Secure Provenance of Electronic Records Based on Blockchain

Qirun Wang[1], Fujian Zhu[2], Sai Ji[2] and Yongjun Ren[2, *]

Abstract: At present, the provenance of electronic records is stored centrally. The centralized way of information storage has huge risks. Whether the database itself is destroyed or the communication between the central database and the external interruption occurs, the provenance information of the stored electronic records will not play its role. At the same time, uncertainties such as fires and earthquakes will also pose a potential threat to centralized databases. Moreover, the existing security provenance model is not specifically designed for electronic records. In this paper, a security provenance model of electronic records is constructed based on PREMIS and METS. Firstly, this paper analyses the security requirements of the provenance information of electronic records. Then, based on the characteristics of blockchain decentralization, and combined with coding theory, a distributed secure provenance guarantees technology of electronic records is constructed, which ensures the authenticity, integrity, confidentiality and reliability of the provenance information.

Keywords: Secure provenance, electronic records, blockchain.

1 Introduction

With the continuous development of social informatization, a large number of electronic records have been produced by various kinds of equipment and software systems. The electronic records are shared and merged among different application systems and organizations through the network [Das, Zeadally and He (2018); Li, Liu, Wu et al. (2018); Wang, Cao, Li et al. (2017)]. Users must verify the credibility of electronic records before making critical decisions, which often requires validation of provenance, version, processing and production methods, and qualification of electronic records [Ren, Qi, Cheng et al. (2020); Wu, Luk, Holder et al. (2019)]. Therefore, over the past decades, the concept of provenance for electronic records has been proposed by academia. World Wide Web Consortium (W3C) believes that the provenance of electronic record is records describing the entity and the processes involved in its production and delivery. The provenance records the ownership of data throughout its life cycle and the processes it undergoes, including storage, processing, and ultimately deletion, utilization or archiving

[1] School of Engineering and Technology, University of Hertfordshire, Hertford, UK.

[2] School of Computer and Software, Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science & Technology, Nanjing, 210044, China.

[*] Corresponding Author: Yongjun Ren. Email: renyj100@126.com.

Received: 24 March 2020; Accepted: 15 May 2020.

CMC. doi:10.32604/cmc.2020.07366 www.techscience.com/journal/cmc

[Factor, Henis, Naor et al. (2009)]. The provenance-aware system produces, stores, processes and disseminates provenance data, allowing people to use provenance to verify the credibility of data, trace the causes of data errors and responsible persons, and achieve automatic reproduction of experimental results in fields such as scientific workflow, business intelligence, health care, etc. [Liang, Shetty, Tosh et al. (2019); Montecchi, Plangger and Etter (2019); Šumilo, Nichols, Ryan et al. (2019); Teke and Tarhan (2019)].

The provenance itself is a metadata describing the historical information of data, but it has the particularity different from ordinary data and other metadata [Castro, Pistoia and Ponzo (2019)]. As a data object, its provenance is semantically immutable. The provenance is non-isolated, and a historical version of a data entity or an instance of a processing process often fails to describe the nature of the provenance information alone. The provenance is usually presented as complex digraphs consisting of causal dependencies between entities and entities. The provenance is of great significance in the organization and management of electronic records long-term preservation system. The primary goal of the long-term preservation system of electronic records is to ensure the authenticity, comprehensibility and accessibility of provenance. If the provenance loses its authenticity, comprehensibility and accessibility will be out of the question. The provenance records the various operations of the long-term preservation system on the electronic records, which can provide evidence for authenticity judgment, and show the changes of electronic records. They will prove whether the electronic records are authentic [Lehman, O'Connor, Kovacs et al. (2019)].

At present, the provenance of electronic records is stored centrally. The centralized way of information storage has huge risks [Huang, Chen, Li et al. (2014)]. Whether the database itself is destroyed or the communication between the central database and the external interruption occurs, the provenance information of the stored electronic records will not play its role. At the same time, uncertainties such as fires and earthquakes will also pose a potential threat to centralized databases. Moreover, the existing security provenance model is not specifically designed for electronic records.

2 Related work

2.1 Development of provenance

Provenance has been the subject of intense interest from a number of universities and research institutions, as detailed below. Bao et al. [Bao, Cohen-Boulakia, Davidson et al. (2009)] developed a tight and effective accessibility labeling scheme to answer questions about workflow traceability running under specified instructions. This labeling scheme is optimal in a sense because it uses logarithmic length, online time and can answer any general time accessibility question. Zhou et al proposed ExSpan (extensible perception of network analysis system) design and application [Zhou, Cronin and Loo (2007)]. ExSpan is traced in the distributed environment can effectively network platform for the gm, extensible framework, and defines a distributed model for the network resource storage, explained by the concept of data source existing in the network of various states, and provides a kind of multi-functional network mechanism. Karvounarakis et al. [Karvounarakis, Ives and Tannen (2010)] proposed a kind of Provenance Query Language based on tuples and semi-circular traceability, which can solve problems related to traceability storage, maintenance and Query.

