Fostering Open Data Using Blockchain Technology

Simon Tschirner[1(✉)], Mathias Röper[1], Katharina Zeuch[1], Markus M. Becker[2], Laura Vilardell Scholten[2], and Volker Skwarek[1]

[1] University of Applied Sciences Hamburg, Ulmenliet 20, 21033 Hamburg, Germany
{simon.tschirner,mathias.roper,katharina.zeuch,
volker.skwarek}@haw-hamburg.de
[2] Leibniz Institute for Plasma Science and Technology, Felix-Hausdorff-Str. 2,
17489 Greifswald, Germany
markus.becker@inp-greifswald.de
http://www.haw-hamburg.de, https://www.leibniz-inp.de

Abstract. While open science is growing in popularity and especially publishing as open access is common and has proven its success in today's research, sharing of research data in terms of open data is still lacking behind. INPTDAT provides a platform to share research data in the field of plasma technology. In the course of this, the project QPTDat aims to increase the incentives to publish, share, and reuse research data, following FAIR principles and fostering the idea of open data. QPTDat identified the following main success factors: Authors need a secure proof of authorship to guarantee that they are credited for their work; a proof of data integrity, to ensure that reused data has not been modified or fabricated; a convincing system for quality curation, to ensure high quality of published data and metadata; and comprehensive reputation management, to give an additional incentive to share research data. This paper discusses these requirements in detail, presents use cases and concepts for their implementation using blockchain technology and finally draws a conclusion regarding utilisation of blockchain technology in the context of open data, summarising the findings in form of a research agenda.

Keywords: Open data · Open science · Research data reuse · Blockchain technology

1 Introduction

In research fields such as plasma technology, data-driven science relies on research data that is findable, accessible, interoperable and reusable [6]. *FAIR* research data, as defined by [29], increase research data reusability [6] and is a first step towards avoiding a phenomenon that is called a reproducibility crisis [1].

The work was funded by the Federal Ministry of Education and Research (BMBF) under the grant marks 16QK03A and 16QK03C. The responsibility for the content of this publication lies with the authors.

E. Bisset Álvarez (Ed.): DIONE 2021, LNICST 378, pp. 209–228, 2021.
https://doi.org/10.1007/978-3-030-77417-2_16

The interdisciplinary data platform for plasma technology INPTDAT[1] aims to bring the *FAIR* principles to research in plasma technology. It provides data publications with a unique digital object identifier (DOI), a plasma source catalogue, faceted search options and API based access to (meta)data [2].

This paper proposes an architecture extending INPTDAT to a blockchain-supported open science platform for plasma technology, adding aspects as proof of authorship, data integrity, quality curation, and reputation management. The presented solution initially focuses on plasma technology. However, the principles are described in a sufficiently abstract way to be easily adaptable to further scientific disciplines.

Open science is a term comprising several aspects – mainly open access, open data, licensing, uniqueness and citation tracking (cf. [25]) – with the aim of opening up scientific structures to make research available to a broader audience [20, p. 9]. Open access usually means publication of research articles to be read free of charge, while open data refers to the availability of data produced during the research process. Open access has now been around for about two decades, becoming quite successful (cf. [21]). Publishing open access has several advantages for researchers, e.g. higher visibility and increased number of citations (cf. [17]).

Open data specifically aims to foster the research process by increasing the number of published research data sets in general and an earlier publication of data during the research process. Additionally, the transparent publication of original data intends to reduce publications with tampered data. However, despite many programs and efforts supporting open science, the success of motivating researchers to share their raw data is still limited, especially when addressing early publication of data in the research process.

As stated by [3], blockchain technology offers new possibilities to open science. Providing de-centrality and information distribution, all participants are aware of data and actions on the blockchain [11, p. 546]. This way, it equips open science systems with transparency and independence from a single ruling authority (cf. [23, p. 8]). Furthermore, blockchains use cryptographic functions to store and chain data. These functions ensure that the data have not been manipulated, changed or dismissed, but maintains integrity. It is even possible to use a blockchain without making data instantly public (cf. [3, p. 14]), but it still allows the author to proof data integrity when the data is published later on.

This article provides a further research agenda towards a comprehensive blockchain-based system for open science. On the path, it evaluates the feasibility of the sketched approach by providing a system architecture in Sect. 2, use cases as starting points for implementation in Sect. 3, and challenges in Sect. 4.

[1] https://www.inptdat.de – latest access: February 6, 2021.

2 Architecture

The proposed system aims to implement the ideas of open access, open data, identification and citation tracking. The main focus is to design a system that motivates researchers to engage in open data and research data sharing.

In this section, the main goal will be broken down into a few main requirements, giving further motivation to explore the blockchain-based approach. Besides, it examines the current applications of blockchain in science. The suggested architecture combines conventional systems for research data management and web databases for accessing data sets with the proposed blockchain-based system. It provides the basis for an open science system implementing the requirements.

2.1 Requirements

The following requirements result from a literature review, focusing on open science, open data and their combination with blockchain technology. Discussions and workshops with researchers experienced in open science or from the domain of plasma technology have validated and further refined these requirements.

The first requirement, aimed at encouraging the sharing of research data early in the research process, is (1) ensuring authorship. Researchers are more likely to share their data if they can prove authorship and when they are assured to be associated with their data. Related to this point is the requirement that (2) different access rights have to be available. E.g. data from a collaboration with a commercial partner might need to remain private for a certain period or permanently.

Several requirements are needed to guarantee a quality standard: On one hand, (3) data integrity has to be ensured, meaning that data did not alter in between its creation until its reuse. This covers the intentional manipulation of data as well as unintentional changes. Important is that researchers can be sure that data, when reused or cited in a future research item, has been in the same form as at the time of recording. If a platform takes measures to ensure data integrity, it is more trustable – (4) this includes critical metadata, e.g. those needed for reproducibility. On the other hand, (5) published research items themselves should meet defined quality standards. In this way, users of the platform can trust the published content, as they can in a peer-reviewed journal with a good reputation.

