Incentivized Research Data Sharing, Reusing and Repurposing with Blockchain Technologies

Oshani Seneviratne
Rensselaer Polytechnic Institute
senevo@rpi.edu

Deborah L. McGuinness
Rensselaer Polytechnic Institute
dlm@cs.rpi.edu

Abstract

Data is the lifeblood of many organizations. Compared to the centralized mechanisms of data sharing (and the subsequent reuse and repurposing), many, if not all, aspects of these processes can be decentralized by using blockchain and provenance semantics. By capturing metadata details at each step of the workflow, data will be easier to audit, verify, and merge with related datasets. It is common in settings where data is either sensitive or valuable (or both) to have formal data use agreements or sometimes less formal rules for reuse, which we have captured in smart contracts. A key innovative aspect of this work is the departure from the traditional natural language-based data use agreements with the aim of making these agreements more computable, resulting in enhanced usability by a broader community. We have also engineered an innovative incentive mechanism for sharing data using an ERC20 token, a popular technical standard for developing fungible tokens on the Ethereum blockchain. The system we developed can be used to track data reuse, thus providing metrics for use in measuring data producers' impact for enterprise reward structures and research measures such as an h-index. As an example application, we discuss how this approach could radically improve the quality and the efficiency of scientific output in the setting of research data sharing. We address the challenge of the costly and time-consuming effort needed to bring an innovative idea from the bench (basic research) to the bedside (clinical level).

1. Introduction

According to Gartner, by 2023, organizations that promote data sharing will outperform their peers on most business value metrics [1]. Experimental approaches that generate data are used in a wide range of settings, from traditional academic science to industrial research, to bring forth innovations from an idea to a consumable product. In fact, data sharing has been recast as a business necessity, with a shift in the mindset from "don't share data unless" to "must share data unless" [2]. Also, as exemplified in this paper, the basis of experimental science as practiced has vital elements, such as studying previous discoveries, coming up with a novel idea, developing a methodology, collecting data, interpreting findings, reviewing peer-reviewed reporting, disseminating research results and data, and incorporating the results into a consumable product. Much of the time spent is critical to test hypotheses and verify outcomes. In some estimates, this process can take as long as 17 years [3]. One vital area that is ripe for significant improvement is *research data sharing in experimental sciences*. We propose using blockchain technologies that provide a platform to verify data and results in a transparent and accountable manner, which could speed up scientific endeavors. The faster the data is shared between the various stakeholders, the more impact the data can have in enabling and accelerating innovation, thus positively impacting society. In fact, since many researchers spend significant time and effort on curating the data, it has been said that they would rather share a toothbrush than data [4]. However, this may change with a better system of trust and accountability.

Our goal in this research is to create a transparent and accountable methodology for sharing data with robust incentives for all the stakeholders involved. To exemplify this goal, we implement a data-sharing application in the context of research data sharing that captures the data life cycle with smart contracts. Such a Decentralized Application (DApp) for research data sharing will enable two or more researchers who have not built up a trust relationship to reuse each other's data in a trustworthy and accountable manner. The critical challenge to pursue this objective is not only enabling the integration of the data spanning heterogeneous data sources and formats but also discovering the data in a decentralized setting and the sustainability of the whole data ecosystem through novel incentive

URI: https://hdl.handle.net/10125/80231
978-0-9981331-5-7

mechanisms. We have designed and implemented a DApp for scientific data sharing as a proof of concept in achieving this ambitious goal.

2. Background

Even though "enterprise blockchain" systems have blossomed in many industry sectors, we have not seen the same level of activity and enthusiasm in using blockchain systems in an area that requires a tremendous amount of transparent, accountable, and global coordination when solving scientific challenges as exemplified by the COVID-19 pandemic. At the same time, enterprise blockchain systems immensely benefit from leveraging experimental data, as that greatly streamlines many of the processes that typically take years to implement. Although we have used research data sharing as an exemplar use case, the lessons learned are equally applicable to many other domains, including business use cases implementing enterprise blockchains. Furthermore, this process can work equally well for data within an organizational consortium managed by a private permissioned enterprise blockchain or a public permissionless blockchain. This section provides the motivation and the background for our work, highlighting both the old and new challenges in science regarding data sharing in academic and industrial research.

The transition process of research data from bench-to-bedside can be compared to supply chain management systems [5]. The specific steps in the bench-to-bedside process as shown in Figure 1 includes many data-related activities. The standard process first involves a researcher having an idea for developing a hypothesis and designing a study. The study is then executed, and the raw material (data) is gathered and stored. It undergoes initial prep (standardization) and advanced processing (analysis). Advanced products (results) are further processed and shipped to suppliers and end-users (peer-reviewed publications). The management and the retrieval of that data are often challenging, and it is often an extra burden for the data quality. Many existing approaches begin with a centralized repository, but if the data are not integrated at a semantic level, access to that data remains challenging, and also, there are no capabilities for semantic search. Due to that, the exploitation of the data remains low.

