View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by AIS Electronic Library (AISeL)

Association for Information Systems

AIS Electronic Library (AISeL)

Wirtschaftsinformatik 2021 Proceedings

Track 14: Data management and data ecosystems

Permissioned Blockchain for Data Provenance in Scientific Data Management

Julius Möller
Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

Sibylle Fröschle
Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

Axel Hahn
Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

Follow this and additional works at: https://aisel.aisnet.org/wi2021

Möller, Julius; Fröschle, Sibylle; and Hahn, Axel, "Permissioned Blockchain for Data Provenance in Scientific Data Management" (2021). *Wirtschaftsinformatik 2021 Proceedings*. 7.
https://aisel.aisnet.org/wi2021/LDatamanagement14/Track14/7

This material is brought to you by the Wirtschaftsinformatik at AIS Electronic Library (AISeL). It has been accepted for inclusion in Wirtschaftsinformatik 2021 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Permissioned Blockchain for Data Provenance in Scientific Data Management

Julius Möller[1], Sibylle Fröschle[1] and Axel Hahn[1]

[1] University of Oldenburg, Ammerländer Heerstraße 114-118, 26129 Oldenburg, Germany
{julius.moeller, sibylle.froeschle, axel.hahn}@uni-oldenburg.de

Abstract. In the age of Big Data, the amount of data-driven research activities has increased significantly. However, when it comes to collaborative data processing in scientific workflows, provenance information of the used data is not always accessible. Especially in complex data ecosystems with multiple decentralized data sources, it is hard to keep track of the processing operations once they are completed. When sharing such data between different researchers and other involved parties, poor traceability of processing steps may also obstruct this process. In this paper, we introduce a blockchain based data provenance information system, which enables decentralized sharing of this information. We then integrate this system into the decentralized data sources context and address trust and traceability issues in the network with an identity-based solution. Furthermore, the system's performance is evaluated, and the concept is examined in a case study on the e-Maritime Integrated Reference Platform (eMIR).

Keywords: data provenance, blockchain, scientific data management.

1 Introduction

The automated recording and storage of huge amounts of data is increasingly important in both research and industry. The management of such *big data* data sets has long since ceased to be trivial and has become a major challenge for research and industry [1]. Additionally, the growing need for high-quality data assets in nearly any branch of industry has led to a new awareness of the actual value of data. In a data ecosystem controlled by different Data Producers, Data Owners, Data Consumers and Data Miners all represented by different physical entities, there is a need for tracking the production, transformation, and provision of data [2]. Also, it is a common scenario in industry and research that project partners agree on a specific objective and work together with different sets of data and data transformation nodes in shared networks. While this often happens in private networks, there are also emerging concepts for the usage of potentially public *data spaces* (cf. [3] for a general description or [4] for a reference architecture). Especially in research, specific data often must be selected, pre-processed, transformed and analyzed from a multitude of data. This digital process known as *e-Science workflow* has been discussed in a great number of publications (see e.g. [5] for an introduction to the topic, [6] for a taxonomy of e-Science workflow systems, and [7] for a more extensive overview). It is really important to be able to track every process step in the e-Science workflow to guarantee a high-quality data-driven research methodology [8]. Moreover, other researchers must be able to verify the authenticity and non-repudiation of the workflow metadata thus created to fully understand the process from which research findings have been made. The enormous value of scientific data for further processing in industrial applications, such as the training of decision-supporting machine learning models cannot be denied. Currently, many of the challenges of collaborative data processing are being addressed by upcoming cloud-based platform solutions [9]. While the cloud platform provider may be trustworthy and reliable, different parties providing, preparing, and transforming the data may not. Keeping track of the creation, the changes and the provision of data is a challenge in platform supported data spaces. Most of these problems can be observed in research activities involving industrial partners with economic interests: For instance, how could shipping companies provide data on vessel movement and fuel consumption as a basis for a collaborative research project on traffic optimization? Also, areas in which multiple partners need to cooperate, as it is done for example in the logistics industry, face similar problems: For instance, how could data from independent storage and transportation companies be securely made available, be processed and analyzed by other companies to gather knowledge about influence factors that can affect efficiency? The goal of this paper is to provide a decentral solution that closes these gaps and fits into the scientific data ecosystem. Our contributions are as follows: Firstly, we describe the setting of data provenance in e-Science. Secondly, with the assumption of an existing data space setup, we elicit the requirements a decentral solution needs to satisfy and motivate how the use of blockchain with an identity-based consensus method is best suited to this purpose. Thirdly, we present the architecture of our system. Finally, a prototypical implementation is evaluated in an example with an existing maritime data space.

2 Scientific Data Management in a Decentralized Context

2.1 Scientific Workflows and Data Provenance

The activity of scientific data management is often presented in cycles or processes. In general, this includes the steps from the import of source data to the extraction of knowledge from the processed data. This procedure is an important element of e-Science (electronic science), which deals with the generation of knowledge using digital infrastructures [10]. A well-founded and detailed model for the scientific data management process is provided by Crowston and Qin [11]. In a comparison of nine data management cycles/process models by Ball [12], the model of Crowston and Qin is identified as one of the most comprehensive models. **Fig. 1** shows a *summary* of the model. For workflow-oriented e-Science, data provenance is a very relevant topic: As a large number of publications with data-driven approaches for problem analysis and solving is emerging, the insufficient availability of trustworthy traceability measures is increasingly becoming a problem [13]. In the case of a poorly documented data processing workflow, other researchers would not be able to reproduce the author's results. Buneman et al. [14] define *data provenance* as follows: "*Data provenance – is the description of the origins of a piece of data and the process by which it arrived in a database*". The work of Simmhan et al. [2] provides a taxonomy of data provenance in e-Science: The application of provenance is subdivided into the sections of *data quality, audit trail, replication recipes, attribution,* and *informational purposes*. It also can be distinguished if the provenance information is related to the *data product* or *the process of its creation*. In this work, we will keep the focus on the audit trail for the whole process from data creation to the final data product.