Dai et al. [Dai, Wang and Zhang (2010)] comprehensively and systematically summarized the data traceability, and introduced the basic research of data traceability and two typical formalized models in the open environment. Wang et al. [Wang, Peng, Luo et al. (2006)] studied the data tracking model in the scientific workflow service framework of object proxy database, and proposed a data tracking method of bidirectional pointer mechanism.

Currently, more and more attention has been paid to the development of data traceability, and International Provenance and Annotation Workshop, Workshop on Data Derivation and Provenance, Workshop on the Theory and Practice of Provenance and other directly related academic conferences have been initiated.

2.2 Provenance model

The traceability model mainly includes stream traceability information model, time-value center traceability model, four-dimensional traceability model, open provenance model, and provenir model.

The steam traceability information model is composed of 6 related entities, mainly including flow entities (change event entities, metadata entities and query input entities) and query entities (change event entities, receive query input entities, including metadata entities). Entities are closely related to each other. Through this close relationship, data traceability can be inferred based on data traceability time.

The Time-Value Centric (TVC) provenance proposed by Marion B, also known as a simple, but useful, hybrid provenance model [Wang, Blount, Davis et al. (2007)]. Since past traceability models, whether annotation-based or process-based, are used in transaction-oriented systems, they are not suitable for high-volume specific needs and continuous medical flows. Therefore, the TVC model that supports the characteristics of data sources in the medical field is proposed to specifically deal with the source information of medical event stream. The sequence of medical events and traces of the original data are inferred from the timestamp and stream ID Numbers in the data.

The four-dimensional traceability model was proposed by Simmhan et al. [Simmhan, Plale and Gannon (2008)] This model views traceability as a discrete set of activities that occur throughout the workflow life cycle and consist of four dimensions (time, space, layer, and data flow distribution). The 4d traceability model differentiates multiple activities in different activity layers in the annotation chain through the time dimension, and then captures the workflow traceability and data traceability that supports workflow execution by tracking the activities in different workflow components.

Open Provenance Model (OPM). At the first International Provenance and Annotation Workshop, participants developed some common ideas about the description of data traceability and proposed an original data model. Later, Moreau et al. [Moreau, Clifford, Freire et al. (2011)] sorted out. The main ideas of the conference and published an article entitled “The Open Provenance Model”. The model mentioned in this paper basically forms the industry information exchange standard and defines the specific format and protocol.

OPM is designed to provide interchangeable traceability information to different systems and allow developers to create and share tools that manipulate the model. OPM also defines traceability from a technical perspective, supports traceability of anything (not just for computer systems), and allows multilevel descriptions to coexist simultaneously. OPM begins by defining three core concepts: Artifact, Process, and Agent. Artifact is used to refer to a state that can be a physical object or a digital expression in a computer system. Process refers to one or a series of actions caused by an Artifact. Agent refers to the catalyst of Process, which is used to promote, control and influence the execution of Process. In addition, OPM has introduced the concept of Role, where a Process can produce multiple artifacts that can have different roles. Taking a division operation as an example, Agent is a calculator (or operation program) and Process is a division operation. Two artifacts involved in the operation belong to the roles of divisor and dividend respectively. The result of the operation also contains two artifacts, and they belong to the roles of quotient and remainder respectively.

2.3 Provenir model

In 2008, the university of Wright, Sahoo et al. [Sahoo, Barga, Goldstein et al. (2018)] presented the Provenir model at the second IPAW conference. Provenir comes from French meaning “to come from”. The model is logically described using the W3C standard. The Provenir model takes into account the details of both the database and workflow fields and forms a complete system from the aspects of model, storage and application. Provenir model effectively solves the problem of data traceability storage by means of materialized view.

Provenir Ontology was given by the Provenir model [Sahoo and Sheth (2009)]. Provenir Ontology defines three main classes as the basic components of the model, which are Data, Process and Agent. The Data category represents raw materials, intermediate materials, final products in scientific experiments and some parameters that affect the implementation of scientific processes. The meaning of Process and Agent is similar to that of OPM. However, Provenir Ontology emphasizes two concepts, namely Occurrent and Continuant. Occurrent refers to the contingency properties that change with time, while Continuant refers to the persistent ones that do not change with time. Provenir Ontology believes that Process is Occurrent and Data and Agent are Continuant.

3 Problem statement: security requirements of provenance

The provenance itself is a kind of metadata that describes data history information, but it is different from ordinary data and other metadata [Xie, Dan, Tan et al. (2013)]. As the history of data object, the provenance is semantically immutable. The provenance is non-isolated. A historical version or process instance of a data entity often cannot describe the nature of the provenance information separately. The provenance usually presents a complex directed graph composed of entities and causal dependencies between entities [Liang, Shetty, Tosh et al. (2019); Montecchi, Plangger and Etter (2019)]. The characteristics of the provenance bring new security challenges, resulting that the traditional security model and mechanism cannot meet the security requirements of the provenance. Provenance security has become one of the key bottlenecks restricting the application of provenance and the spread of provenance awareness system.