Similar to unpublished research, if the data in shared research is not made available in some accessible format (e.g., available in a data repository with the appropriate access control), the data is likely to be lost over time [6]. There may be valid reasons for not making the data available, such as for protecting individual privacy with medical or survey data, threats from over-exploitation of species or resources, national security concerns, or matters subject to legal action. However, in cases where the data is not subject to such restrictions or can be handled with access control, if the sharing of data does not transpire within 12 months, then the likelihood of the data publication decreases with time [7], and it may benefit the data creator to publish the data to get credit for their work and to support enhanced data reuse. We believe that the successful deployment of a research data sharing DApp will reduce repetitive work in collecting and reproducing scientific data, as the data would have already been provided in a provenance-aware accountable way. Such a mechanism will enable researchers and their institutions to receive credit for high-quality research data and methods even before publishing and sharing their research under the traditional peer-review process while reducing the number of errors in the scientific literature as the researchers would vet the data thoroughly as they would now be more accountable for their research products.

Furthermore, intellectual property issues and confidentiality are cited as the top reasons researchers [8] and enterprises [9] did not share their data. In particular, researchers in less privileged countries may be afraid of getting scooped by more resource-rich institutions [10]. Furthermore, researchers with more insightful perspectives on interpreting data may use others' data to their advantage [11]. One of the most famous cases of a scientist being disadvantaged in this way is the story of Rosalind Franklin and Nobel Prize winners Watson and Crick, who were all among several labs working to identify DNA structure. "Franklin had been working on the DNA model question using X-ray crystallography to determine the structure. Just a few months before Watson and Crick's landmark paper, Franklin came tantalizingly close to discovering the correct model, but she did not recognize it in the photographs she took. Frustrated with the attitude of her college toward women, Franklin was preparing to leave King's College when fellow scientist Maurice Wilkins, a close friend of Crick, showed Watson a critical photograph." [11] Watson and Crick assessed the double helix structure and published this finding that they may have never reached on their own. Watson and Crick won the Nobel Prize. Franklin got no recognition for her contribution to the discovery because the data sharing was not done ethically. A blockchain-based solution could capture all the stages of the data preparation and the fair and just use of the data through expressive, autonomous, and sustainable data sharing policies encoded in data sharing smart contracts that will avoid such issues.

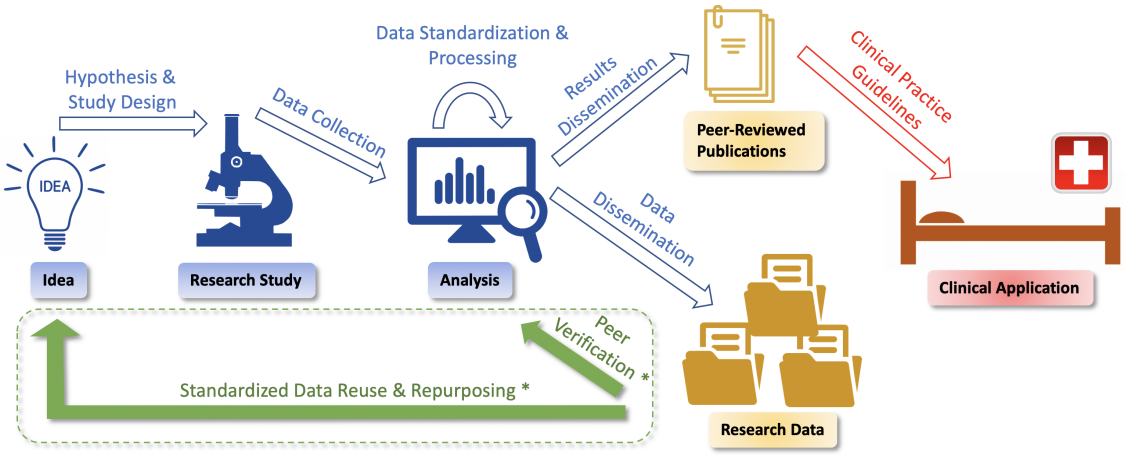

Figure 1. The progression of research in a typical scientific research project. This research focuses on the peer verification of the research data, and the standardized data reuse and repurposing aspects.

When requesting access to data, most often, people are required to have project-specific applications, and approvals. This process can take months to years, effectively excluding many reuse opportunities. For example, the UK-based Biobank's approval process can take several months [12], and this was stated as one of the better examples. Furthermore, researchers rarely consider the possibility of someone unknown to them working with the data, 20 or 50 years later, to correctly interpret the results and derive correct conclusions from the data. Therefore, it is beneficial to the community if the research data were published to facilitate integration with other data and for both intended and unintended reuse into the foreseeable future. Just as it is vital in scientific communication to be clear and precise when writing papers, it is also crucial to communicate effectively about data and their context. However, the quasi-release of data by attaching conditions to their use as it is currently done in centralized repositories is unnecessarily cumbersome that presents a disincentive to others to explore their potential, and these conditions are often impractical to enforce.

There are currently many diverse data sharing policies instituted by journals [13, 14, 15, 16, 17] and funding agencies [18, 19, 20, 21, 22]. These policies are a significant first step towards getting scientific research data shared, but they are insufficient to achieve reliable preservation and subsequent reuse of data [23]. Some journals choose to archive data in their online supplemental materials, but they are typically not curated in a standard way. There is also URL decay in the datasets provided [24] or often unclear protocols on how to share the data [25]. Authors' lack of responsiveness to data requests has led to many journal editors reporting draining experiences dealing with authors who submitted papers under clear data-sharing policies but who subsequently refused to share those data when contacted by other researchers [26]. There is also a change in the perceived willingness to share data over time [27]. Shared research datasets through interoperable protocols enable more transparent science, with better error checking and verification of results.