To function as a base for researchers to explore and reuse existing data, (7) the system needs to follow the FAIR principles. One requirement on open data is to (8) allow for proper citation of research data. Therefore, the unique identification of research data is needed, which allows citation tracking (cf. [25]).

A system needs to give its users the right incentives to add their high-quality content. Following the tradition of financial applications based on blockchain technology, some researchers suggest establishing an alternative way to fund and merchandise research (e.g. [15]). The approach presented here sees reputation as the main incentive for researchers. Work is submitted to conferences and

journals to share knowledge, make an impact and eventually to increase reputation. Therefore, (6) the system should add another mean to generate impact and gain reputation.

Finally, there is an implicit requirement, leading back to the motivation to use blockchain technology. Its mechanics provide transparency and security. The system should provide proof of authorship, guarantee data integrity and quality and monitor reputation. All these points need a secure, manipulation-proof implementation.

The requirements in summary:

1. Ensure authorship
2. Facilitate different access rights
3. Enable the validation of data integrity
4. Guarantee the integrity of significant information (metadata)
5. Ensure basic quality of data and metadata
6. Implement a robust reputation system
7. Published items must be searchable, accessible, interoperable and reusable (FAIR)
8. Research data must be citable

2.2 Related Work

Leible et al. [16] give a systematic review of the general suitability of blockchain technology for the field of open science. They conclude that open science can benefit from various aspects blockchain technology offers, especially from its tamper-proof recording of transactions and its decentralised nature, introducing a new level of trust. Nevertheless, the review states to consider a blockchain as one component of many. All of these have to be combined in a meaningful manner to create successful solutions fostering open science.

Tennant et al. [26] conducted an extensive multi-disciplinary study about innovations in the peer review domain. They see blockchain technology suitable as an enabler for new approaches, such as new incentive systems and authentication/certification methods for research data to prevent fraud and protect authorship.

Specifically dedicated to the area of science is the Bloxberg-blockchain[2] that has been developed by Max Planck Digital Library in 2019. It offers smart contracts for certification and verification of research data, governance and voting and consensus mechanisms.

Current approaches to blockchain-based open science systems mostly target subsets of these requirements or have a commercial background.

CryptSubmit [9] approaches the issue in research processes of not having persistent evidence of data existence. Therefore, it creates this evidence using trusted timestamps on the Bitcoin blockchain.

Focusing on the scientific publishing process, a blockchain-based system including document submission, review and publication, including a reputation

[2] www.bloxberg.org – latest access: November 6, 2020.

management has been proposed by [27]. Pluto offers a decentralised research network for publishing research data in general. It includes its verification, peer review process and a token-based reputation system [19].

Frankl [7] presents a commercial open science platform specialised in cognitive assessments. It offers blockchain-based data management and an app-based data sharing marketplace, using a token-based incentivisation mechanism.

ARTiFACTS[3] is a commercial blockchain-based platform to create a proof-of-existence of research data. Also included are the possibilities to get citations on work in progress and linkage of corresponding data via metadata. ResearchHub proposed the ResearchCoin[4] to represent the scientific reputation of an individual. It resembles a currency, earned by community votes.

2.3 Related Concepts

For those new to the topics, this section briefly introduces two key concepts used throughout this paper: blockchain technology and hash values of data.

Blockchain Technology combines cryptography, data management, peer-to-peer (P2P) networking and consensus mechanisms. It creates a trusted environment for the verification, execution and recording of transactions between parties. In the context of "cryptocurrencies", transactions mainly are of financial nature, while in other contexts, they can be used to add information to the blockchain. One way to look at a blockchain is to regard it as a ledger with its content organised in blocks, following an append-only logic. It is distributed and synchronised among the network peers (called nodes). Cryptography – especially cryptographic hashes – ensures the ledger's integrity. Some blockchains allow deploying code, which's execution can be triggered by transactions. Executable code deployed to a blockchain is called smart contract. Smart contracts can implement functions and data structures on the blockchain and enable the creation of decentralised applications [31, p. 3 ff.].

To summarise, the main properties of blockchain technology in the context of this paper are: (1) they work append-only, meaning that information once stored on the blockchain, can usually not be altered or removed later on, (2) they are handled by a P2P network, meaning that no centralised authority is needed or even desired, and (3) smart contracts add additional logic, leading to decentralised applications.

Hash Functions are mathematical functions taking input data of any length and produce an output of fixed size. This output is called a hash or hash value, comparable to a fingerprint of the input data. Hash functions are one-way functions and always generate the same hash out of the same input data.

[3] www.artifacts.ai – latest access: November 6, 2020.

[4] www.researchhub.com/paper/819400/the-researchcoin-whitepaper – latest access: November 6, 2020.

Hash values are of a certain length, while the input data is usually not limited. Thus, in theory, different input data could create the same hash value. But since hash functions aim for an even distribution of all possible inputs among the possible hash values, a slight change to the input results in an entirely different hash value. Therefore, the chance that two inputs lead to the same hash (this is called a collision) is very low [12, p. 246 f.]. Also, it is practically impossible to find a corresponding input which leads to a specific hash value.

2.4 Proposed Architecture

The proposed architecture is an open science system with a blockchain backbone. It lies in the nature of blockchain-systems, that storage is expensive, due to its redundancy. Therefore, blockchain-based systems strive to minimise data stored on-chain, but to store most of the data off-chain. Usually, integrity of off-chain data is secured by storing a hash of the data on-chain. This principle is also used by the proposed system architecture (cf. Fig. 1) that handles data in three different categories and stores them accordingly in different places: (1) research data, stored on the researcher's personal computer or affiliations research data management system (RDM), (2) metadata, stored on a traditional web platform (INPTDAT), and (3) secured data, meaning hash values of research data or metadata, stored on the blockchain (BC Network).