The particularity of provenance brings new problems and challenges. In this section, the security requirements of provenance are analyzed based on information security theory. They are shown in Fig. 1.

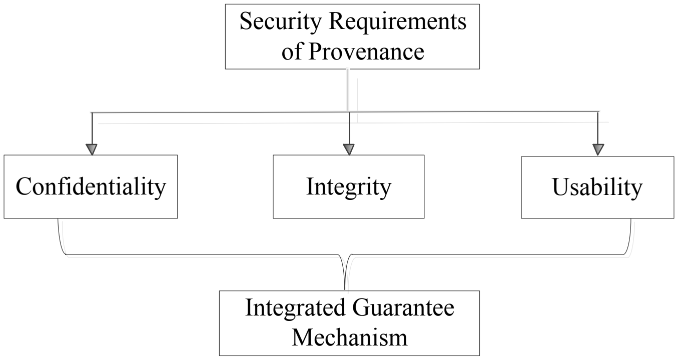
Security Requirements of Provenance
Confidentiality
Integrity
Usability
Integrated Guarantee Mechanism

Figure 1: Essential security requirements of provenance

Data integrity refers to ensuring that information or data is not tampered with un-authorized entity. And the changes can be quickly detected after tampering in the process of transmitting, storing information or data. The integrity of provenance involves not only the integrity of the provenance entity but also the integrity of the dependencies between the provenance entities. On the one hand, the provenance entity must be semantically immutable and must not tamper with or delete existing records or introduce new falsified records. On the other hand, the dependency between related provenance entities cannot be ignored under the premise of ensuring the right, and its causal order cannot be reversed and confused.

Data confidentiality is an index that is not compromised to unauthorized users, entities, or processes. The confidentiality of provenance requires preventing all sensitive information related of provenance from being illegally obtained or used. Because the entities, dependencies and origins sub-graphs of the provenance map may be sensitive. It is worth noting that there is no necessary connection between the confidentiality of data and the confidentiality of its provenance [Syalim, Nishide and Sakurai (2010)]. On the one hand, the provenance of public data may be sensitive. For example, the content of the announcement issued by the government is public, but the provenance information such as the author and draft of the discussion may be sensitive. On the other hand, the provenance of sensitive data may be open. For example, although the postman knows the origin information of the letter, such as the sender, the addressee and the transport process, he does not know the content of the letter.

The usability of data means that authorized users can obtain the data when they need it, which means that the computing system that processes the data, the relevant security protection mechanism and the data transmission mechanism must be correct. The usability of provenance includes two aspects: usability of provenance related infrastructure and usability of provenance information. The usability of provenance infrastructure is similar to the usability of common software systems. This paper discusses the usability of provenance information, that is, the degree to which the provenance information is obtained by users and meets their business needs after various processing. In fact, in order to answer the user's provenance accurately, the provenance query result not only needs to contain most relevant information, but also cannot contain too much redundant information; otherwise it will affect the usability of the provenance.

4 Preliminaries

4.1 Preservation metadata: implementation strategies (PREMIS)

In March 2000, Online Computer Library Center (OCLC) and Research Library Group (RLG) jointly initiated and created a working group to develop the infrastructure for the digital storage domain. The working group is composed of prominent experts in the digital preservation field and a wide range of institutional and geographical backgrounds. The primary responsibility of the working group is to bring together their expertise and experience to develop a metadata storage framework that can be applied to a wide range of digital storage practices. The working group published a document on its official website, i.e. Preservation Metadata for Digital Objects: A Review of the State of the Art. It defined and discussed the concept of preservation metadata, and reviewed the current theories and practices of preservation metadata. Moreover, it identified the starting point of the common foundation of the field.

In 2003, OCLC and RLG launched the project after listening to the experts of the preservation metadata, i.e., Preservation Metadata: Implementation Strategies (PREMIS). The main objective of the project is to pay attention to the implementation of metadata preservation in practice on the basis of metadata preservation framework, and propose specific guidance scheme for metadata preservation in the long-term preservation of digital resources. In May 2005, Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group was released. In November 2015, the working group released PREMIS Data Dictionary for Preservation Metadata version 3.0. At present, PREMIS has become the factual standard for metadata storage in the world, which has been accepted and adopted by more and more existing or under construction long-term storage systems.

Based on the OAIS reference model, PREMIS defines its own data model from the perspective of implementation, which can be regarded as the translation framework from the OAIS conceptual model to the executable unit. PREMIS defines a subset of all saved management metadata (as shown in Fig. 2), involving digital object description metadata, saving metadata, business rules, technical metadata in specific formats and other aspects. These data correspond to different parts of the AIP that hold description information and present information respectively.