3. Related Work

A theoretical model of data sharing shows how, starting with a certain population participating, different parameters surrounding sharing cost and incentive cost affect whether more people choose to participate or stop participating [28]. The authors conclude that if sharing and participation costs are too high, participation will not continue no matter how many people are participating in the beginning. With positive sharing and participating costs, participation will continue if the beginning population of participants is over a certain threshold. The most important condition has a negative participation cost, meaning the user is being provided an incentive. In this model, full participation happens no matter the beginning population of participants. A more practical application is described in COEUS [29] as a "semantic web in-a-box" tool developed for research applications capable of simplifying and speeding development of semantic web applications as well as providing an ontology that had terms related to scientific research. It has been applied as a framework for connecting rare disease registries in different locations and formats [30].

There are several blockchain-based DApps for research data sharing in medical research described in the literature. For example, the motivations, advantages,

and limitations, as well as barriers and future challenges faced when applying blockchain technology for sharing data for medical research with a special focus on oncology, are discussed in [31]. The authors have indicated that while there are several existing prototypes of blockchain-based healthcare systems, there is a lack of implementation and evaluation in real-world settings due to the existing barriers of the adoption (e.g., legal, social, and technological limitations).

A blockchain-based pilot study called the Cancer Gene Trust (CGT) was developed for the dissemination of de-identified clinical and genomic data with a focus on late-stage cancer [32]. The CGT addresses the problem of researchers' reluctance to share data to preserve patient privacy, which unfortunately results in little to no learning from real-world data. However, the CGT does not address a scalable incentive mechanism for the stakeholders involved.

A mobile DApp that allows a research study participants to share useful features of their location data derived from geo-coordinates, with a third party research team, without revealing their raw coordinates is described in [33]. Since this is a highly specialized DApp, it is unclear how the methodology described can be applied to general scientific research data sharing settings nor how the researchers and participants would be incentivized to use the DApp.

To address the data management challenges pertaining to protocol compliance, patient enrollment, transparency, traceability, data integrity, and selective reporting in clinical trial management, a solution that combines blockchain and a decentralized file management system called Inter-Planetary File System (IPFS) is discussed in [34]. The main focus of this system is consent management, which is only one side of the equation, and they lack an incentive mechanism that would drive broad-scale adoption.

As electronic health records are increasingly utilized in medical research, a blockchain-based framework with fine-grained access control in which an authorized researcher can search and gain access to the data is introduced in [35]. This system, too, lacks a description of scalable incentives that may ultimately decide the success or the failure of the DApp.

There is also a fair-sized list of startups and organizations attempting to increase data-sharing or data-use via blockchain. The Ocean Protocol [36] is a market built on Ethereum [37] that allows for the sale, consumption, and stake of data. The organization has a native token called OCEAN, an ERC20 token [38], which can be used for the purchase or stake of a data asset. The market provides data scientists a simple platform to "list" datasets. When a dataset is added, a new ERC20 token is minted for the dataset, known as a "Datatoken." To access the dataset, one must own 1.0 of the Datatokens. The dataset, for the most part, is never stored on-chain (on the blockchain). The market, instead, provides theoretical "on and off-ramps" to the data sets. Thus, the interoperable tokens can be sent from wallet to wallet and exchanges without the data ever leaving the hands of the original owner. The use of smart contracts allows for different parameters of agreement when one purchases a Datatoken (access to a dataset). There are multiple types, but the two prominent types, "time-bounded access," which allows for a certain period of access, and "compute-to-data," an on-site computation of data with a provided model or algorithm, can be very attractive in incentivizing data sharing. Although a data marketplace is different from our goal of an ecosystem, we have taken several ideas from the Ocean Protocol. One of the most important aspects in DApps is to create tokens, which represent access to stored datasets instead of storing large datasets on-chain. Decentralized Machine Learning (DML) [39] builds on the idea of "compute-to-data" implemented by the OCEAN protocol, which allows data receivers to run algorithms on datasets remotely without ever receiving the data. In many cases, data contributors would typically much rather do this than share their entire dataset and this approach could be key in increasing participation in our data ecosystem.

4. Application Capabilities

4.1. Smart Contract Powered Autonomous Data Use Agreements

We propose a mechanism to capture the data life cycle with smart contracts. A research study's design directly influences how a study's data is collected and recorded, including associated metadata that can assert data context and ownership. We provide the details for standards needed in developing solutions that will preserve the study's data and metadata integrity, including a protocol facilitated by blockchain technology for data sharing. This methodology will facilitate proper data stewardship, including long-term care in a sustainable way that enables data to be discovered and reused for downstream investigations, either alone or in combination with newly generated data. It will improve the chance that the data will generate correct new insights into new frontiers in scientific research and reduce the number of avoidable questions that the original researcher might have to answer from other researchers about the data. Researchers will be clear about the allowed and disallowed usages of the data that is already shared.

Data agreements encoded in the smart contracts will handle several facets, as discussed below.

Providing Context: When people reuse data that have been made available to them through the DApp, they typically must understand the work required to create those data and the greater insight into the context of those data that their originators had. This process enables them to acknowledge the debt to those original data creators properly and better understand the data and methods, and thus better enables them to use the data appropriately. If the data reusers cite their sources, as they would for print publications, it adds credibility to the data used. If they do not cite data sources, then, as in print media, they may be guilty of plagiarism. We facilitate the requirement for capturing context through the use of nanopublications [40] along with computable provenance that can be directly ingested in smart contracts.