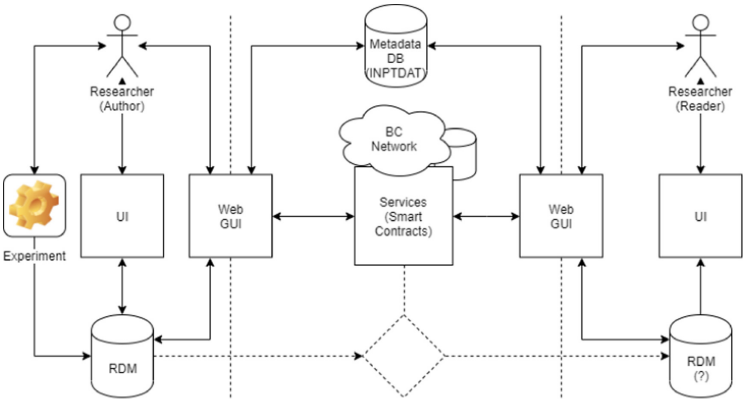

Fig. 1. The proposed architecture, combining a common metadata database and a blockchain-based solution

Six use cases have been identified to meet the main requirements. They will be deployed as services implemented by smart contracts on the architecture. The desired functionality will be shortly presented in the following, mapping the use cases to the architecture. In the subsequent section, the use-cases are discussed in-depth.

In the RDM, researchers can store their data, e.g. retrieved from an experiment. Although RDMs usually provide a user interface (UI), possibly a web interface, all three data storage should allow interaction via the same user interface (Web GUI). That way, users do not have to cope with several different systems. The BC network offers the service to certify research data stored in the RDM. Certification of data is then the basis for data verification by readers later on (cf. Sect. 3.1). From this moment, authorship and integrity of the research data can be proven; note that the data does not have to be made public immediately. This can happen later, by publishing the record on the web platform (INPTDAT).

Metadata databases allow fellow researchers and other interested persons to find interesting data sets. Therefore, researchers should add relevant metadata to the metadata database on research data publication (cf. Sect. 3.2). Next, an automated curation process (cf. Sect. 3.4) should assure the data quality. If necessary, this is complemented by a manual review process that can be supported by the blockchain-based system (cf. Sect. 3.5). Besides, other reused research data may receive additional reputation (cf. Sect. 3.6).

When readers identify interesting research data via the web interface, they can request a copy. Depending on the configuration or author's settings, a record can be copied directly from the author's RDM to the reader's RDM (or personal computer). Alternatively, the system prompts the author for the allowance of data sharing (cf. Sect. 3.3). After a successful transfer, data can be again checked for integrity, using a BC service and the data's reputation can get increased by a read.

3 Use Cases

Based on the general architecture, this section presents use cases meeting the requirements. The detailed description of the use cases helps to evaluate if the architecture supports them. The goal is to determine the degree, to which blockchain technology is useful for their implementation. Directions for further research are identified based on a discussion of the state-of-the-art and open challenges regarding each use case.

3.1 Certification and Verification

The central use case for applying a blockchain-based solution is the certification of research data. The aim is to certify that a specific researcher or a group of researchers has possessed specific research results at that time. If researchers certify their data directly after collection, this establishes authorship. The suggestion is to store authorship information on a blockchain. Thus, after certification, it will be almost impossible to manipulate authorship.

Additionally, the certified research item has to be stored. However, placing the research item itself on-chain is not feasible. First, the amount of data stored on the blockchain needs to be as small as possible. Second, information stored

on a blockchain typically is publicly available. Whilst, somewhat contradicting the main idea of open data, sharing research data might not be desired under certain conditions, e.g. very early on in the research process, before a proper analysis. Therefore, usually, only the hash of the data is stored on-chain. Storing a hash gives a sufficient compromise of a minimal amount of stored data and security. Figure 2 shows this use case.

Researchers can proof their authorship by providing the research item and the blockchain address where its hash-value was stored during certification. The verification function compares the given information. It also confirms the exact date of certification, because every stored block includes a verified timestamp. The verification function further ensures data integrity: If research data got modified after certification, verification would fail, since the hash-value of the presented research data would not match the stored one. Ensuring data integrity fosters research data sharing and reuse, as this avoids future investigations being carried out based on altered or incorrect reference data.

At its core, this use case has been implemented many times before. Several services exist, where users can prove authorship on certain data, e.g. by adding a hash to a Bitcoin transaction (see e.g. OriginStamp [10]). This use case is also the first use case implemented by Bloxberg (cf. [13]).

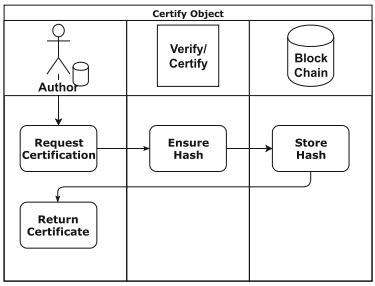

Fig. 2. When a user requests certification of a research item, the verify/certify service is responsible for creating a trustworthy hash value of the research item. The hash value will be stored tamper-protected on the blockchain.

Seeming relatively simple, a couple of design decisions imply further challenges. One challenge is related to authorship; a research item can have one or several authors, whose identities should be validatable. At the moment, in Bloxberg, authorship is just a text string that is hashed and stored together with the hash-value of the research item. This simple solution has some disadvantages. First, it is not possible to evaluate the provided identity. Thus it is not possible to prevent someone from adding information under a false identity. Second, it is not easily possible to include a unique identity. Third, if authorship would be stored trackable on the blockchain, this can lead to issues related to the General Data Protection Regulation (GDPR; for GDPR-issues regarding blockchain technology, see e.g. [14]). These issues – identity and GDPR-conformity – are research

objectives on their own and placed on the research agenda developed throughout this paper. For now, the recommendation is to use an external identity-provider that delivers a solution using a GDPR-compliant structure, see [30] for such.