PREMIS is a general data model for considering and organizing metadata storage. It can be used as a checklist of core metadata in the storage system to guide local implementation, or as a standard for packet exchange between the storage systems. However, PREMIS is not a straightforward solution that needs to be instantiated into metadata elements in the system. It simply defines what the system should know and be able to export to other systems. PREMIS recommends an XML representation for data exchange and provides a simple XML schema directly corresponding to the data dictionary to describe objects, events, actors, and power declarations.

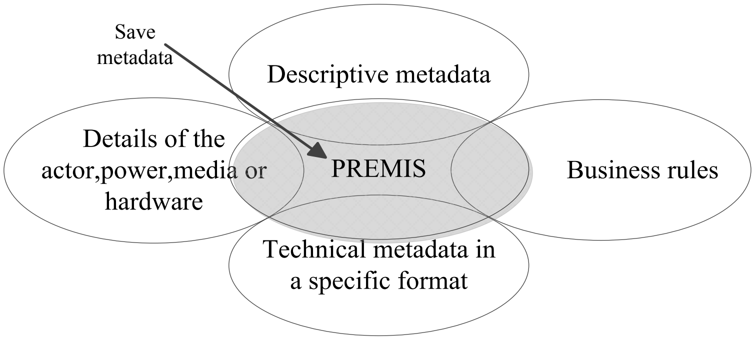
Save metadata
Descriptive metadata
Details of the actor,power,media or hardware
PREMIS
Business rules
Technical metadata in a specific format

Figure 2: Relationship between PREMIS and long-term storage metadata

4.2 Metadata encoding and transmission standard (METS)

Metadata Encoding and Transmission Standard (METS) is the most influential and widely used metadata packaging method of electronic records in the international field. According to a survey conducted in the field of long-term preservation abroad, 64 percent of libraries, 42 percent of archives and 35 percent of other types of institutions are undergoing METS packaging.

METS is a framework for storing all associated metadata of digital objects. Therefore, it can submission information package (SIP), dissemination information package (DIP) in OAIS. Thus, it is archival information package (AIP) crucially. METS standard provides the overall framework scheme with great flexibility. It can completely encapsulate digital objects together and is compatible with a variety of metadata standards. Be platform-and software-independent. As shown in the following Fig. 3. METS contains four major components.

(1) A file library containing all digital object files (such as image files, text, video, or audio files).

(2) Manage metadata parts (such as document-related technical information, rights management information, object source information and data source information).

(3) Descriptive metadata section (including bibliographic information and any information that can criticize the intellectual property value of the object content).

(4) A structure diagram that illustrates the interrelationships between the components of a product in a hierarchical manner, thereby allowing users to navigate by the components.

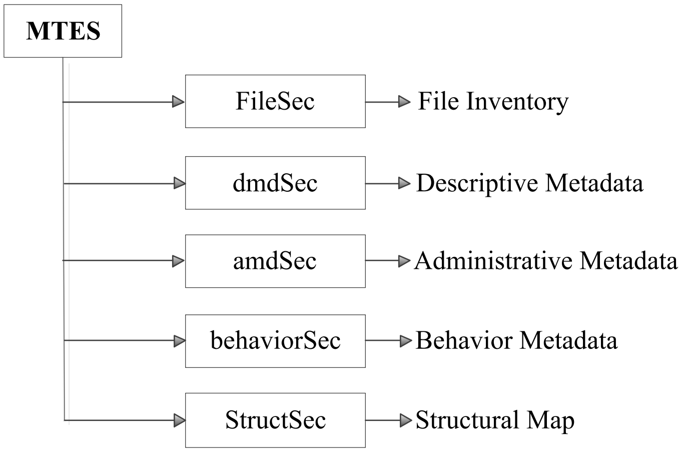

Figure 3: METS file structure

5 Our scheme

Due to the change of technology and information environment, as well as the existence of black interest chain, the provenance data are more vulnerable to damage or even loss. Therefore, the provenance preservation of electronic records is very important to preserve the original data. In the section, the security provenance of electronic record is studied. Based on PREMIS and MEST standards, and using blockchain [Lin, He, Huang et al. (2018); Ren, Leng, Cheng et al. (2019); Ren, Zhu, Sharma et al. (2020)], the provenance protection model and integrated safeguard technology of electronic record are proposed.

5.1 Secure provenance model

Provenance information is necessary for the long-term use and preservation of electronic record. PREMIS is a widely accepted standard for digital preservation. Provenance information is an important part of it. PREMIS is defined as a general and high-level provenance standard, lacking specific means of implementation. Therefore, this paper combines PREMIS with METS to realize the long-term preservation of provenance information for electronic record based on blockchain.

Metadata lifetime is an important factor of metadata long-term preservation. If metadata is to be preserved for a long time, the provenance information must be preserved for a long time. Therefore, provenance should be recorded and managed completely. On the one hand, provenance is a digital object; on the other hand, provenance is a logical data entity independent of any specific physical representation. PREMIS is a widely accepted standard for the life of digital objects in the industry. However, there is no mature model or longevity standard that regards metadata as a logical data entity. Combining PREMIS and METS, this paper proposes a security provenance model from the perspective of provenance lifetime.