Consolidating Scientific Research Data: There are many data repositories in specialist data centers managed by funding agencies, journals, and institutions from all over the world [41]. We believe a uniform data-use agreement methodology will help to make these rich scientific research data sources appear more unified and accessible. To achieve this objective, we integrate the data access and reuse on disparate data sources using smart contracts on the Ethereum blockchain [37]. The utilization of Ethereum would further enable the widespread adoption of DApp for scientific research, similar to the revolution it has had in the Decentralized Finance (DeFi) space where one could build more complex financial products using a "money lego" architecture [42]. Following a similar methodology, we will create a class of smart contracts that enable researchers to build "science legos" using research artifacts shared with them.

Handling Legal Aspects: Copyright issues are not very likely to compromise data publication and dissemination extensively, because facts and short statements are not copyrightable, although some names and phrases may be trademarked. Thus, data can be routinely extracted from various sources with terms and conditions without infringing copyright through manual or automated means. Furthermore, smart contracts provide a mechanism for an exact representation of the terms and conditions of data sharing and repurposing, which is not subject to interpretation as in the traditional data use agreements.

4.2. Incentive Engineering

The main problem in data availability is not a lack of policy, technology, financial resources, or publication outlets, but rather, the reward system has not kept pace with the new technological opportunities [7]. Trying to get the incentives right is not merely a matter of looking at the monetary incentives. There are different types of non-monetary incentives for those contributing to science and different over-arching incentives for the various stakeholder groups. The quest for knowledge, findings that add to humanity's understanding of the world and ourselves, is what drives and inspires most researchers. Traditionally academic science is not a widely lucrative endeavor for most involved. But the excitement of discovery and having rare or unique knowledge of the world, coupled with the contribution to the advancement of knowledge, and getting credit for said contributions, are some of the non-monetary incentives researchers have. There is also an increasing shift towards highlighting individual contributorship, as opposed to authorship in research, to help identify influential individual researchers [43]. That will also improve our ability to identify the right group of researchers needed to advance research.

Similar to the idea of presenting a token in the blockchain world, in the first half of 2012, the Center for Open Science (COS) introduced "Open Science Badges" for incentivizing open data sharing [44]. While studies are unsure as to whether badges increase sharing practices, the COS claims that badges increased the reported open data rate at the journal from 1.5% to 39.4% over three years [45]. Therefore, a system of badges or rankings based on the reputation index could further increase the incentives to share reproducible and well-annotated data. Our incentive methodology through tokens would follow the same path, and we expect to see similar results.

The h-index is a well-known metric used often for hiring and grant decisions in academia. Since its introduction by Hirsch in 2005 [46], the h-index has turned the academic community into the reputation-based economy we see today. Hood, Sutherland has proposed "Data-index," [47] which is analogous to the h-index, but meant to measure contributions of an individual as a data contributor rather than as a data synthesist or paper author. The Data-index is calculated similarly to the h-index, but the set of papers it is calculated on differs because it uses the cited shared datasets, which gives us a well-defined metric that will differentiate between a data contributor and a data synthesist.

Rewarding Reproducible Research: We are providing a robust mechanism to report the data's usefulness and the research methods' reproducibility. The usefulness and reproducibility will improve the *reputation* of the researcher or organization contributing the dataset in

our incentive model, and thus have a higher reward. Since we aim to build a data community consisting of a network of researchers who share and use their data, the incentives offered to the data community is to reuse the research data of colleagues and others in a field of common interest, with the goal of building upon one another's work, and possibly the commercial application of that research that will have greater monetary benefits. One standard metric of peer recognition is the citation of papers. Citation also shows who is responsible for the information cited and provides its authority, an essential aspect of quality assessment. There is a concern that datasets will not be cited in the same way that print publications would be when they are the source of information. This concern is justified, as most online databases do not provide a citation for each dataset like that of print media, and data users tend to cite the website's URL where the dataset is found rather than the actual dataset and its authors or editors, regardless of whether this information is available. Such incorrect citation is equivalent to authors' citing a journal rather than the papers published in that journal. Our DApp enables reporting of the impact of scientific data at a granular level and the reproducibility of the research methods through the use of nanopublications [40], thus providing incentives to researchers to supply well-annotated, highly reusable data and reproducible methods, which will specifically address the difficulty of attribution across multiple cycles of research. This process has the potential to revolutionize how data is cited and potentially create a new kind of academic reputation index akin to the h-index for data citation.

Rewarding Peer Verification: Researchers can make mistakes either intentionally or unintentionally. Consider the case of 31 retracted cardiology papers that contained fabricated data and infected the literature for more than a decade [48]. Scientists can also unintentionally perpetuate bad results by merely running an array of tests on a dataset until something interesting or statistically significant shows up (commonly referred to as p-hacking) [49]. If such bad research results have been published, we need a mechanism to enable interested parties to rapidly track, identify, and appropriately amend every article that had referenced the original bad result. However, more importantly, if the falsified data was available, others could verify the data and report any issues before the fraudulent results are perpetuated. We believe that practices that improve reproducibility will be more effective than a small percent chance of receiving punishment for practices that reduce reproducibility.