Another challenge is tamper-resistance of the hashing mechanism. It must be infeasible to generate data to match a certain hash-value. Otherwise, dishonest researchers could add any hash-value to the blockchain and later fabricate data around a core data-set. The created data would then result in the same hash value stored on the blockchain and would, in consequence, be verified by the system, even if it is not the data that has originally been certified. Until today, SHA-256 is, correctly implemented (see e.g. [8]), seen as sufficiently tamper-proof. Still, technical progress might imply that hash-algorithms that are seen as secure today might not be secure in a few years from now (cf. [18]). A sustainable blockchain solution might require the implementation of mechanisms that allow for a later exchange of the used hash-algorithm.

Additionally, research items processed by the hash-algorithm have to be kept private but still hashed with the correct algorithm. Privacy is hard to guarantee if data has to be sent to the entire blockchain to get hashed. In turn, it is hard to secure the intended execution of the hash-algorithm, if it is only executed locally on the researcher's system. An interesting but payment-oriented approach is the usage of off-chain state channels as implemented by Perun [4].

3.2 Publication Process

The second use case covers the step of publication (see Fig. 3). It could be publishing research data that has already been certified but otherwise kept private or the publication of research articles. When a (data) publication is accepted, it will be published via the conventional metadata database (INPTDAT) and becomes publicly available for search on the web.

When authors request publication, the proposed platform will forward the request to the quality curator described in Sect. 3.4 (UC4). When the quality control passed successfully, the research item is ready for publication. Metadata usually supports the possibility to reuse data. In plasma science, e.g., metadata should document the experimental setup and environmental conditions under which the data has been collected. This information is critical for reproducibility and to relate different pieces of work to each other. This importance implies that essential metadata should be immutable afterwards. Therefore, on publication, even metadata will be secured by a hash-value written to the blockchain.

The last step projects the classical citation of research articles to research items. I.e. other research items that have been influential for the item to be published are gaining reputation. Reputation is handled by the use case (UC5) in Sect. 3.6.

This use case is straight forward and does not add significant tasks to the proposed research agenda. A conventional web platform will implement the main functionality. The contribution of blockchain technology in this use case lies in the additional protection of critical metadata.

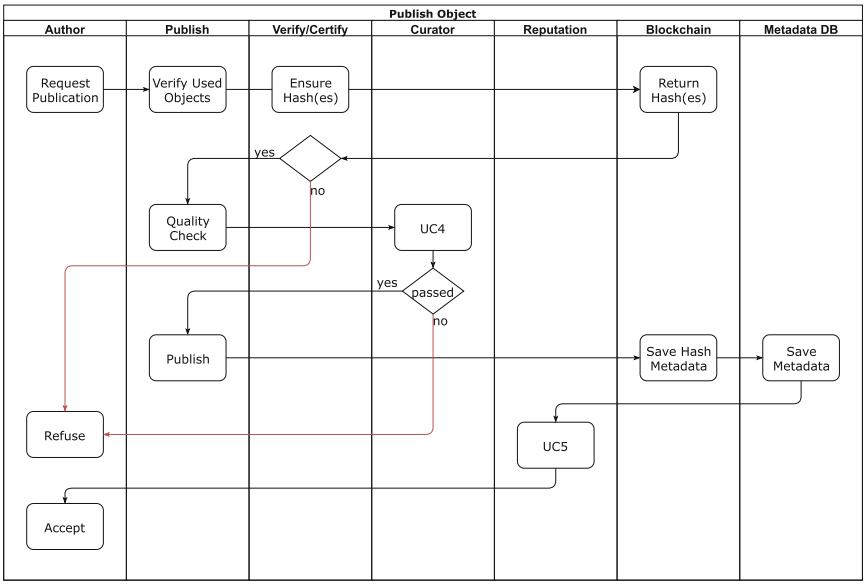

Fig. 3. For publication, (1) validation of all relevant certificates (of cited or reused work), (2) ensuring the quality of the research item, and (3) incrementation of the related reputation counters (e.g. of cited work).

3.3 Access

A published research item (see Fig. 4) will be findable and (indirectly) accessible via the conventional web database. The author can give access restrictions to an item. It can either be publicly available, i.e. accessible without any constraints, or available on request, meaning that a reader needs to request the item from its author.

The system can directly resolve a link to the location of publicly available items (e.g. on the author's RDM); it is immediately possible to create a copy. Otherwise, the author receives a request, e.g. in the web UI. If he or she grants access, the requesting researcher gets a unique, one-time link for download. In both cases, after completion, the verification smart contract validates the copy. It will only be possible to access the copy after successful validation. Via the reputation-service, the author receives a virtual incentive for his or her read reputation. If verification fails, the system removes the copy and informs the author and reader. Possible issues are that the item got corrupted during transfer, that data integrity got violated, or simply that the host is off-line. The author will have the opportunity to resolve the situation, e.g. by providing the item again. If access fails continuously, this can have an impact on the item's reputation. It has to be validated manually. Ultimately, it might be de-listed.

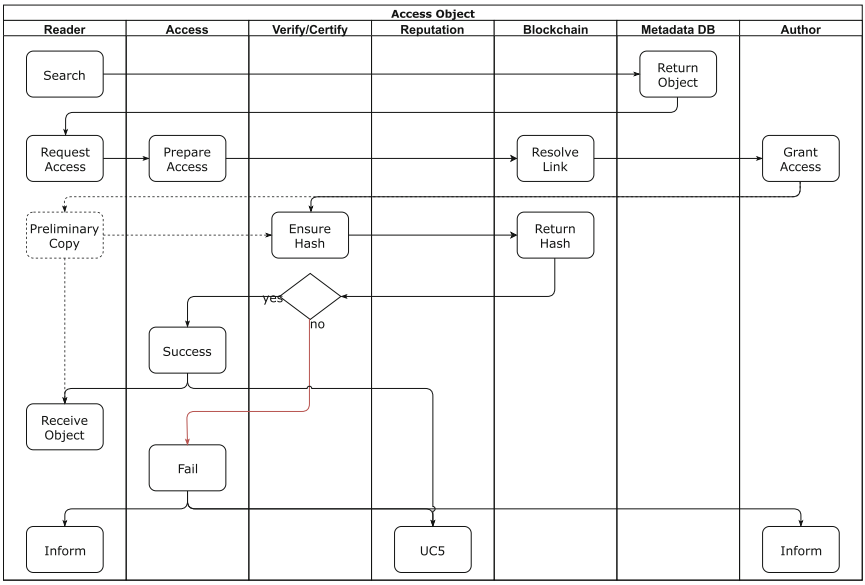

Fig. 4. The steps to resolve access to a research item.