According to the nature of provenance, there is also data about provenance in provenance information. It describes the provenance from the perspective of structure and semantic definition. The provenance instance is created as a digital or logical data instance of the provenance, which are represented as independent digital objects or embedded in digital objects. Provenance is important information related to metadata lifetime. In the following Fig. 4, we combine PREMIS with METS to preserve provenance information.

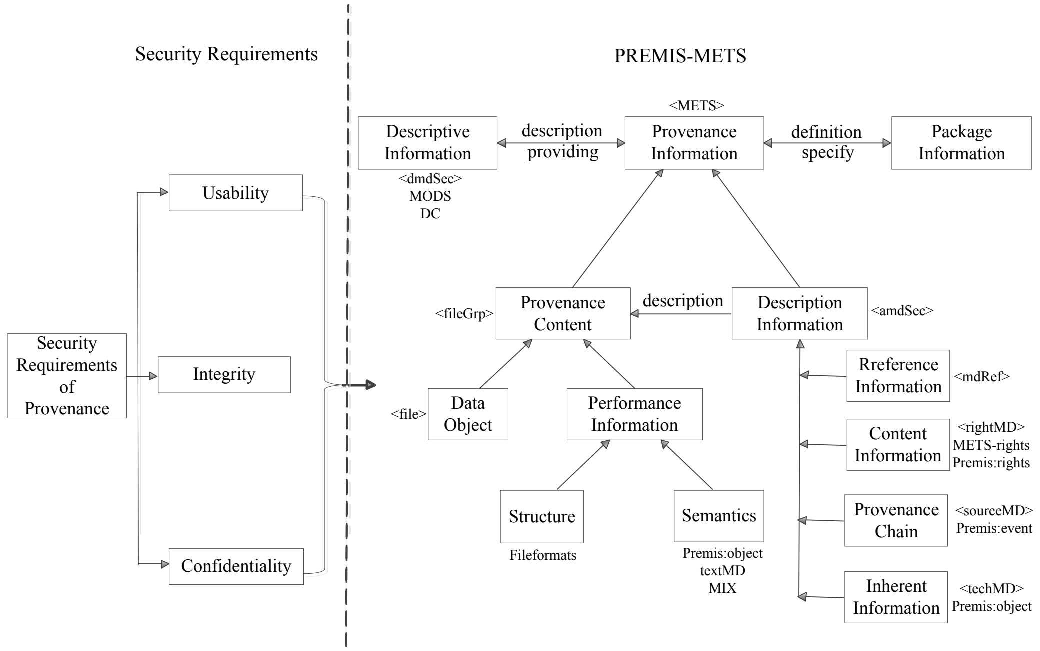

Figure 4: Secure provenance model of electronic record

METS does not define specific descriptive and managerial provenance schemes. But it allows the use of externally developed provenance schemes for its two defined provenance components. Therefore, the description information <amdSec> of METS can be implemented by PREMIS. The <rightMI>, <digiProvMD>/<SourceMI> and <techMD> of <amdSec> can be realized by the <rights>, <event> and <object> of PREMIS. The <dmdSec> can be implemented by extension of MODS and DC data.

5.2 Provenance preservation based on PREMIS-METS and blockchain

Decentralization is an important feature of blockchain technology. As new information technology, timestamps and cryptography technology are utilized to record transactions in blockchain. These data blocks are composed of time series. The consensus mechanism is used to store the data in the distributed database. The generated data record is unique, permanent and irreversible. Therefore, credible transactions can be achieved without relying on any central agency. This distributed data storage structure determines the decentralization of blockchain.

5.2.1 Timestamp guarantees the authenticity of provenance

Blockchain is data chains consisting of countless blocks connected from beginning to end. Each block is automatically timestamped when it is generated. Timestamp can be used as a key parameter of proof of existence, which can confirm that certain data must exist at a certain time. Using the timestamp characteristic of blockchain, it can solve the problem of traceability of electronic records and anti-counterfeiting of provenance.

At first, the timestamp feature helps to identify the authenticity and ownership of provenance. The interactivity of electronic records enables users to fabricate, modify and forward information at will. After many times of forwarding, it is difficult to distinguish who owns the information. Moreover, it is difficult to distinguish whether it is true or not. Each user on the blockchain automatically stamps a timestamp when he publishes an original message. Every forwarding, modification and other operations by other user are also recorded in the blockchain, thus forming a time-series data link. The authenticator can trace the origin of the original content and its owner in the data link, and form a real and complete the provenance of electronic records.

Secondly, timestamp can ensure that electronic records cannot be tampered with or forged after being saved, and improve the stability of electronic records. Electronic records, unlike paper files, can be easily tampered with. The timestamp determines the openness and transparency of blockchain. It is like multiple duplicate invoices or receipts, modifying or destroying single data cannot change the content of other electronic records. That is to say, every electronic record is in the process of monitoring and reviewing the whole network. Tampering with data and operating records on block chains will cost unthinkable costs.