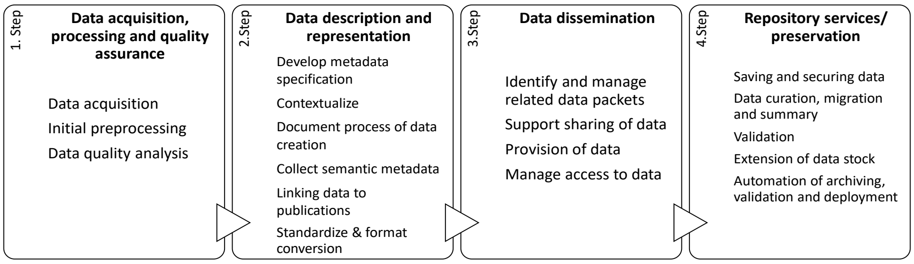
1. Step
Data acquisition, processing and quality assurance
Data acquisition
Initial preprocessing
Data quality analysis
2.Step
Data description and representation
Develop metadata specification
Contextualize
Document process of data creation
Collect semantic metadata
Linking data to publications
Standardize & format conversion
3.Step
Data dissemination
Identify and manage related data packets
Support sharing of data
Provision of data
Manage access to data
4.Step
Repository services/ preservation
Saving and securing data
Data curation, migration and summary
Validation
Extension of data stock
Automation of archiving, validation and deployment

Fig. 1. Summary of the Scientific Data Management Process as described by [11].

2.2 Decentralized Data in Data Spaces

The term "data space" is widely used in different contexts. In the scope of this paper, we use the definition of data spaces given by Franklin et al. [3] who define a data space as a co-existent amount of data which is linked by a "data space support system" (or specifically a "data space support platform (DSSP)"). This system must fulfil a set of requirements to be recognized as such. Firstly, it must support a wide range of data types and formats covering all data in the data space. Secondly, it must offer means of searching, querying, updating, and administrating the data space. Data space queries are not required to result in a complete result of available data, an approximation is sufficient. And lastly, it must support tools to create a tighter integration of the data in the data space.

Data spaces can be found in situations where partial control over or knowledge of several data sources is available to a central entity. This central entity, however, is not able to maintain full control over the data sources and therefore tasks like data ingestion and harmonization are not trivial. Additionally, data spaces typically contain sets of syntactically and semantically different data. [15]

Data space architectures have already been realized in several publications, e.g. as a vehicular data space [16], IoT data space [17] or maritime data space [18].

2.3 Identity-Based Blockchain

The blockchain concept has increasingly been applied in a large number of cases for enhancing cyber security and decentralizing control structures and has also been investigated for usage in a scientific research context e.g. in [19]. A blockchain typically works like a distributed database with some special functional principles, such as finding a network consensus on adding new information to the blockchain. Most consensus algorithms for blockchain applications require the cooperation of a vast number of nodes in the blockchain network. This often leads to slow performance when a new block needs to be accepted. Assuming that a smaller group of nodes with trusted identities, and only these nodes are used to determine a consensus, the performance can be improved significantly. Consensus algorithms utilizing this assumption are called Proof-of-Authority consensus [20]. Another important factor in the application of blockchain technology is the permission policy of the network. Common policies for blockchain are public, consortium-based, and private. These approaches mainly differ in the degree of centralization. Furthermore, permissions for reading data from and writing data to the blockchain may also be restricted depending on the permission policy of the blockchain [21].

2.4 Related Work

For the literature review, we analyze work in the area of data provenance in scientific data management with special regards to security, architecture, and workflow models. Additionally, we discuss work that uses blockchain technology in the context of storing data provenance information. The importance of data provenance for scientific data processing has already been discussed in a significant number of publications (see e.g. [22] for an overview, and [23], [24] for applications). Additional work on the security of data provenance has also been conducted in the past years. The work of Bertino et al. [25] gives a good overview of this topic and presents an architecture framework and methods for the secure exchange of data provenance. However, collaborative editing of this information is not considered. Hasan et al. [26] introduce a formal model for a secure provenance chain, in which document editing steps are cryptographically signed by their originators. In addition to that, hashes of the changed data are appended to the blocks of the editing chain. The model relies on public key cryptography and provides a good baseline for the secure provenance documentation. A framework for finding a consensus on a valid edit in a network of editing users is not discussed. Closely related to the work in this paper are the approaches of Ramachandran et al. [27] and Liang et al. [28], which both use a blockchain-based approach for securely organizing data provenance information. Ramachandran et al. use the Ethereum blockchain and smart contracts with the Open Provenance Model (see [29]) as their base. The consensus on a change of a document is determined by voting with all nodes or by randomized threshold voting and therefore seems very comprehensible for participants. The approach is evaluated with two real-world use-cases and the performance is considered applicable by the authors of the paper. Liang et al. also propose a blockchain based architecture, which, however, aims at integrating a *central* cloud-provider that stores the data that is being edited. An action-based method for tracking the changes in documents is utilized for creating the data provenance information. The blockchain is used to carry a distributed database which includes the tracked changes of the documents. Both works do not solve the problem of unidentifiable entities and only partially discuss the challenges of decentralized data sources. Apart from these papers, there are some others that *partly* address some of the discussed problems. Chen et al. [30] present a formal model for a blockchain data structure for efficient sharing of scientific workflow provenance data. Neisse et al. [31] discuss different design choices of a blockchain-based data provenance approach and their compliance with the GDPR. Finally, Tosh et al. [32] compare different consensus methods for cloud-based data provenance and come to the conclusion, that a Proof-of-Stake consensus seems to be the best method in such a setup.