As for the previous use case, access is a relatively straight forward implementation, that mostly depends on other use cases (namely verification and reputation) or the conventional metadata database INPTDAT.

3.4 Curation

The quality of available research items is a crucial success factor for a platform providing research data. If data is inconsistent or the process of its creation not well-documented, it is not reproducible and its reuse value low. Including such items on a platform will be frustrating for its users. They would have to skip through a couple of data sets to identify the content of reasonable quality. As the aim is to establish a platform with a strong reputation, positioning it as a real alternative to commercial websites, to guarantee a certain level of data quality is a primary requirement. Traditionally, a review process involving human reviewers is the curation process of choice. While this process has its clear advantages, it is usually lengthy and tedious. The presented approach aims to support this process with automation, if possible.

This paper suggests an algorithmic analysis based on certain key-information as metadata. Realistically, this can only result in a pre-check of research publications, e.g. monitoring the bibliography. When the object to review is data, already available in a machine-readable form, the premises for (partly) automation are positive. The next section covers classical peer-review.

The automatic curation (cf. Fig. 5) is initiated by the publication use case shown in Sect. 3.2. At this point, the integrity of the current data set is already

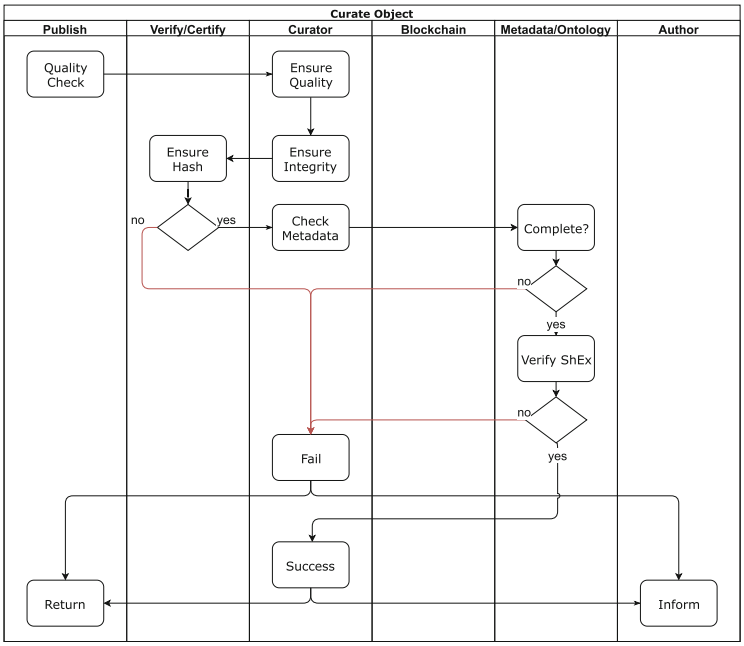

Fig. 5. The suggested process to execute automatic data curation.

confirmed (cf. Fig. 3). Now, the system cycles through (cited) data sets that have influenced the current research, to confirm their integrity, too. To do so, the reused data sets and related blockchain transactions (those data sets' certificates), have to be submitted. If this check succeeds, the curator continues to check the metadata of the current data set.

In general, metadata gives further context to a data set and thus, e.g., making it easier to find and interpret. Available criteria for metadata differ a lot in different disciplines. A distinctive example exists within the COMBINE community, dealing with standardisation and modelling in clinical and biomedical domains [28]. Here, comprehensive standards exist, e.g. to represent models used for data generation and processing. Given such a basis, it would eventually be possible to analyse models and data automatically.

Such achievements are usually results of decades of collaboration. It is one aim of the research project "Quality assurance and linking of research data in plasma technology – QPTDat", to foster the process of generating a joint, comprehensive metadata schema for plasma technology. Given such a schema would open up two possibilities of automated checks: (1) metadata can be checked for completeness, (2) metadata can be tested for soundness of the given metadata. The current approach suggested within QPTDat is to use shape expressions (ShEx) [24] for such a quality check.

Usage of blockchain technology to support automated quality curation of research data seems promising. Based on the certification and verification, it is possible to prove the integrity of reused data sets. It is furthermore possible to protect critical metadata from undesired alteration. Finally, the system allows logging the status of a research item's quality check.

A challenge is to develop a system for automated quality curation in detail and its meaningful connection to the blockchain – due to computational complexity, it would not be possible to run the check on the chain, meaning that each node performs the calculation. A solution offering transparency and tamper-resistance for the automated quality check is still desirable. However, even a comprehensive metadata schema will not fully replace a review by a human expert. Therefore, consideration of mechanisms for human interaction is necessary.

3.5 Peer-Review

The automated quality curation described in the previous section is especially interesting for the sharing of research data. However, substantial peer-reviews of research articles are usually a strong advantage of traditional paper processing offered by conferences and journals, guaranteeing the quality of published research items. To publish articles, a peer-review is still necessary. Even though peer-review is not a primary concern in the presented research project, it is still briefly outlined here.

It is possible to trigger peer-review during the publication use case. The latter will be stalled, until a result of the peer-review is available. First, suitable reviewers, e.g. having a positive reputation and experience in the field, are selected. Advantageous is a track record of good quality reviews submitted within the desired time frame. Reviewers should also not have conflicts of interest with the item to review. The system then notifies the reviewers and waits for their acceptance. The reviewers then have time to react and finish their review. After an agreed time or if a reviewer resigns, it is possible to reassign. A sufficient number of reviews leads to a verdict like accept, reject or request for changes.