5.2.2 Entire data link ensuring the integrity of provenance

Blockchain is a data structure that combines data blocks in a chain manner. The front-end of each data block contains the compression value of the transaction information of the previous block. And a long chain of blocks and blocks is formed. All of the blocks contain the reference structure of the previous block. It allows the existing set of blocks to form a dynamic global data link. It is shown in the Fig. 5.

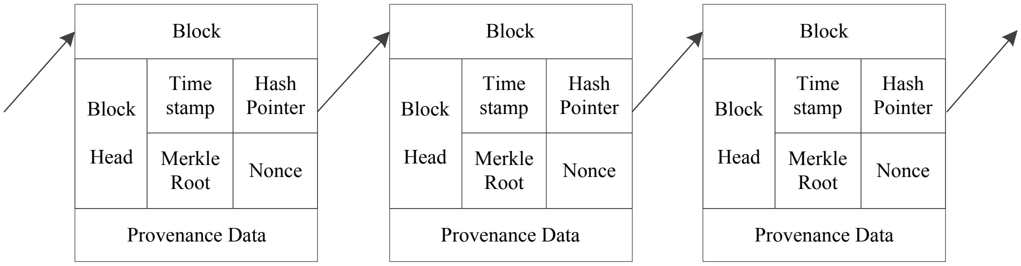

Figure 5: Data links ensure the integrity of provenance

There is no doubt that the content of electronic records itself is the most valuable part of data, with a certain value of evidence. Blockchain dynamically integrates published content with provenance into a complete data entity. Each data entity can be considered as a founding block in the blockchain, in which such as creator, release date, geographic data are stored. The interactive operation around the information produces other blocks. These blocks are arranged in time series after the original blocks, forming a one-way, irreversible complete data link. Therefore, using blocks as the basic unit of information acquisition can solve the problem of separation of provenance, operation data and publishing information data. Therefore, the integrity of provenance for electronic records is technically solved.

5.2.3 Asymmetric encryption guarantee the confidentiality of provenance

Asymmetric encryption is a new key protocol that allows communication parties to exchange information on insecure networks. Thus, it solves the problems of public transmission of information and key management. Asymmetric encryption uses two different keys, public key and private key, to encrypt and decrypt. Data is usually encrypted with a public key, and only the corresponding private key can be decrypted. Correspondingly, only the public key can verify the signature if the data is signed with the private key.

Provenance information contains a large amount of information and constitutes an important part of electronic records. It is necessary to archive and preserve it. Moreover, much provenance information itself involves personal privacy and sensitive information of the unit. Asymmetric encryption algorithm can effectively solve the problem, using public and private keys to encrypt and decrypt data in block chain. Archiving subjects such as archives and libraries can obtain the public key of provenance, and owners of provenance can keep the private key. The archiving subject encrypts each data on the blockchain with public key, and only the owner of the private key can decrypt it. Similarly, the owner of provenance can sign his own information with the private key. Only the archiving subject who has the public key can check the signature. Through asymmetric encryption data of blockchain, provenance is stored on blockchain network nodes in the form of encryption, and form a distributed cloud. Only those who have mastered the secret key can view the provenance. Therefore, when an organization or individual sends a provenance requirement, it can use the private key to authenticate the identity. If the private key held by the archiving subject cannot decrypt the data, it means that the archiving subject does not have archiving authority over the provenance, as shown in the following Fig. 6.

The provenance information is designed and operated according to the de-centralized mode at the bottom of the technology. At present, a large number of electronic records are produced in the mass-dominated “point-to-point” information dissemination. And ordinary people can voice freely through the Internet. Blockchain technology and electronic records have the characteristics of de-centralization, which provides a natural application advantage for blockchain technology in the protection of the provenance of electronic records. On the basis, the application of blockchain technology in archiving the provenance of electronic records will help to promote the reform of electronic records management mode.

5.2.4 Fountain coding guarantee the usability of provenance

In order to ensure the usability of provenance, fountain code is used to encode provenance information, and the corresponding hash values of each data block are stored to the location of tampered provenance data. The scheme is divided into three stages: encoding, validation and recovery. In the coding stage, the provenance is encoded, and the hash value of provenance is calculated. In the validation phase, users first calculate the tags related to the holding validation, and then use the tags to verify whether the storage node holds its provenance correctly. Once the validation fails, the user will ask the storage node for data recovery. In the recovery stage, the storage node first extracts the valid data and decodes it, then validates the recovered provenance. After validation, it waits for the user to confirm the decoded provenance, and finally recovers the remaining tampered provenance.