Handling "Researcher's Dilemma": We investigate a quasi-optimal solution to the classic "Prisoner's Dilemma" [50] as it applies to scientific research data sharing, which we will call the "Researcher's Dilemma." Consider two researchers' behavior deciding whether to share materials or data (this use case is similar to the Franklin vs. Watson and Crick story [11] introduced in Section 2). If one scientist shares data, the likelihood of the other scientist solving the problem increases. Sharing has the potential benefit that the other scientist may share their material in the future. Both researchers would be better off if they shared, but neither does so unless the situation is expected to arise again and often. When the prospect of sharing occurs repeatedly, the researchers weigh the current gain from refusing to share against the expected loss from the lack of access to materials or data in the future. The best possible outcome is multilateral cooperation among all the researchers. Since the researchers with shorter time horizons will be less likely to share, their incentives would be higher in the protocol for incentives. If the environment is very competitive (i.e., multiple researchers attempting to be the first in the field), the reward incentives for the researchers who share the data would be made higher. These game-theoretic aspects are taken into account in the design of the ERC20 token associated with the DApp.

5. Use Case

To illustrate the application capabilities described in Section 4, we implemented a use case in a scientific research area that has seen rapid interest in data reuse recently. Data collected as part of the COVID-19 National Collaboration (N3C) [51] was an ideal candidate for this purpose. We mainly investigated the data use agreements and the contribution principles available on the N3C website, which helped us build foundations for creating our blockchain-based decentralized application. Through the N3C research group notes, we learned more about some real-world use cases that utilize COVID-19 data. For example, the Immuno-Suppressed/Compromised Clinical Domain Team researched the COVID-19 infection data to gain a better understanding of how COVID-19 affects patient populations with suppressed or compromised immune systems. This included what kind of data researchers were exploring and what application they built for solving a certain research problem. The user agreement on N3C set several limitations for the users applying for access to data. For example, the data must be used for research purposes, and the N3C sets different data levels for different users. If the users want to get a more detailed and

accurate dataset, there would be more restrictions and requirements.

We analyzed five policies listed in Section 5.1 and implemented them in smart contracts using solidity. The code is available at `https://github.com/sharing-science/data-sharing-dapp`.

5.1. Modeling N3C Data Sharing Policies

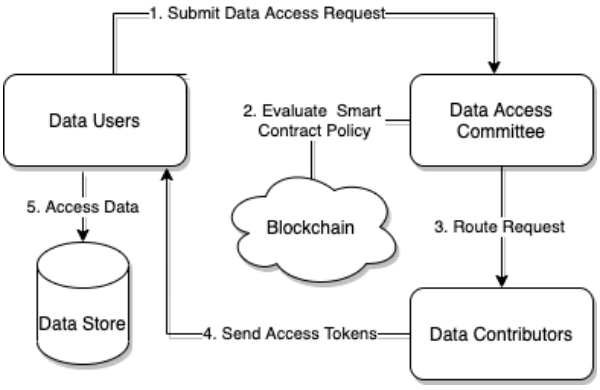

Figure 2. The Modeled N3C COVID-19 Data Request Life Cycle

First, we created a checklist for how the N3C data access committee, a governing body within the N3C that evaluates data access requests, reviewed and approved the data use requests. We also set two major policies for violations, and users should not only agree to obey the rules themselves, but they are also responsible for reporting the violations from any of their collaborators or a third party they share data with. Moreover, in our smart contract implementation, we modeled how the data access committee sets suspension to users, collaborators, and third-party in a more detailed way according to the timely report by users and verification by the data access committee. As an example, **Policy 1** we implemented using a smart contract states:

> *For each proposed Research Project, User(s) agree(s) to submit a Data Use Request to the Data Access Committee for review and approval to access the Data.*

To implement this policy, we utilize a combination of eight functions described below.

1. `CreateContract()`: Create a new contract, initiated by the data access committee by changing the state from `NotReady` to `ReadyforRequirement`.
2. `RequireRequest()`: Data access committee's request for the submission of the request by the user, changing the state to `ReadyforSubmitRequest`.
3. `UserSubmitRequest()`: Submit the request to the data access committee with the change of state to `ReadyforReview`.
4. `Check_isResearcher()`: Check names of project personnel before approving a request.
5. `Check_hasResearchStatement()`: Check the non-confidential research statement.
6. `Check_hasProjectProposal()`: Check the project proposal contained in the request.
7. `Check_hasDataLevel()`: Check if the user has the required data access level.
8. The committee will review the request by checking each requirement for a completed request. If the checklist contains all correct stated information, the request will be approved using `ApproveRequest()`, in which case, the contract will be active by changing to `Active`. Otherwise, the request will be aborted under the function, and then the contract will go through `selfdestruct()`.

The process of a successful data access request implementing the above policy is depicted in Figure 2. The user interface of the DApp given in Figure 3 presents the applicable policies to the user, and for this first policy, they are requested to supplement the file containing information about the request. As can be seen in Figure 3, when the user first gets to the contract page, the contract state is *NotReady* as displayed in the upper right corner. The user can click on the *Agree* button for the first policy and upload their data use request. At this point, the contract state will change to *ReadyforReview*. The data use request the user sent is then routed to the data access committee, who verifies the required information on their end, as demonstrated in Figure 4. If the data access committee ascertains that the user request contains all the required information, the contract state will be changed to *Active*, and the user gets the permission to the requested Data Set.