2.5 Research Objective

There is an increasing need for data provenance solutions for scientific data management. Especially in data space environments, solutions with the ability to handle a high degree of decentralization of data need to be developed. In existing work, the problem of permissions to edit provenance information or having the right to vote for appending new data provenance information is not solved entirely. An identity-based permission system could possibly solve this issue and establish trust in a system of different editing parties. Also, the existing architectures cannot address the decentrality of the actual data in a data space setup. Requirements for scientific data processing are also only being discussed partially. Therefore, a permissioned, identity-based blockchain seems to be a candidate technology for establishing a data provenance information system. This seems to fit best to the presented scenario, as a central ledger infrastructure may not be able to establish overall trust. Moreover, it would require an independent organization to govern the ledger and protect it against security risks. While it is still possible to have authenticated and authorized entities, a central party would need to take the responsibility for the system, which would be a problem with several parties that may have conflicting interests that would have to be resolved for each workflow individually. A data provenance information system must be able to ensure the secure documentation of data provenance information and its consistency with the actual data in a trustworthy and reliable way for authorized entities. Additionally, it should be visible to researchers who authored the data provenance information. The system should be adapted to the needs of a data-driven science process, being able to track single workflow steps of data processing.

3 Design of the Data Provenance Blockchain with Identifiable Entities

3.1 Architectural Components

To introduce a secure documentation of data provenance in scientific workflows, several architectural components are required. **Fig. 2** gives an overview of the involved entities and components and their interactions. We assume, that an existing data space is present and has a DSSP as the corresponding support system (cf. section 2.2) as this setup is one of the most common solutions.

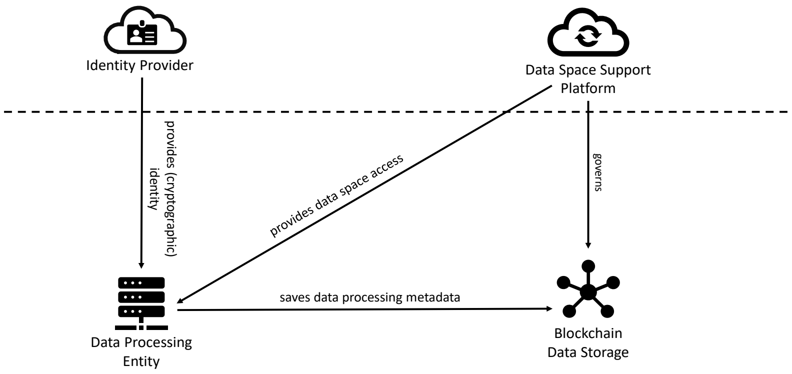
Identity Provider
Data Space Support Platform
provides (cryptographic) identity
provides data space access
governs
saves data processing metadata
Data Processing Entity
Blockchain Data Storage

Fig. 2. Architecture overview of the data provenance management concept.

The architectural components can be described as follows:

- **Identity Provider:** The Identity Provider is assumed to be a trusted entity with the function of providing cryptographic key-pairs linked to legal entities. Prerequisite for this is the existence of a Public-Key Infrastructure (PKI). The Identity Provider is needed for the identity-based consensus in our blockchain setup.
- **Data Space Support Platform (DSSP):** The Data Space Support Platform is the access point for data space access requests. It may also fulfil the functions of a Data Processing Entity as workflow steps of the e-Science Workflow may also be executed on the platform.
- **Data Processing Entity:** The Data Processing Entity is processing data from or provides data to the data space, which it accesses via the DSSP. Several Data Processing Entities can be involved in the processing of a single data set.
- **Blockchain Data Storage:** The Blockchain Data Storage contains the actual data provenance information and may also be used to organize data space access rights via smart contracts.

When a dataset is created, the originator, i.e. the first Data Processing Entity, provides first information on the data and makes the data available to the data space via the DSSP. The metadata of the data creation is then stored in the blockchain and can be retrieved by the next Data Processing Entity in the workflow. The data is then again processed, made available to the data space and the metadata is stored in the blockchain. Access to the blockchain always requires a cryptographic identity, provided by the Identity Provider.

3.2 Data Provenance Model

The classes and attributes of a data provenance model always depend on their use-case and the domain they are applied to. The data provenance model in our approach should describe the creation and processing of scientific data in e-Science workflows. We assume, that every transformation of the data can be partitioned into a chain of single processing steps. As a proof of concept, we use a simple workflow-oriented model whose steps are derived from the tasks of Crowston and Qin's model (see section 2.1). We design this model in such a way that it can act as a template and can be extended further easily. Hence, we deliberately keep the attributes in our model general. We generalize the steps of the e-Science workflow to the following tasks: *Data Acquisition Process*, *Anonymization* (to comply with data protection regulations), *Data Quality Analysis*, *Preprocessing and Transformation* and *Conversion and Validation*. A formalized model of the proposed tasks is used to represent the data provenance information. Instances of this model for workflow steps can be serialized and then stored in the body of a blockchain block. The stakeholders in the processing of scientific data in our data space set-up can be modelled through the following roles (cf. [4]):

Data Owner. The Data Owner is considered possessing the actual data. This can be interpreted in a legal or technical sense and is not further specified for our approach. The Data Owner determines the access rights to the data.

Data Provider. The Data Provider is an entity which provides the technical means to access a specific data set. The Data Provider must be authorized by the Data Owner and only provide the data to other entities with access rights granted by the Data Owner.

Data Consumer. The Data Consumer is accessing a data set as a client of the Data Provider.

Physical entities in this model can also have multiple roles at the same time. Refer to section 4.1 for an example.

3.3 Blockchain Architecture

Identities. In a collaborative research scenario, the anonymity (as e.g. found in crypto-currency blockchains) of Data Processing Entities would lead to less traceability and trust between different parties as manipulations of the data would not cause any negative reputation for the guilty parties. Furthermore, the research community would benefit from a secure and transparent documentation of data processing workflows as investigations become easier reproducible. In our context, it is not a given that the transformations on a data set always can be reproduced and verified (against a hash) by any participant. Hence, for our system we require technological measures to be in place so that an entity that has processed data cannot repudiate their processing step and can be held responsible for the result. These considerations lead to the conclusion that a blockchain, applied to this problem would only fulfil its purpose if Data Processing Entities in a scientific workflow can be identified. We assume that physical identities are bound to cryptographic key pairs. To obtain such a key pair, Data Processing entities must fulfil several requirements, which are defined through the Identity Provider. These could be for example the evidence that a Data Processing Entity is part of a legally registered organization. After obtaining a key pair and a certificate stating its validity from the Identity Provider, the Data Processing Entity can participate in the blockchain network (see **Fig. 2**). Every transaction in the network that is committed by a Data Processing Entity must be signed with its private key, so that other entities can trace his interactions with the processed data.