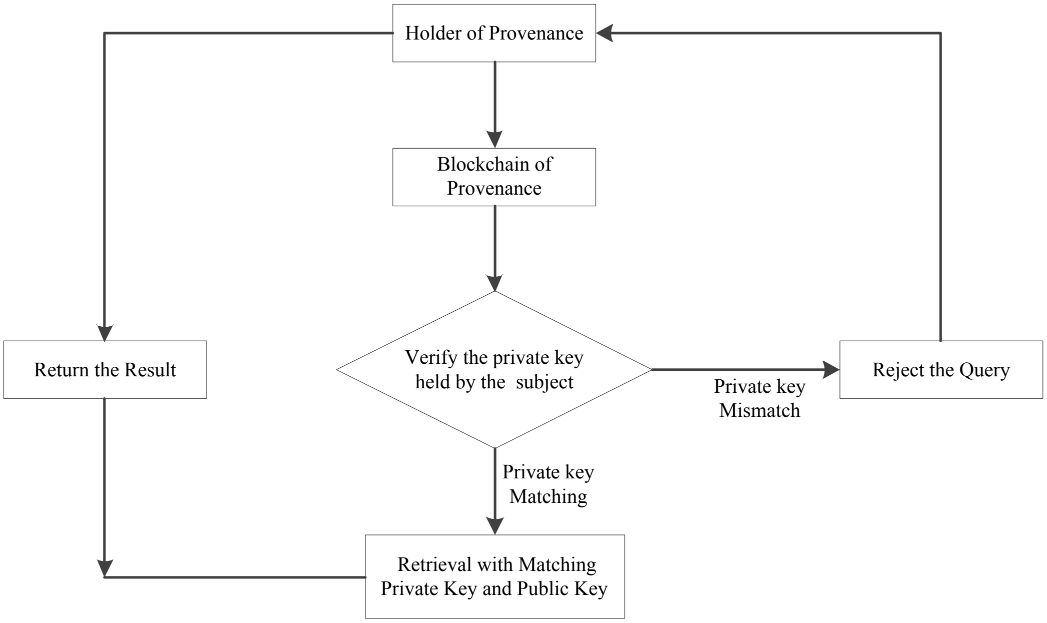

Figure 6: Provenance query based on blockchain

Provenance coding

Firstly, the provenance D is divided into data blocks $D_1, D_2, \cdots D_k$. And then D_i is coded by the encoding matrix G and the symbol C_i $(C_i \in C, C_i = D_i G)$ is achieved. G is a $k \times n$matrix and consists of $G_1, G_2, \cdots G_n$. When decoding, k symbols are randomly selected from C and combined into Q symbols, and the corresponding encoding matrices are selected from G to form the decoding matrix P in sequence. According to$D = QP^{-1}$, the provenance information D can be obtained. If $D_{m\times k}G_{k\times k} = C_{m\times k}$, and G is reversible, $D = CG^{-1}$is achieved. When G is a $k \times n$ matrix, $D_{m\times k}G_{k\times k} = C_{m\times k}, D = (D_1, D_2, \cdots, D_k), D_i = (d_{i1}, d_{i2}, \cdots, d_{im})^T$, $G = (G_1, G_2, \cdots, G_n)$ $G_i = (g_{i1}, g_{i2}, \cdots, g_{ki})^T$, $C = (C_1, C_2, \cdots, C_n), C_i = (c_{i1}, c_{i2}, \cdots, c_{mi})^T$. If$c_{ij} = \sum_{h=1,u=1}^{h=k,u=m} d_{hi}g_{uj}$, k columns are selected from C to form the decoder Q, and the corresponding k columns from G is composed of the decoding matrix $P_{k\times k}$. So, if P is reversible, $D = QP^{-1}$.

According to the Vandermonde Matric, when α_i is unequal and not equal to zero, the square matrix composed of arbitrary k columns is reversible. Let$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)$. When $\alpha_i \neq 0, i \in (1,2,\cdots,n)$ and $\alpha_i \neq \alpha_j (i \neq j)$, the matrix G can be obtained.

$$G = \begin{bmatrix} a_1^1 \, a_2^1 & \cdots & a_n^1 \\ a_1^2 \, a_2^2 & \cdots & a_n^2 \\ & \cdots & \\ a_1^k \, a_2^k & \cdots & a_n^k \end{bmatrix}$$

So, the square matrix consisting of arbitrary k-columns from G is reversible. As long as k is large enough, any large data can be coded. But, with the increase of k, α_i^kwill increase exponentially. If the storage character set is $[0,1,\cdots,255]$, data is divided into eight data blocks, and G is a 8×10 matrix, so the bit width required for each element in G is 1B. Thus, the minimum bit width is10^7. This will result in a waste of storage space. Adny k-order reversibility is selected from G. The probability of extracting a k-order square matrix from a $k \times (k+\varepsilon)$ binary matrix is $1-\delta$, according to the stochastic linear fountain, where $\delta = 2^{-\varepsilon}$. Thus, binary matrix is used as encoding matrix.

