]. The reasons for a lack of reproducibility range from insufficiently defined biological reagents and reference materials, poorly documented study designs, problems in data analysis and reporting, and in laboratory protocols [27].

There are several areas which should be addressed to solve the reproducibility issues. Fact is that at the end of the day, we will need a technology to implement a satisfiable solution. Since the BCT offers means to reach many desired properties, it can make sense to investigate its applicability deeper and to counterweight it to "standard" technical solutions. To be more specific, BCT enables us to achieve integrity, authenticity and nonrepudiation; to verify validity and to establish level of reliability of a content through a consensus. It works in a distributed environment and all actions performed on BCT are transparent and auditable. To learn more about current research dealing with BCT in biomedical domain we refer to a scoping review [30]. But as described before, it needs to be shown whether BCT brings advantages compared to traditional technologies.

7. Conclusions

In this paper we discussed the usefulness of blockchain technology in its two flavours, permissionless and permissioned, for possible applications in the research domain. On purpose, we did not discuss its relevance for commercial and governmental applications. There is no doubt that new technologies

[11] This resembles the example of the tight control of the supply chain of milk powder bins mentioned earlier.

are needed to address the two basic motivations mentioned in the introduction. However, most of the arguments in favour of using BCT one can find in publications cover statements such as

"BCT is meant for transactions of ownership of real assets."
"BCT can help to document a permanent trace of value in the digital age."

Halunen [31] is critical even in these respects arguing that BCT seldom adds much value compared to existing technologies while Valivaara [32] expects many BCT applications just being there in future without being noticed by the user to control and document the many (small) transactions. A systematic overview [33] of BCT application comes to a critical conclusion too: "*However, while many try to propose blockchains as a panacea and an alternative to databases, this is far from true. As already discussed, there are many scenarios where traditional databases should be used instead.*". But we will focus on the research domain where Lannom stated that BCT does not address the major issues we are currently faced with [34].

As Koureas pointed out [35] we need to redefine "trust" as a major challenge in the research domain to come to a culture of data sharing. Data is being created locally, but is going to be used globally. BCT does not address this dimension of trust. The dimension it addresses is "security of the documentation of transactions". The other big challenge in research is to increase FAIRness [36] and Openness of data which is focusing on a broad and efficient reuse of data and services. Also this dimension is not addressed by BCT directly.

In the research domain we have areas where highly sensitive data is being exchanged such as in the medical sector which requires secure technologies and where the regulations require detailed documentation. Since this area has direct impact on human lives, we believe that this might be an area where BCT could play a role and where a combination of FAIR DOs and BCT could make sense. Most probably permissioned BCT would be chosen, but the closeness of P-BCT to other existing technologies should motivate projects to carefully look at details. It is the ready-made bundling of a number of technologies in BCT which could make BCT attractive for projects.

9. References

[1] M. Wilkinson et al.: FAIR Principles; https://www.nature.com/articles/sdata201618
[2] G. Strawn; Open Science, Business Analytics, and FAIR Digital Objects; http://doi.org/10.23728/b2share.6ceeed13eb6340fcb132bcb5b5e3d69a
[3] Petr Holub, Florian Kohlmayer, Fabian Prasser, Michaela Th. Mayrhofer, Irene Schlünder, Gillian M. Martin, Sara Casati, Lefteris Koumakis, Andrea Wutte, Łukasz Kozera, Dominik Strapagiel, Gabriele Anton, Gianluigi Zanetti, Osman Ugur Sezerman, Maimuna Mendy, Dalibor Valík, Marialuisa Lavitrano, Georges Dagher, Kurt Zatloukal, GertJan B. van Ommen, and Jan-Eric Litton.Biopreservation and Biobanking.Apr 2018.ahead of printhttp://doi.org/10.1089/bio.2017.0110
[4] D. Yaga, P. Mell, N. Roby, K.Scarfone: Blockchain Technology Overview, NISTIR 8202; https://doi.org/10.6028/NIST.IR.8202
[5] Bitcoin; https://de.wikipedia.org/wiki/Bitcoin
[6] Bitcoin Enrgy Consumption: https://www.theverge.com/2019/7/4/20682109/bitcoin-energy-consumption-annual-calculation-cambridge-index-cbeci-country-comparison
[7] P. Nikander: Blockchains and IoT: a reality check; IoTWeek Bilbao; https://github.com/GEDE-RDA-Europe/GEDE/tree/master/Blockchains
[8] E. Andoulaki, M. Jarke, J-J. Quisquater: Introduction to the special Theme Blockchain Engineering; https://ercim-news.ercim.eu/en110/special/introduction-to-the-special-theme-blockchain-engineering
[9] CoreTrustSeal: https://www.coretrustseal.org/
[10] P. Lyons, R. Kahn: Blocks as digital entities: A Standards perspective, CNRI; https://doi.org/10.3233/ISU-180021