Transactions. Storing data provenance information in the blockchain can be achieved in several different ways. Nevertheless, it must be kept in mind that any data, which is stored in a conventional blockchain is replicated by every node in the network and therefore causes traffic. Typically, data that is stored in a blockchain can be represented as transactions. From the perspective of a data space entity, the smallest monitorable change of a data set in our model is a single workflow step. For an entity who is executing a workflow step, there may be smaller processing steps as parts of its implementation of the workflow step, but these are normally not visible to any other entities. Hence, it is appropriate to define a workflow step as a transaction in the blockchain. Transactions also must contain a hash of the processed data to later enable other entities to verify that a transaction was related to a specific workflow step. This binds the data provenance information in the blockchain to the actual data set. In addition to the hash, meta-information (see section 3.2) for processing the data by the system is also stored. However, this information cannot always be expected to include exact descriptions of used processes, as they may contain proprietary algorithms.

Blockchain Structure. It must be kept in mind, that in a data space, there is not only a single data set, that is being processed by its entities. Thus, there are several chains of transactions that must be stored in the blockchain network. For this reason, we propose to use multiple shorter chains, each representing a workflow for a single source data set (see **Fig. 3**).

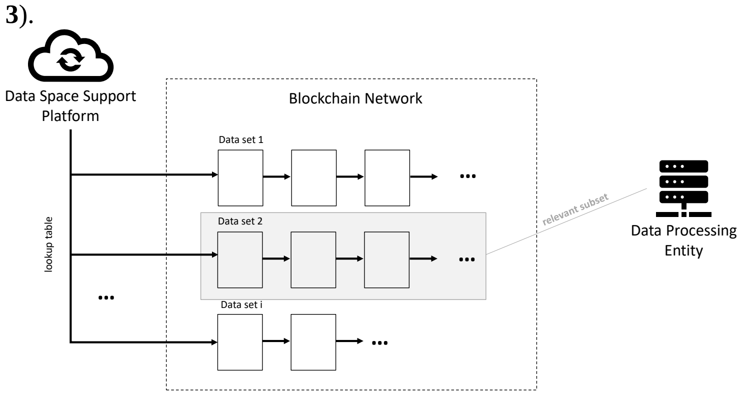

Fig. 3. Data set specific blockchain setup with multiple chains.

This has some advantages over a conventional, single blockchain: First, permissions can easily be set for every data set separately. Also, entities do not need to keep track of data sets, which they are not permissioned to access or not interested in. This reduces the locally used storage of the blockchain instance and prevents entities from wasting their computational resources to track transactions, in which they do not have any interests. To keep track of the different chains in the network, the DSSP can provide a central lookup table or any other means for optimizing access to the blockchain network. This task falls directly within the remit of such a platform. The permission model and deployment of the blockchain should follow a standardized process. In a more proprietary setup, Data Owner, Provider and Consumer may also have problems on finding a consensus on a process, even in a small group due to conflicting interests or because data exchange setups can also be dynamic or even fully automated. Standardization will largely prevent the occurrence of these problems and support the balance between administrative burden and benefits of the proposed method. Standardized procedures can also be supported by the DSSP.

Consensus and Smart Contracts. In the defined setup, the stakeholders in the process of data processing in the data space have been clearly identified. The existence of an Identity Provider now makes it possible to use a Proof-of-Authority (or identity-based) consensus method. The Data Owner of a source data set has been defined as the entity which holds the rights to distribute and modify the data set. We propose that the Data Owner nominates a subset of entities in the network that are authorized to vote on data provenance auditing for the data provenance information related to the specific data set (see **Fig. 4**).

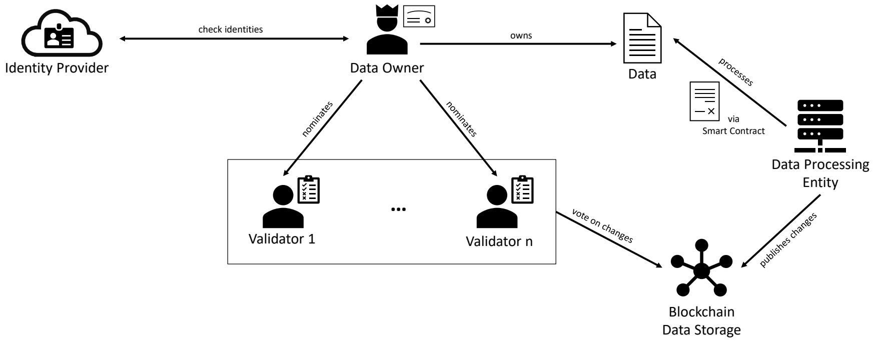

Fig. 4. Conceptual overview of Proof-of-Authority consensus mechanism.

The Data Owner can verify the identities of these *validators* via the Identity Provider. It is left open how the Data Owner determines this subset, as there can be several legal, organizational, or technical requirements which will be specific to the case. For example, the authorization of being a validator may include contractual agreements, which require validators to pay penalties, if agreements are violated. In return, a Data Owner may provide access to its data or act as a validator for the other party. Also, for scenarios with a stronger need to protect data, entities may also consider paying an independent organization to provide validation facilities. However, in the case of working with highly sensitive data, validators must be included in the process of the data processing and may be selected from the set of existent and authorized Data Processing Entities for validating the data processing steps of other authorized Data Processing Entities. Also, in less critical workflows, validators may also base their decisions on data processing metadata, without requiring access to the actual data sets. Anytime a new block will be added to the blockchain, only the validators vote on the changes included in this block. In this way, our setup fulfils the definition of a consortium blockchain. Consensus algorithms like e.g. Aura or Clique can be used to implement the building of a consensus [20].