Designing the peer-review-process with blockchain technology needs a clear distinction between the parts of the process that require a blockchain architecture and those better supported by traditional technology. This distinction depends, e.g., on the philosophy, one follows regarding open peer review [22]. One could choose to design a system, where only critical data, such as review scores, are collected and stored on the blockchain, or a system where the reviews themselves are stored, or even one where the complete review process is modelled and managed via smart contracts. The Decentralized Science project is currently working on a conversion of the classical peer review process onto a blockchain. Preliminary results support the claim, that transparency and decentralisation provided by blockchain technology is an enabler for the shift towards open access [27].

3.6 Reputation Management

The last main requirement is a robust reputation management system. Citations of scientific articles and publication in journals of high reputation are still among – if not the – most important reputation factors for scientists. This leads to three implications: (1) Principle incentive for the researchers to use a blockchain-based system for open science will be to gain additional reputation. (2) Designing reputation management should probably centre around citations. And (3) a newly designed approach allows to include further factors to represent researchers' reputation, factors that often might be less visible, as in the case of research data sharing, the amount of data shared, quality and reuse of that data.

Finding a suitable, motivating, and fair new reputation index is a research question on its own and out of scope for this article. Instead, the following four factors are recommended as elementary values for a probable reputation index:

- Citations or number of reuses: How often research items of a researcher get reused and cited by others.
- Reads: How many times their research items get downloaded and read.
- Number of publications: The number of research items published.
- Number of reviews: The number of reviews given for research items of other scientists

An optional fifth factor could be the quality of the submitted/published research items. Possibly, the number of citations/reuses already indicates this factor.

The use case diagram depicted in Fig. 6 contains three independent tasks. First and upmost, the retrieval of reputation scores, which itself could be an array containing four values representing the factors mentioned above. The scores are stored verifiable on the blockchain and can be accessed, e.g. for inclusion in result lists on INPTDAT. The middle and bottom tasks are increments of the reputation, based on reads or publications. During publication, even the reputation score of the cited/reused research objects is increased.

An advantage is that a blockchain can store these values with high integrity and that their accumulation is stored transparently. However, this system will put a lot of stress on the tamper-resistance of the smart contracts that manage the reputation counters. Here the scepticism expressed by Leible et al. [16] has to be heard and accepted as a challenge.

The next challenge is to find the right representation of the reputation counters on the blockchain. Here, it will be particularly problematic to manage reputation related to the researchers' identity. If reputation scores are stored related to research items, these have to be connected to the correct author, to allow proper aggregation of reputation. If it is stored related to authors, its calculation and the items adding to it need to be transparent. Even ways to alter the score calculation at a later date have to be considered. Be it that errors or fraud have been revealed or for reasons connected to GDPR. It even has to be considered that attackers might generate arbitrary content to hide tampering with reputation indicators.

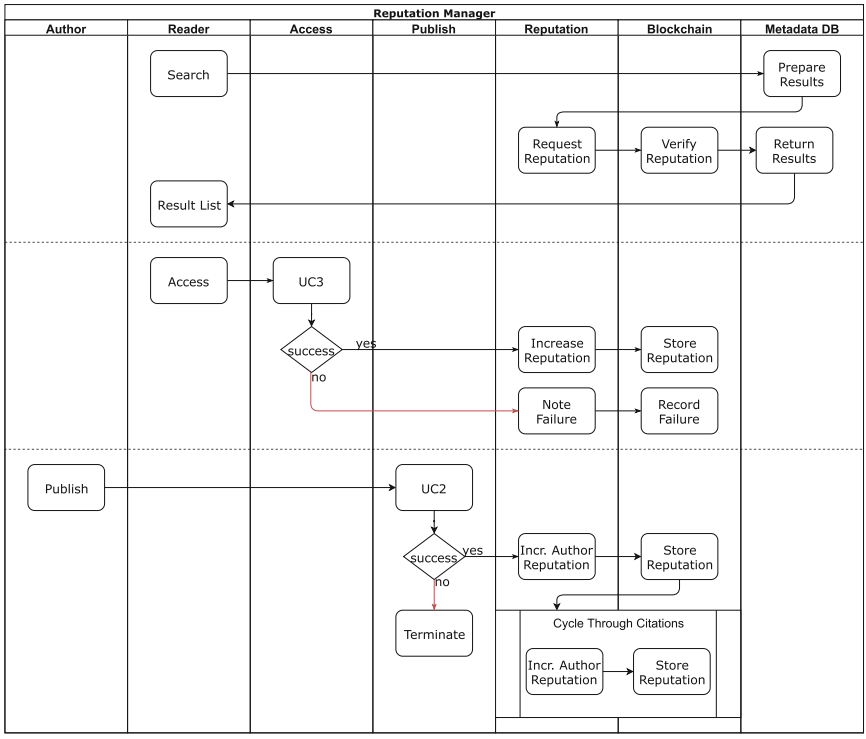

Fig. 6. The suggested process to execute automatic data curation.

4 Results

Compared to other work, describing possibilities to support science and publication of research results with blockchain technology in general, this work takes a step closer to its implementation. The purpose is to investigate the potential of blockchain technology to foster open data. Special attention was paid to the opportunity of increasing the attraction of sharing research data, especially prepublication. For this purpose, the main requirements, which resulted from the related QPTDat research project, have been listed. The main contribution of this paper lies in the proposed architecture and use cases. These show a starting point for the implementation of a blockchain system supporting open data. The implication is that the application of blockchain technology in research data sharing seems promising. The almost tamper-proof data structure and the transparent and decentralised nature of blockchain networks have the potential to replace common, centralised, and often commercial structures in the publication of research results in favour of an open science approach. However, a complete implementation is still pending, and additionally, quite some challenges still lie ahead. An overview of the latter is given in the subsequent section.