5.2. Incentivizing Data Sharing and Reuse

We propose an ERC20 [38] token implementation for "data credits," which users receive upon verification and exchange for data. By minting a token and supplying verified users with the given token, we create a theoretical marketplace by which transactions occur. Upon registration and verification, participants receive an initial set amount to use. We have implemented a two-step transaction for this process. The first step is a transaction between the sharer and receiver–a set amount, either static or set by the sharer when they list their dataset. The second step involves a set of confirmations by other participants in the community.

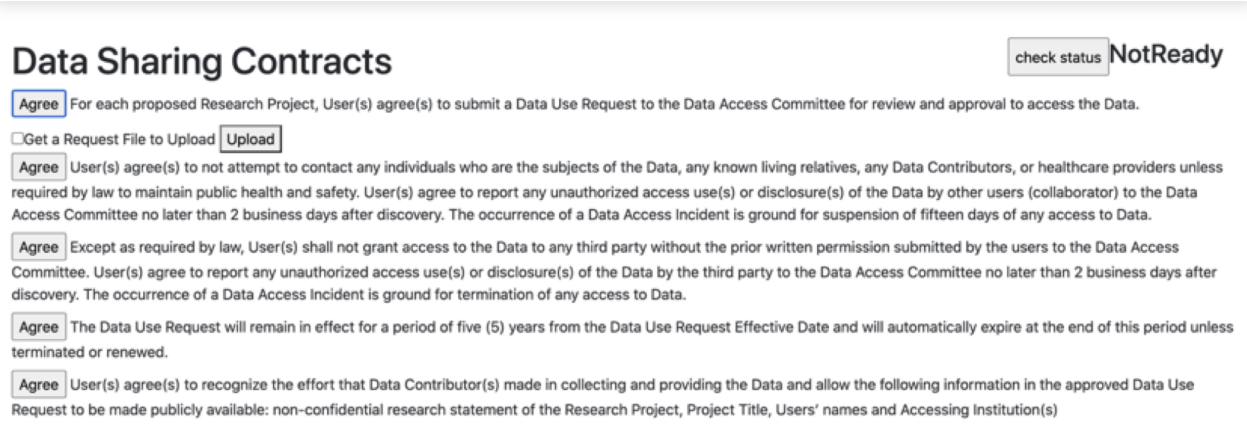

Figure 3. Data Sharing Contracts Available to the User

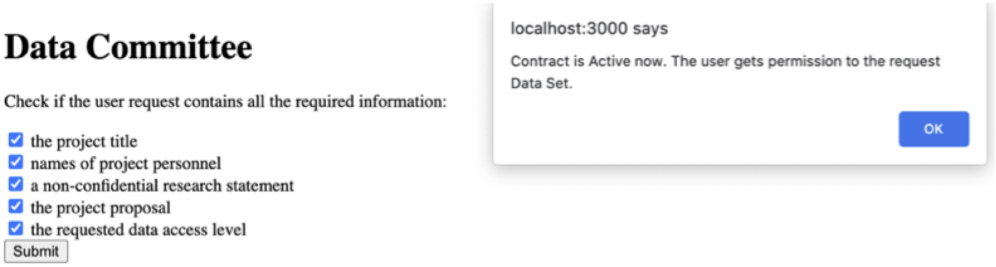

Figure 4. Data Access Committee Granting Permission After Verifying the Data Access Request Proposal

Many of our data use policies we reviewed (introduced in Section 5.1) call for a third party to confirm proposals and the use of datasets. We define this third party as any given set of other participants in the DApp. When a transaction is presented, others in the system are notified of it and receive information about the proposal and requested use. If they choose to participate, they approve or deny the proposal, and for this, we can reward them with "data credits." This assumes that most of the community participating is honest and fair and wants to uphold the integrity of their community. Once the transaction receives a certain number of confirmations, the transaction is approved, and data is shared.

We have implemented a proof of concept ERC20 token to implement the "data credits." This contract acts as a mint for the tokens and holds a mapping to users' accounts to store information about participants. It acts as a factory by which it creates a new smart-contract instance of "Account.sol." The factory then allows us to look up different information about a user using just their linked Ethereum address. It also manages the transfers of data through a function. After verifying the transaction, it sends the ERC20 token from the data recipient's account to the data sharers token. It further handles confirmations by sending new tokens to confirmers' accounts as they participate. The second contract handles single instances of a user. It allows us to store data on-chain about the user, such as their name, the amount of data they have shared or received, and their "data credit" balance. It will also handle the updating of the Data-index.

6. Conclusion

Our contributions are built on blockchain technologies, primarily smart contracts to implement a decentralized application methodology and a novel incentive mechanism for data sharing. We demonstrate the value of our approach by implementing a proof-of-concept decentralized application for scientific researchers to contribute high-quality data in academic and industrial research workflows. However, this methodology is not limited to research data sharing. A key innovative aspect of this work is the departure from the traditional data use agreements to make them more computable, resulting in improved data availability for the broader scientific community. Our application also provides a mechanism for capturing the contribution and intellectual property details of

research data in a trackable and virtually immutable way. Through this research, digital data objects of many kinds may become "first-class" citizens and are not relegated to optional supplementary material. We demonstrate this in the setting of collecting, sharing, and reusing scientific data.