In the near past and with the approach of the Ethereum blockchain, so-called smart contracts have been in the focus of blockchain researchers and developers [33]. Smart contracts are pieces of code, which run on the blockchain and execute contract terms that have been defined in the code [34]. In the arrangement in **Fig. 4**, the Data Processing Entity accesses and processes the data, provided by the Data Owner via the data space. We propose to use the data provenance blockchain to deploy smart contracts between the Data Owner and Data Processing Entities for the determination of access, modification, and distribution rights of data sets. The smart contracts will be deployed on the blockchain that is linked to the corresponding workflow. The DSSP can then subscribe to these smart contracts and manage data space access accordingly. Additionally, the nomination of validators may also be carried out via a smart contract. This formalizes the processes of data and right management and makes it decentrally available to all authorized parties. This completes our system design.

3.4 Security Analysis

The proposed system stores data provenance information without giving unauthorized parties the possibility to tamper with this information or its consistency with the data sets it relates to. Moreover, the system provides traceability (transactions can be traced to a legal identity) and non-repudiation (a participant cannot deny having carried out a transaction on the data set): this is implemented via the signatures within the blockchain structure and the binding of the access control to the data sets to the permissions given through the blockchain structure. We analyze what can happen when an unauthorized or authorized entity is compromised as well as blockchain specific attacks:

Threat 1: Unauthorized entities. When an unauthorized entity tries to add false information to the data provenance blockchain, this will be detected by the validators of the blockchain and the transaction will be discarded due to invalid signatures. Similarly, unauthorized entities will not have access rights to tamper with the data set.

Threat 2: Compromised validators. In general, there must be a significant amount of compromised validators [20], which is relatively unlikely in a data space setup with independent validators. However, if the Data Owner nominates a set of highly dependent validators this can become a security issue if he does not ensure that they are highly trustworthy at the same time. Also, conspiring validators will suffer a loss of reputation and possibly legal consequences if this attack is detected.

Threat 3: Compromised Data Owner. If a Data Owner is compromised, then he will perhaps be able to provide fake data within the original data set. However, since he must sign the data provenance information in the blockchain it cannot be denied having made the claim that it is real data later. Hence, when someone discovers that the data is not authentic, the compromised Data Owner risks his repudiation as Data Provider or could even be made liable if damage is caused. If a compromised Data Owner tries to tamper with the transactions or adds transactions, he is not authorized for then this will be spotted by the validators. Even though as the Data Owner he could nominate conspirator validators this is unlikely (cf. Threat 2). The case of a compromised Data Processing Entity is analogous.

Threat 4: Compromised DSSP lookup table. If the lookup table would contain false information, this would only lead to false access in the blockchain, which would be detected by any entity verifying the signatures or hash-values in the blockchain by cross checking with the identities of the expected entities with the help of the Identity Provider. The attacker might still duplicate a chain or prefix of a chain. However, this will not cause any harm as each block (describing data transactions) contains the hash of the actual data, and this data hash is cryptographically bound together with the metadata by the signature of the processing party. Moreover, if a regular party or the attacker tries to add blocks to a duplicated blockchain or prefix in a way that would lead to a fork of the workflow with respect to the respective blockchain then this will not pass the consensus algorithm as usual. If the attacker tries to add a new block with an inconsistent match between metadata and an existing data set, then this will be detected by the validating nodes as usual. At most, if the attacker duplicates a prefix of a chain the information that a data set was deleted might be lost, and a regular party who follows the corresponding link will not be able to access the data set as expected. In general, duplication is less of a problem here than in currency blockchains since the data sets and their provenance records are not "consumed" but rather a derived data set has to be deleted explicitly.

Threat 5: General blockchain attack scenario. There are a few general attack scenarios against a blockchain instance [35]. Attacks in which single nodes flood the blockchain with transactions are possible. Not all these attacks are always applicable. Since we deploy multiple chains with different permissions this will typically affect only a small section of our proposed blockchain network.

Threat 6: Compromised Identity Provider. A compromised Identity Provider would have fatal consequences for the proposed system. An attacker could invalidate the identities of authorized nodes, masquerade as an existing identity, and create new, malicious identities. A countermeasure for this attack would be the utilization of an identity provider with decentralized structures (see also section 4.1).

This high-level analysis is only meant to show that the presented system is also promising with respect to security. We will provide a detailed design and state-of-the-art verification of the cryptographic architecture together with a resilience analysis in case of key compromises in future work.

4 Evaluation

4.1 Case Study: AIS Data Processing in a Maritime Data Space

In 2002 The Automatic Identification System (AIS) was introduced by the IMO SOLAS Agreement. It facilitates the submission of dynamic and static vessel properties (such as position, speed, destination, size, etc.) by vessels via VHF. Several publications make use of historical AIS data in their research process (see e.g. [36] for AIS-based collision risk analysis, [37] for anomaly detection or [38] for route prediction). Even though AIS is not encrypted and theoretically can be recorded by anyone, it requires powerful equipment to record it for larger areas. For this reason, it is often the case, that AIS data for a specific area needs to be exchanged between the recorder and users of the data, which is a typical business case in the maritime data domain (see e.g. MarineTraffic[1]). As a representative scientific workflow for a maritime data space, we use the process of creating a heat map for vessel traffic density

[1] https://www.marinetraffic.com/en/p/ais-historical-data

for the German Bight from raw AIS data (in the NMEA0183 format, see [39]). Traffic density heat maps can especially be useful for traffic optimization and could, for instance, be used as a data source by Vessel Traffic Services (VTS) for optimizing traffic efficiency. For implementing this process, the e-Maritime Integrated Reference Platform (eMIR) [40], which offers an open modular research and test environment used for scientific analysis of maritime systems and data generation with a variety of maritime data sources, is utilized. The workflow is illustrated in **Fig. 5**: eMIR[2] uses a network of distributed sensors to continuously record the raw AIS messages from vessels in the German Bight (approx. 2.700.000 messages per day). This data is persisted in a PostgreSQL database. As AIS data contains public, unique ship identifiers (MMSI and IMO numbers), which must be treated as personal information, the data is anonymized by replacing the MMSI and IMO with a hashed value. All of this is done by an arbitrary Organization A (Data Owner / Data Provider) to create the source data set, which is distributed via the data space. In our example, Organization B (Data Consumer/ Data Provider) uses this data set as a starting point for enhancing the data for further processing: Faulty entries are removed, and unnecessary attributes are filtered. A data science researcher from Organization C (Data Consumer) can then use the prepared data set to create a heat map for his research project.