4.1 Research Agenda

Identity and GDPR are closely related. While knowing the researchers' identity is needed, regarding their research items and, to some extent, even reviews and reads, GDPR requires that identities and all related personal information can be removed on request. Authorship of a research item is such personal information. In this case, the nature of blockchain technology making it almost impossible to remove or alter information once added to the blockchain, which is one of the main reasons why this technology is feasible for proof of authorship and integrity of research data, makes it harder to comply with GDPR.

A thorough discussion of the relation between blockchain and GDPR can be found in [5]. One mentioned solution is to store the personal data off-chain, where it could be deleted on request. Such external identity management would only need to store a pseudonym on the blockchain. Editable off-chain storage would handle the relation between identity and pseudonym. A consequent requirement would be that each entry on the blockchain comes with a unique identifier so that removing authorship from one research item would not revoke authorship of other research items.

Even if there are already some solutions available, this topic is complicated. Solving the issue of identity and GDPR-conformity might impact other advantages of blockchain technology. The connection to reputation management seems particularly problematic.

Hashing is the central underlying mechanism to prove the authorship and integrity of research items via blockchain. There are two main issues. (1) During implementation, it has to be made sure that the hash algorithm is exchangeable in the future, in case the used algorithm becomes insecure. This would eventually require a complete re-hashing of all certified research data, which is a challenge by itself. Due to its computational complexity, (2) the hash function cannot be executed on the (complete) blockchain network (by using a smart contract). However, this is not even desired, as that would mean that the research data cannot be kept private (which is a requirement, e.g. for sensitive data prepublications). However, performing the hashing off-chain could increase the risk of manipulated hash values entering the system, which eventually could cause problems.

Another open question is how to perform complete integrity checks for research items, including the integrity checks of reused data. First of all, a basic technical solution would require an author to provide all reused data sets. Additional details, e.g. certificates, would be needed to verify integrity via the blockchain automatically. Specific challenges lie in the confirmation that data has been reused soundly. E.g. that the considered data sample is representative of the whole data set. This might currently be a limitation of automatic quality curation, requiring human experts as reviewers.

Details of Reputation Management. Regarding the reputation system, basic general questions are: which data structure should be used to store reputation, how should it be managed on the blockchain and (how) should a reputation index be calculated.

Security is a central topic that is part of almost all items on the research agenda. Data structures and smart contracts will have to be developed and tested very carefully to prevent tampering with the system. The proof of authorship using hash values is a relatively well-explored field and thus security risks are easier to avoid.

However, the situation regarding reputation management is quite the opposite. Reputation has been identified as one of the key incentives to use the proposed open science system. It must be impossible to manipulate reputation scores. Therefore, very cautious development and testing of the related data structures and smart contracts is mandatory.

Possible scenarios might be: attackers taking over reputation scores which do not belong to them (a question related to the identity management), reputation not being appropriately registered, faked research items or identities could be used to gain additional reputation, to name a few.

Others. Two additional topics should be part of the research agenda: The detailed design of the used blockchain itself and challenges in the area of peer-review (as far as not included in the previous topics). However, this paper does not discuss these topics. Instead, the usage of Bloxberg as blockchain is recommended. Bloxberg provides basic functionality as well as a strong community and network. With its scientific background, it is for the time being an excellent choice for the proposed system. The Bloxberg consortium handles challenges related to the blockchain infrastructure. The Decentralized Science project, mentioned earlier, covers the field of peer-review well.

4.2 Success Factors

Even a perfect technical solution cannot guarantee adoption by its potential users. Without adoption, the main incentive, gaining reputation from the publication of research data, will be non-existent. Resembling the chicken or the egg dilemma, this leads to the last two main requirements for the system to be developed. (1) It has to be easy to use. In the best case, the system integrates seamlessly with the researcher's workflow. QPTDat aims to add the functionality directly in a solution for research data management. (2) The researchers need to see clear benefits of the solution. The aim has to be to design the final platform useful right away. Combining the most important of the before mentioned requirements: The system is easy to use, so researchers add their data. Added data is of high quality and as such easy to find and reuse. Ideally, the quality curator would help researchers to improve data quality. Finally, scientists will reuse the published research data and increase their reputation from their data publication – and, all of this is done on a transparent, decentralised platform, open to everyone.

5 Summary and Conclusion

This paper has presented requirements towards an open science platform aiming to foster open data in plasma technology. The proposed architecture shows the integration of such a platform with a blockchain structure. Several presented use cases depict possible solutions to the blockchain integration and outline further challenges. A summary of these challenges suggests the direction of future efforts in research and development. These are the challenges that have to be faced to finally fulfil the requirements and lead to an open science system that, with the help of blockchain technology, fosters sharing of research data even early on in the research process. The platform aims to increase willingness to share research data by giving researchers the security of authorship, lowering the threshold to reuse data, ensuring data quality and integrity, and eventually giving deserved reputation for researchers sharing their research data. A brief analysis of security issues implies that security of reputation management is a concern to be further considered.