The future outlook of our research includes the possibility to enhance data sharing environments to move from self-reporting websites (e.g., projectreporter.nih.gov and clinicaltrials.gov), and other scientific data repositories (e.g., Genbank [52]) to smart contract enabled automatic query, delivery, and compliance checking. Our research could also create computable data management plans that facilitate scientific data, code, and methodological procedures to be shared in data management plans in a Findable, Accessible, Interoperable, and Reusable (FAIR) manner [4].

Using a fully autonomous process to capture the data life cycle, we will improve how data is collected and give more value to the overall effort undertaken in the experimental data creation and collection process through subsequent reuses of the data. With lower barriers to data access and the preservation of provenance in the setting of publications, under-represented researchers from resource-poor institutions will have better access to quality data, improving science equity and the overall data access to do impactful work. Similarly, in the setting of enterprise data, data contributors who may not be favored or connected in portions of the enterprise may have their data more broadly recognized and reused, thus potentially helping underserved minorities. Our research would also facilitate a more fine-grained and transparent system for data contributors for receiving credit, potentially addressing current problems with credit allocation for data generation.

In conclusion, this research aims to improve scientific data practices (collection, sharing, and reuse) to develop better data-informed products in a shorter period that will benefit society. By providing a robust end-to-end data infrastructure to researchers, we reduce the time it takes to bring a scientific innovation from bench to bedside, allowing innovations to reach vulnerable populations that would greatly benefit from faster miracles. While we focused on scientific data sharing, which often happens through the scientific process, we believe nothing about our approach is limited to scientific settings. A similar approach could be implemented in an in-house enterprise, access-controlled setting. With the realization and large-scale adoption of the work outlined in this paper, data creators worldwide, who are not just in scientific research settings but also many other domains, would contribute their data, receive kudos, and discover new related datasets seamlessly and accountably.

Acknowledgements

We wish to thank our student researchers Kacy Adams, Yicheng (Catherine) Wang, and Ruoyi (Jennifer) Zhan, for their undergraduate research project contributions to the work presented in this paper.

References

[1] Master Data Management - A Geek's Point of View, "The importance of data sharing in organizations," 2020.

[2] Laurence Goasduff, "Data sharing is a business necessity to accelerate digital business," 2021.

[3] Z. S. Morris, S. Wooding, and J. Grant, "The answer is 17 years, what is the question: understanding time lags in translational research," *Journal of the Royal Society of Medicine*, vol. 104, no. 12, pp. 510–520, 2011.

[4] M. Wilkinson, "Interoperability with Moby 1.0-it's better than sharing your toothbrush!," *Nature Precedings*, pp. 1–1, 2008.

[5] "Unravelling complexity: The challenge of compliance in the life sciences supply chain," April 2018.

[6] P. B. Heidorn, "Shedding light on the dark data in the long tail of science," *Library trends*, vol. 57, no. 2, pp. 280–299, 2008.

[7] M. J. Costello, "Motivating online publication of data," *BioScience*, vol. 59, no. 5, pp. 418–427, 2009.

[8] Alice Meadows, "To Share or not to Share? That is the (Research Data) Question...," Nov. 2014.

[9] C. Arthur, "Businesses unwilling to share data, but keen on government doing it," June 2010.

[10] L. Bezuidenhout and E. Chakauya, "Hidden concerns of sharing research data by low/middle-income country scientists," *Global Bioethics*, vol. 29, no. 1, pp. 39–54, 2018.

[11] R. Fernandez, "Barriers to open science: from big business to Watson and Crick," Sept. 2010.

[12] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, *et al.*, "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *Plos med*, vol. 12, no. 3, p. e1001779, 2015.

[13] "Springer Nature Data Access Policy." `https://www.springernature.com/gp/authors/research-data-policy`.

[14] "Wiley's Data Sharing Policies." `https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/data-sharing-citation/data-sharing-policy.html`.

[15] "Taylor & Francis - Data availability statements." `https://authorservices.taylorandfrancis.com/data-sharing-policies/data-availability-statements/`.

[16] "Science Editorial Policies - Data and Materials Availability after Publication." https://www.sciencemag.org/authors/science-journals-editorial-policies.

[17] "Plos One - Data Availability." https://journals.plos.org/plosone/s/data-availability.

[18] National Institutes of Health, "Final NIH Statement on Sharing Research Data," Feb. 2003.

[19] Horizon Europe, "The Next EU Research & Innovation Investment Programme," Apr. 2019.

[20] G8, "Open Data Charter," June 2013.

[21] Horizon Europe, "China Open Science and Open Data Mandate Released," Mar. 2018.

[22] Belmont Forum, "Belmont Forum Data Accessibility Statement and Policy," Oct. 2018.

[23] L. Jones, R. Grant, and I. Hrynaszkiewicz, "Implementing publisher policies that inform, support and encourage authors to share data: two case studies," *Insights*, vol. 32, no. 1, 2019.

[24] J. D. Wren, "URL decay in MEDLINE—a 4-year follow-up study," *Bioinformatics*, vol. 24, no. 11, pp. 1381–1385, 2008.

[25] M. A. Noor, K. J. Zimmerman, and K. C. Teeter, "Data sharing: how much doesn't get submitted to GenBank?," *PLoS Biol*, vol. 4, no. 7, p. e228, 2006.