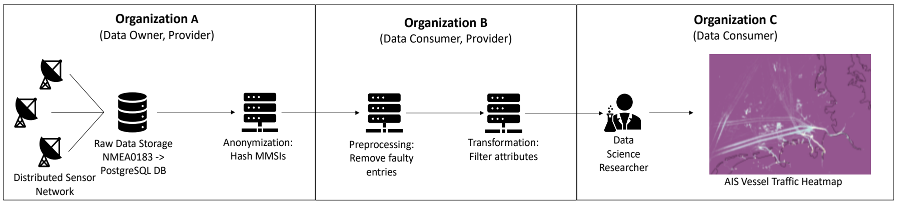
Organization A
(Data Owner, Provider)
Distributed Sensor Network
Raw Data Storage NMEA0183 -> PostgreSQL DB
Anonymization: Hash MMSIs
Organization B
(Data Consumer, Provider)
Preprocessing: Remove faulty entries
Transformation: Filter attributes
Organization C
(Data Consumer)
Data Science Researcher
AIS Vessel Traffic Heatmap

Fig. 5. Workflow for creating a Vessel Traffic Heatmap in a maritime data space.

For the realization, we instantiated the proposed concept in section 3 in the eMIR data space setup to operate with its available resources. **Fig. 6** illustrates a technical overview of the realization for this case study: The Maritime Connectivity Platform[3] (MCP) is a platform to support the implementation of digital services for supporting the maritime industry and was selected as an identity provider as it features a decentralized management of identities (cf. Threat 6 in section 3.4). For the realization of the blockchain network, R3's Corda Open Source[4] was selected as it provides interfaces to model real-world relations independently from crypto-currency features. Furthermore, it was optimized to run as a permissioned blockchain network and allows easy integration of blockchain peers and data processing logic. The process for creating the heatmap is completely automated and can for example be executed daily for tracking changes in traffic. In our case, we worked with data sets of 1.000.000 AIS data points. The documentation of the workflow in our system starts with the internal data acquisition, conversion, and anonymization of the data by Organization A. The data set

[2] https://www.emaritime.de/
[3] https://maritimeconnectivity.net/
[4] https://www.corda.net/

is then made available to the data space for further processing by Organization B and C.

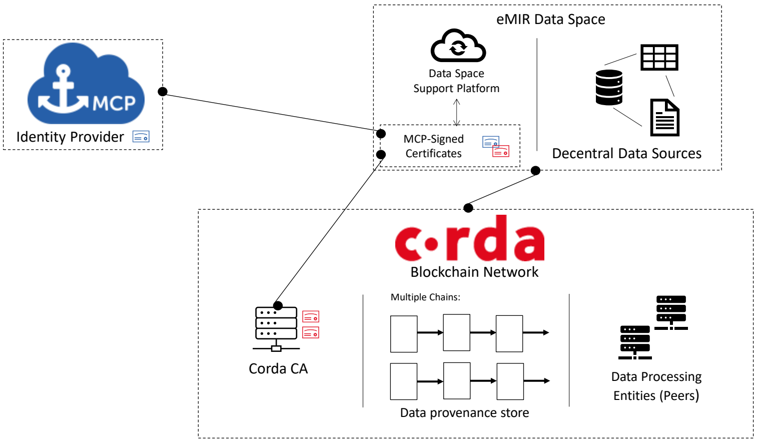

Fig. 6. Realization of the data provenance information system with Corda OS and the MCP.

Note, that the data acquisition, conversion, and anonymization cannot be observed by other participants of the data space. However, as the data is bound to the transactions by its hashes and the transactions are signed by their originator, the data and their processing could be revealed later in case of discrepancies to proof the validity of the data provenance information. For a single instance of the workflow, a total of six workflow steps were added to the data provenance system adding ~3KB of provenance data to the blockchain (per million AIS data points). Assuming ~2.7 million AIS data points are recorded per day, this would create ~9KB of provenance data in the blockchain per day, for the presented workflow. Finally, the case study successfully showed the applicability of the proposed concept. It could be seen that our concept can be integrated into existing workflows and the workflow model can be used to represent typical data processing steps. However, the integration of existing infrastructure, such as external identity providers is not always trivial. In our example, MIR keys could not directly be integrated into to the Corda OS framework, as they did not match the cryptographic requirements. For this reason, we authorized the Corda CA certificates making use of the MIR keys for each entity and made this information available via the DSSP for validation of signatures (as shown in **Fig. 6**). Finally, it could be shown that the system can be used to support typical data exchange problems that can be found in the maritime data domain and close the gaps of existing work. Gaining global data coverage for larger areas is very important to the international maritime industry. As this task is often not achievable for a single entity, data needs to be exchanged and analyzed collaboratively.

4.2 Performance Evaluation

We have implemented and evaluated a network of nodes that allows us to consider complete workflows and several validators. We used a typical windows machine (AMD

Ryzen 7 1700 @ 3 GHz, 16 GB RAM) for performance testing. Our focus in this evaluation is on changes in the node-setup, to derive implementation/setup-independent performance insights. Therefore, we mainly used a network of 10 nodes with different workflows and role set-ups (as shown in **Table 1**).