[11] Kahn & Wilensky, 1995: http://www.cnri.reston.va.us/k-w.html
[12] Kahn & Wilensky, 2006: https://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf
[13] RDA DFT Core Terms and Model; http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318
[14] Schultes & Wittenburg, DAMDID Paper: https://github.com/GEDE-RDA-Europe/GEDE/blob/master/Digital-Objects/Papers(finished)/FAIR-DO-Relation/damdid2018_paper_schultes_wbg-final.pdf
[15] DOIP: https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf
[16] EC FAIR Exp Group Report: https://doi.org/10.2777/1524
[17] Handle System: https://en.wikipedia.org/wiki/Handle_System
[18] DOI: https://de.wikipedia.org/wiki/Digital_Object_Identifier
[19] EIDR: https://en.wikipedia.org/wiki/EIDR
[20] ePIC: https://www.pidconsortium.eu/
[21] X1255: https://www.itu.int/rec/T-REC-X.1255-201309-I
[22] RDA DTR: https://rd-alliance.org/group/data-type-registries-wg/outcomes/data-type-registries
[23] CORDRA: https://cordra.org/
[24] Ethereum: https://de.wikipedia.org/wiki/Ethereum
[25] Slavicek, K., Dostal, O., Javornik, M. and Drdla, M., 2010, January. MEDIMED-Regional Centre for Medicine Image Data Processing. In *2010 Third International Conference on Knowledge Discovery and Data Mining* (pp. 310-313). IEEE.
[26] W. Kuchinke, P. Wittenburg: Blockchain and Data; http://doi.org/10.23728/b2share.09a3b784761244819a0c2dc390ff9c7a
[27] Freedman LP, Cockburn IM, Simcoe TS (2015) The Economics of Reproducibility in Preclinical Research. PLoS Biol 13(6): e1002165. https://doi.org/10.1371/journal.pbio.1002165
[28] Begley CG, Ioannidis JPA. Reproducibility in science: Improving the standard for basic and preclinical research. Circulation Research. 2015. pp. 116–126. 10.1161/CIRCRESAHA.114.303819
[29] Freedman LP, Inglese J. The increasing urgency for standards in basic biologic research. *Cancer Res*. 2014;74(15):4024–4029. doi:10.1158/0008-5472.CAN-14-0925
[30] DROSATOS, George a Eleni KALDOUDI. Blockchain Applications in the Biomedical Domain: A Scoping Review. Computational and Structural Biotechnology Journal [online]. 2019 [cit. 2019-06-07]. DOI: 10.1016/j.csbj.2019.01.010. ISSN 20010370.
[31] K. Halunen: Using Blockchains to incrase trust in AI; https://vttblog.com/2018/02/19/using-blockchains-to-increase-trust-in-ai/
[32] V. Vallivaara: Can Blockchains help Changing the Data Culture?; IoTWeek Bilbao; https://github.com/GEDE-RDA-Europe/GEDE/tree/master/Blockchains
[33] CASINO, Fran, Thomas k. DASAKLIS a Constantinos PATSAKIS. A systematic literature review of blockchain-based applications: Current status, classification and open issues. *Telematics and Informatics* [online]. 2019,**36**, 55-81 [cit. 2019-06-07]. DOI: 10.1016/j.tele.2018.11.006. ISSN 07365853.
[34] L. Lannom: commentary to an early version of this paper
[35] D. Koureas, Digital Object Cloud for linking natural science collections information; https://github.com/GEDE-RDA-Europe/GEDE/blob/master/Digital-Objects/DO-Workshops/Workshop-Philadelphia-2019/koureas-do-p13.pdf
[36] M. Wilkinson: https://www.nature.com/articles/sdata2018118?draft=collection