References

1. Baker, M.: Is there a reproducibility crisis? Nature **533**(7604), 452–454 (2016). https://doi.org/10.1038/533452a. http://www.nature.com/articles/533452a
2. Becker, M.M., Paulet, L., Franke, S., O'Connell, D.: INPTDAT - a new data platform for plasma technology, October 2019. https://doi.org/10.5281/zenodo.3500283. https://doi.org/10.5281/zenodo.3500283
3. Bell, J., LaToza, T.D., Baldmitsi, F., Stavrou, A.: Advancing open science with version control and blockchains. In: 2017 IEEE/ACM 12th International Workshop on Software Engineering for Science (SE4Science), pp. 13–14. IEEE (2017). https://doi.org/10.1109/SE4Science.2017.11. http://ieeexplore.ieee.org/document/7964307/
4. Dziembowski, S., Eckey, L., Faust, S., Malinowski, D.: Perun: virtual payment hubs over cryptocurrencies. In: 2019 IEEE Symposium on Security and Privacy (SP), pp. 106–123, May 2019. https://doi.org/10.1109/SP.2019.00020. ISSN 2375-1207
5. Finck, M.: Blockchain and the General Data Protection Regulation: Can Distributed Ledgers be Squared with European Data Protection Law?: Study. European Parliament (2019)
6. Franke, S., Paulet, L., Schäfer, J., O'Connell, D., Becker, M.M.: Plasma-MDS, a metadata schema for plasma science with examples from plasmatechnology. Sci. Data **7**(1), 439 (2020). https://doi.org/10.1038/s41597-020-00771-0
7. Frankl: Frankl - An open science platform (2018). https://docsend.com/view/gn8t7k9. Library Catalog: docsend.com
8. Gilbert, H., Handschuh, H.: Security analysis of SHA-256 and sisters. In: Matsui, M., Zuccherato, R.J. (eds.) SAC 2003. LNCS, vol. 3006, pp. 175–193. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24654-1_13
9. Gipp, B., Breitinger, C., Meuschke, N., Beel, J.: CryptSubmit: introducing securely timestamped manuscript submission and peer review feedback using the blockchain. In: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 1–4. IEEE (2017). https://doi.org/10.1109/JCDL.2017.7991588. http://ieeexplore.ieee.org/document/7991588/

10. Hepp, T., Schoenhals, A., Gondek, C., Gipp, B.: OriginStamp: a blockchain-backed system for decentralized trusted timestamping. IT Inf. Technol. **60**(5–6), 273–281 (2018)
11. Janowicz, K., et al.: On the prospects of blockchain and distributed ledger technologies for open science and academic publishing. Semant. Web**9**(5), 545–555 (2018). https://doi.org/10.3233/SW-180322. https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-180322
12. Kizza, J.M.: Guide to Computer Network Security. CCN. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-55606-2
13. Kleinfercher, F., Vengadasalam, S., Lawton, J.: Bloxberg - the trusted research infrastructure [whitepaper]. Technical report, Max Planck Digital Library, February 2020. https://bloxberg.org/wp-content/uploads/2020/02/bloxberg_whitepaper_1.1.pdf
14. Kunde, E., et al.: Faktenpapier Blockchain und Datenschutz. Technical report, Bitkom e.V. (2017)
15. Lehner, E., Hunzeker, D., Ziegler, J.R.: Funding science with science: cryptocurrency and independent academic research funding. Ledger **2**, 65–76 (2017)
16. Leible, S., Schlager, S., Schubotz, M., Gipp, B.: A review on blockchain technology and blockchain projects fostering open science. Front. Blockchain **2**, 16 (2019). https://doi.org/10.3389/fbloc.2019.00016. https://www.frontiersin.org/article/10.3389/fbloc.2019.00016
17. McKiernan, E.C., et al..: How open science helps researchers succeed. eLife **5**, e16800 (2016). https://doi.org/10.7554/eLife.16800
18. Mosca, M.: Cybersecurity in an era with quantum computers: will we be ready? IEEE Secur. Priv. **16**(5), 38–41 (2018).https://doi.org/10.1109/MSP.2018.3761723. Conference Name: IEEE Security Privacy
19. Network, P.: Pluto - breaking down the barriers in academia [whitepaper]. Technical report, Pluto Network (2018). https://assets.pluto.network/Pluto_white_paper_v04_180719_1355_BSH.pdf
20. OECD/OCDE: Making Open Science a Reality (2015). https://doi.org/10.1787/5jrs2f963zs1-en. https://www.oecd-ilibrary.org/science-and-technology/making-open-science-a-reality_5jrs2f963zs1-en. Series: OECD Science, Technology and Industry Policy Papers
21. Piwowar, H., et al.: The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. PeerJ **6**, e4375 (2018). https://doi.org/10.7717/peerj.4375. https://peerj.com/articles/4375
22. Ross-Hellauer, T.: What is open peer review? A systematic review. F1000Research **6** (2017)
23. Rossum, J.V.: Blockchain for research. Technical report, Digital Science (2017). https://doi.org/10.6084/M9.FIGSHARE.5607778.V1. https://digitalscience.figshare.com/articles/Blockchain_for_Research/5607778/1. Artwork Size: 2269031 Bytes
24. Staworko, S., Boneva, I., Gayo, J.E.L., Hym, S., Prud'Hommeaux, E.G., Solbrig, H.: Complexity and expressiveness of ShEx for RDF. In: 18th International Conference on Database Theory (ICDT 2015) (2015)
25. Taylor, S.J.E., et al.: Open science: approaches and benefits for modeling & simulation. In: Proceedings of the 2017 Winter Simulation Conference. WSC 2017, IEEE Press (2017)
26. Tennant, J., et al..: A multi-disciplinary perspective on emergent and future innovations in peer review [version 3; peer review: 2 approved]. F1000Research **6**(1151) (2017). https://doi.org/10.12688/f1000research.12037.3

27. Tenorio-Fornés, A., Jacynycz, V., Llop-Vila, D., Sánchez-Ruiz, A., Hassan, S.: Towards a decentralized process for scientific publication and peer review using blockchain and IPFS. In: Proceedings of the 52nd Hawaii International Conference on System Sciences (2019)
28. Waltemath, D., et al.: The first 10 years of the international coordination network for standards in systems and synthetic biology (combine). J. Integr. Bioinform. **17**(2–3) (2020)
29. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci. Data **3**(1), 160018 (2016). https://doi.org/10.1038/sdata.2016.18. http://www.nature.com/articles/sdata201618
30. Wirth, C., Kolain, M.: Privacy by blockchain design: a blockchain-enabled GDPR-compliant approach for handling personal data. In: Proceedings of 1st ERCIM Blockchain Workshop 2018. European Society for Socially Embedded Technologies (EUSSET) (2018). https://doi.org/10.18420/blockchain2018_03. https://dl.eusset.eu/handle/20.500.12015/3159
31. Xu, X., Weber, I., Staples, M.: Architecture for Blockchain Applications. Springer (2019). https://doi.org/10.1007/978-3-030-03035-3