Table 1. Performance evaluation results.[5]

Number of Nodes	Workflow Setup	Avg. Time per Transaction
10	1 Workflow, 1 Validator	1860 ms
10	1 Workflow, 2 Validators	2366 ms
10	1 Workflow, 4 Validators	2801 ms
10	1 Workflow, 6 Validators	3379 ms
10	2 Workflows, 2 Validators each	1364 ms
10	3 Workflows, 1 Validator each	790 ms
5	1 Workflow, 2 Validators	1195 ms

In the first four trials we set up a single workflow and constantly raised the number of validators for that workflow. Consequently, the average time per transaction also significantly increased. This is obviously due to the higher number of nodes that need to communicate to find a consensus on adding a new workflow step. As stated in section 3.4, a higher number of validators increases the security of the system. Finding the right balance of security and performance in terms of validators therefore can be identified as a challenge of the proposed system and could lead to poorly configured systems. Secondly, we investigated how parallel-running workflows affect the performance of the system. With the same number of nodes and validators (cf. rows 3 and 5 of **Table 1**), we already have a 52% faster transaction speed with two parallel workflows. Additionally, we conducted a test with 5 nodes and 2 validators in a single workflow. It was seen that two parallel running workflows almost have the same performance as a single workflow (cf. rows 5 and 7 of **Table 1**). We interpret this as a result of our permissioned approach with multiple chains. Lastly, it could be seen that the 'time per transaction'-measurements for our case study were already relatively high. According to R3 Ltd. [41], this seems to be a general problem of the open source implementation of Corda and is probably not related to our consensus mechanism. Also, due to our blockchain architecture, we do not expect scenarios in which thousands of participants issue transactions at a single blockchain instance. For an increasing number workflows, the system can easily and efficiently be scaled horizontally by adding additional blockchain instances as the results of the performance evaluation could show.

4.3 Extended Example: Setting with Confidential Data Sets

We now extend our case study to illustrate how our design can handle a setting where two participating organizations are competitors and have therefore an interest in keeping some of their data sets confidential. Assume there are three more participating

[5] Our implementation is available under: https://doi.org/10.5281/zenodo.3960262 .

organizations D, E, and F: both, D and E, are companies that specialize in algorithms to optimize AIS data sets for use in ship navigation systems; F is a company that develops ship navigation systems (potential client of both D and E). Moreover, E wishes to provide a demo service, where F could view up to three results of their latest algorithm run on data sets selected by F from the data space. Naturally, E does not wish that competitors such as D have access to the resulting data sets. With many demo data sets publicly available a competitor could at some point be able to reengineer the algorithm. The participants in this example will choose a legal contract (from a set of standardized templates) where every industrial participant is allowed to restrict access to data sets that result from one of their processing steps to other industrial participants of their choice. These in turn are then also bound to confidentiality (by the legal contract). Technically, E will establish a secret key K with F, encrypt the confidential data set under K, and only store it in encrypted form on the data space. The data hash for the provenance blockchain can be computed over the encrypted data set. Hence, the workings of the blockchain system are as usual. Naturally, this is also an example for the case when the validators will neither be able to nor obliged to verify that the processed data set is indeed the result of the transformation described in the data provenance information. Other scenarios where confidential data must be accessed by several participants can make use of multi-party key establishment schemes.

5 Discussion and Conclusion

In this work, we designed a blockchain-based data provenance system and integrated it into an existing data space setup. For this purpose, a scientific workflow-based model was utilized to track each data processing step of an e-Science approach. We used an external Identity Provider and a Proof-of-Authority-like consensus method to secure the blockchain against attacks and make the process of data provenance for scientific workflows more transparent and verifiable. Our multi-chain concept for separating data provenance information by their belonging workflows improved security and performance of the system. However, we identified the need to carefully consider certificate and performance requirements for implementations of our system. Also, the cases of several data sets being merged by a processing step or forks on the chain of data processing need to be evaluated further. We expect that our framework can easily be extended to these cases since it seems the best strategy to generate a new data set in such cases. The scenario of continuously changing data processing entities as permissioned users also should be investigated further as transactions in a blockchain-setup are immutable. In general, we aim to further integrate our concept into the data architecture of the eMIR Platform to provide an overall architecture for collaborative data science and integrated data provenance tracking. As the volume, variety and velocity of available data is increasing continuously, we cannot deny that data provenance management will play an equally important role. Collaborative e-Science has a big impact on today's research methodologies and needs solutions for trust issues and the problem of decentralized data. We expect concepts like ours to fill these gaps in the future and provide a secure and efficient possibility to track data provenance.