[26] C. J. Savage and A. J. Vickers, "Empirical study of data sharing by authors publishing in PLoS journals," *PloS one*, vol. 4, no. 9, p. e7078, 2009.

[27] E. G. Campbell, B. R. Clarridge, M. Gokhale, L. Birenbaum, S. Hilgartner, N. A. Holtzman, and D. Blumenthal, "Data withholding in academic genetics: evidence from a national survey," *jama*, vol. 287, no. 4, pp. 473–480, 2002.

[28] S. Xuan, L. Zheng, I. Chung, W. Wang, D. Man, X. Du, W. Yang, and M. Guizani, "An incentive mechanism for data sharing based on blockchain with smart contracts," *Computers & Electrical Engineering*, vol. 83, p. 106587, 2020.

[29] P. Lopes and J. L. Oliveira, "COEUS:"semantic web in a box" for biomedical applications," *Journal of biomedical semantics*, vol. 3, no. 1, pp. 1–19, 2012.

[30] P. Sernadela, L. González-Castro, C. Carta, E. Van Der Horst, P. Lopes, R. Kaliyaperumal, M. Thompson, R. Thompson, N. Queralt-Rosinach, E. Lopez, *et al.*, "Linked registries: connecting rare diseases patient registries through a semantic web layer," *BioMed research international*, vol. 2017, 2017.

[31] A. Dubovitskaya, P. Novotny, Z. Xu, and F. Wang, "Applications of blockchain technology for data-sharing in oncology: results from a systematic literature review," *Oncology*, vol. 98, no. 6, pp. 403–411, 2020.

[32] B. S. Glicksberg, S. Burns, R. Currie, A. Griffin, Z. J. Wang, D. Haussler, T. Goldstein, and E. Collisson, "Blockchain-authenticated sharing of genomic and clinical outcomes data of patients with cancer: a prospective cohort study," *Journal of medical Internet research*, vol. 22, no. 3, p. e16810, 2020.

[33] M. Johnson, M. Jones, M. Shervey, J. T. Dudley, and N. Zimmerman, "Building a secure biomedical data sharing decentralized app (dapp): Tutorial," *Journal of medical Internet research*, vol. 21, no. 10, p. e13601, 2019.

[34] I. A. Omar, R. Jayaraman, K. Salah, M. C. E. Simsekler, I. Yaqoob, and S. Ellahham, "Ensuring protocol compliance and data transparency in clinical trials using blockchain smart contracts," *BMC Medical Research Methodology*, vol. 20, no. 1, pp. 1–17, 2020.

[35] J. Sun, L. Ren, S. Wang, and X. Yao, "A blockchain-based framework for electronic medical records sharing with fine-grained access control," *Plos one*, vol. 15, no. 10, p. e0239946, 2020.

[36] Trent McConaghy, "How ocean can benefit data scientists," Aug. 2020.

[37] V. Buterin *et al.*, "Ethereum white paper," *GitHub repository*, vol. 1, pp. 22–23, 2013.

[38] "Ethereum improvement proposal-20: Token standard."

[39] DML Developers, "Decentralized machine learning white paper," Dec. 2017.

[40] P. Groth, A. Gibson, and J. Velterop, "The anatomy of a nanopublication," *Information Services & Use*, vol. 30, no. 1-2, pp. 51–56, 2010.

[41] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. P. Hill, R. Kania, M. Schaeffer, S. St Pierre, *et al.*, "The future of biocuration," *Nature*, vol. 455, no. 7209, pp. 47–50, 2008.

[42] S. M. Werner, D. Perez, L. Gudgeon, A. Klages-Mundt, D. Harz, and W. J. Knottenbelt, "SoK: Decentralized Finance (DeFi)," *arXiv preprint arXiv:2101.08778*, 2021.

[43] A. O. Holcombe, "Contributorship, not authorship: Use CRediT to indicate who did what," *Publications*, vol. 7, no. 3, p. 48, 2019.

[44] Center for Open Science, "Open science badges enhance openness, a core value of scientific practice.."

[45] A. Rowhani-Farid, M. Allen, and A. G. Barnett, "What incentives increase data sharing in health and medical research? a systematic review," *Research integrity and peer review*, vol. 2, no. 1, pp. 1–10, 2017.

[46] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National academy of Sciences*, vol. 102, no. 46, pp. 16569–16572, 2005.

[47] A. S. Hood and W. J. Sutherland, "The data-index: an author-level metric that values impactful data and incentivises data sharing," *bioRxiv*, 2020.

[48] I. Oransky and A. Marcus, "Harvard and the Brigham call for more than 30 retractions of cardiac stem cell research," Oct. 2018.

[49] M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions, "The extent and consequences of p-hacking in science," *PLoS Biol*, vol. 13, no. 3, p. e1002106, 2015.

[50] R. Axelrod, "Effective choice in the prisoner's dilemma," *Journal of conflict resolution*, vol. 24, no. 1, pp. 3–25, 1980.

[51] M. A. Haendel, C. G. Chute, T. D. Bennett, D. A. Eichmann, J. Guinney, W. A. Kibbe, P. R. Payne, E. R. Pfaff, P. N. Robinson, J. H. Saltz, *et al.*, "The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment," *Journal of the American Medical Informatics Association*, 2020.

[52] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "Genbank," *Nucleic acids research*, vol. 41, no. D1, pp. D36–D42, 2012.