References

1. McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D., Barton, D.: Big data: the management revolution. Harvard business review. 90, 60–68 (2012).
2. Simmhan, Y.L., Plale, B., Gannon, D.: A survey of data provenance in e-science. ACM Sigmod Record. 34, 31–36 (2005).
3. Franklin, M., Halevy, A., Maier, D.: From databases to dataspaces: a new abstraction for information management. ACM Sigmod Record. 34, 27–33 (2005).
4. Otto, B., Steinbuß, S., et al.: International Data Space - Reference Architecture Model, https://www.internationaldataspaces.org/ressource-hub/publications-ids/, (2019).
5. Belloum, A., Inda, M.A., Vasunin, D., Korkhov, V., Zhao, Z., Rauwerda, H., Breit, T.M., Bubak, M., Hertzberger, L.O.: Collaborative e-science experiments and scientific workflows. IEEE Internet Computing. 15, 39–47 (2011).
6. Deelman, E., Gannon, D., Shields, M., Taylor, I.: Workflows and e-Science: An overview of workflow system features and capabilities. Future generation computer systems. 25, 528–540 (2009).
7. Taylor, I.J., Deelman, E., Gannon, D.B., Shields, M.: Workflows for e-Science: scientific workflows for grids. Springer (2007).
8. Rousidis, D., Garoufallou, E., Balatsoukas, P., Sicilia, M.-A.: Metadata for Big Data: a preliminary investigation of metadata quality issues in research data repositories. Information services & use. 34, 279–286 (2014).
9. Jayaraman, P.P., Perera, C., Georgakopoulos, D., Dustdar, S., Thakker, D., Ranjan, R.: Analytics-as-a-service in a multi-cloud environment through semantically-enabled hierarchical data processing. Software: Practice and Experience. 47, 1139–1156 (2017).
10. Humphrey, C.: e-Science and the Life Cycle of Research. (2006).
11. Crowston, K., Qin, J.: A capability maturity model for scientific data management: Evidence from the literature. Proceedings of the American Society for Information Science and Technology. 48, 1–9 (2011).
12. Ball, A.: Review of data management lifecycle models. University of Bath, IDMRC (2012).
13. Verbert, K., Manouselis, N., Drachsler, H., Duval, E.: Dataset-driven research to support learning and knowledge analytics. Journal of Educational Technology & Society. 15, 133–148 (2012).
14. Buneman, P., Khanna, S., Wang-Chiew, T.: Why and where: A characterization of data provenance. In: International conference on database theory. pp. 316–330. Springer (2001).
15. Halevy, A., Franklin, M., Maier, D.: Principles of dataspace systems. Presented at the Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (2006).
16. Rettore, P.H., Maia, G., Villas, L.A., Loureiro, A.A.: Vehicular Data Space: The Data Point of View. IEEE Communications Surveys & Tutorials. 21, 2392–2418 (2019).
17. Curry, E., Derguech, W., Hasan, S., Kouroupetroglou, C., ul Hassan, U.: A real-time linked dataspace for the internet of things: enabling “pay-as-you-go” data management in smart environments. Future Generation Computer Systems. 90, 405–422 (2019).
18. Berre, A., Rødseth, Ø.: From digital twin to maritime data space: Transparent ownership and use of ship information. Presented at the September 27 (2018).
19. Shrestha, A.K., Vassileva, J.: Blockchain-based research data sharing framework for incentivizing the data owners. In: International Conference on Blockchain. pp. 259–266. Springer (2018).
20. De Angelis, S., Aniello, L., Baldoni, R., Lombardi, F., Margheri, A., Sassone, V.: PBFT vs proof-of-authority: Applying the CAP theorem to permissioned blockchain. (2018).
21. Zheng, Z., Xie, S., Dai, H., Chen, X., Wang, H.: An overview of blockchain technology: Architecture, consensus, and future trends. Presented at the 2017 IEEE international congress on big data (BigData congress) (2017).

22. Bowers, S.: Scientific workflow, provenance, and data modeling challenges and approaches. Springer (2012).
23. Chen, P., Plale, B., Aktas, M.S.: Temporal representation for scientific data provenance. In: 2012 IEEE 8th International Conference on E-Science. pp. 1–8. IEEE (2012).
24. Bowers, S., McPhillips, T., Ludäscher, B., Cohen, S., Davidson, S.B.: A model for user-oriented data provenance in pipelined scientific workflows. In: International Provenance and Annotation Workshop. pp. 133–147. Springer (2006).
25. Bertino, E., Ghinita, G., Kantarcioglu, M., Nguyen, D., Park, J., Sandhu, R., Sultana, S., Thuraisingham, B., Xu, S.: A roadmap for privacy-enhanced secure data provenance. J Intell Inf Syst. 43, 481–501 (2014). https://doi.org/10.1007/s10844-014-0322-7.
26. Hasan, R., Sion, R., Winslett, M.: The Case of the Fake Picasso: Preventing History Forgery with Secure Provenance. (2009).
27. Ramachandran, A., Kantarcioglu, D.: Using blockchain and smart contracts for secure data provenance management. arXiv preprint arXiv:1709.10000. (2017).
28. Liang, X., Shetty, S., Tosh, D., Kamhoua, C., Kwiat, K., Njilla, L.: ProvChain: A Blockchain-Based Data Provenance Architecture in Cloud Environment with Enhanced Privacy and Availability. Presented at the May 16 (2017).
29. Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J.: The open provenance model core specification (v1. 1). Future generation computer systems. 27, 743–756 (2011).
30. Chen, W., Liang, X., Li, J., Qin, H., Mu, Y., Wang, J.: Blockchain Based Provenance Sharing of Scientific Workflows. In: 2018 IEEE International Conference on Big Data (Big Data). pp. 3814–3820 (2018). https://doi.org/10.1109/BigData.2018.8622237.
31. Neisse, R., Steri, G., Nai-Fovino, I.: A blockchain-based approach for data accountability and provenance tracking. In: Proceedings of the 12th International Conference on Availability, Reliability and Security. pp. 1–10 (2017).
32. Tosh, D., Shetty, S., Liang, X., Kamhoua, C., Njilla, L.: Consensus protocols for blockchain-based data provenance: Challenges and opportunities. Presented at the October 1 (2017). https://doi.org/10.1109/UEMCON.2017.8249088.
33. Alharby, M., van Moorsel, A.: Blockchain Based Smart Contracts : A Systematic Mapping Study. Presented at the August 26 (2017). https://doi.org/10.5121/csit.2017.71011.
34. Zheng, Z., Xie, S., Dai, H.-N., Chen, X., Wang, H.: Blockchain challenges and opportunities: A survey. International Journal of Web and Grid Services. 14, 352–375 (2018).
35. Xu, J.J.: Are blockchains immune to all malicious attacks? Financial Innovation. 2, 25 (2016). https://doi.org/10.1186/s40854-016-0046-5.
36. Silveira, P., Teixeira, A., Soares, C.G.: Use of AIS data to characterise marine traffic patterns and ship collision risk off the coast of Portugal. The Journal of Navigation. 66, 879–898 (2013).
37. Ristic, B., La Scala, B., Morelande, M., Gordon, N.: Statistical analysis of motion patterns in AIS data: Anomaly detection and motion prediction. In: information fusion, 2008 11th international conference on. pp. 1–7. IEEE (2008).
38. Pallotta, G., Vespe, M., Bryan, K.: Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction. Entropy. 15, 2218–2245 (2013).
39. Langley, R.B.: NMEA 0183: A GPS receiver interface standard. GPS world. 6, (1995).
40. Rüssmeier, N., Lamm, A., Hahn, A.: A Generic Testbed for Simulation and Physical Based Testing of Maritime Cyber-Physical System of Systems. Presented at the November 14 (2019).
41. R3 Ltd.: Corda - Sizing and performance, https://docs.corda.net/docs/corda-enterprise/3.3/sizing-and-performance.html, last accessed 2020/11/05.