Although fountain code has the ability of error detection, it cannot locate the location of tampered provenance itself. Therefore, when some provenance is tampered with, the decoding efficiency is greatly reduced. In this paper, the hash values of each block are calculated to locate the errors in the stored provenance. If the data is tampered with in the data block, the storage node of provenance storage can distinguish the tampered data only by recalculating the hash value of the data and comparing it with the hash value of the stored provenance.

Provenance validation

Firstly, the provenance is encoded by fountain code and divided into n data blocks. And then three secret keys k_1, k_2, k_3 are calculated. Let t is the times of challenge-response, and the number of challenged data blocks every time are s.$s = k + \varepsilon, \varepsilon = -\ln\delta$, $^{\delta}$ is the probability of decoding failure. To validate the provenance label for each response, the encoded provenance data block, hash value and encrypted validation label are stored in the storage nodes. The process of validating provenance labels is as follows.

(1)The two preprocessing keys pk_i, ck_i are computed as the following: $pk_i = AES_{k_1}(i)$,$ck_i = AES_{k_2}(i)$; And the two key are shared with storage storage nodes.

(2) Computing block index using ck_i and pseudo-random function g.

$I_j = g_{ck_i}(j), I_j \in [1,2,\cdots,t]$,$1 \le j \le n$

(3) After computing the tag to be encrypted.

$Tag_i' = hash(pk_i, block(I_1), block(I_2), \cdots block(I_s)), Tag_i = AES_{k_3}(i, Tag_i)$.

The following is the data validation phase. The storage nodes compute the tags as the following.

$Tag_i' = hash(pk_i, block(I_1), block(I_2), \cdots block(I_s)), Tag_i = AES_{k_3}(i, Tag_i)$. And then storage nodes return the tags to the verifier. The verifier decodes the tags returned by the storage node. If decoding succeeds, validation is considered successful; otherwise, the validation fails. In the process of verification, correct verification has no effect on the system. Therefore, this paper will analyze the probability of false verification.

Assuming that part of the provenance is tampered with, and the probability of passing the validation is P_e. Therefore, there is $P_e = (1 - t/n)^s$, where t denotes the number of blocks deleted or tampered with; n is the total number of data blocks, s is the number of verified data blocks. When the ratio of data deletion or tampering is $t/n = 1\%$and$s = 512$,$P_e \le 0.6$. And the larger is s, the smaller the probability of error detection. In our scheme, let $s = k + \varepsilon$. It will not only ensure the efficiency of verification, but also ensure the high probability of restoring the provenance. The scheme also considers the failure of provenance validation. When the provenance validation fails, the storage node restores the provenance.

Tamper detection and recovery

When user validation fails, the feedback of error k is provided immediately to the storage nodes. k denotes the size of the original data block, i.e., the least needed data block for recovery or decoding. The storage nodes immediately use the stored hash values to verify with the data blocks. If the validation passes, it is considered that the block has not been tampered with; otherwise, it is considered that the block has been tampered with. The storage nodes count the number of blocks that have been validated v. If $v < k$, it is considered that it is impossible to recover data; otherwise, tamper detection will be carried out again by using complete blocks.

If the provenance D is perfectly correct before encoding, the corresponding symbol C is also correct. Therefore, the corresponding provenance information tampered with only is considered after encoding. Both the encoding matrix and the symbol can be tampered with. If partly symbol C is tampered with, ΔC is data tampered with, and new symbol C^*can be expressed as $C^* = C + \Delta C$. When data tampering occurs in line i of C, tampering may occur in line i of Q accordingly. Assuming that $\Delta Q$is the tampered information contained in Q, i.e., $Q^* = Q + \Delta Q$, the corresponding decoding process is $D^* = Q^* P^{-1} = (Q + \Delta Q)P^{-1} = D + \Delta D$. $\Delta D$is the tampered information contained in the decoding process. Therefore, the i-th elements of k blocks contained in the decoded information D^* are tampered with. When some data is tampered with in G, there are the following formulas. $P^* = P + \Delta P$, $\Delta P$is tampered information contained in P. Correspondingly, $D^* = Q(P^*)^{-1} = Q(P + \Delta P)^{-1} = D + \Delta D$. When data tampering occurs in both C and G, it is $D^* = Q^*(P^*)^{-1} = (Q + \Delta Q)(P + \Delta P)^{-1} = D + \Delta D$ correspondingly. Therefore, tampering detection is necessary.

The detection principle is as follows. Since the correct check block B_k, which is different from the decoder Q, can be selected randomly from the data, it can be detected by comparing $DG_k$with C_k. If $DG_k = C_k$, there is no tampering in decoded data; otherwise there is tampering. The storage nodes receive the feedback parameter k, which is sent from the user, and calculate the hash of the data block to get the correct set S. And then k data blocks are selected from S for self-checking. If tamper detection is successful, then the hash value of k data blocks can be obtained by calculating the index. Then the index order and hash value are sent to the user. According to the index order, the user calculates the hash values, and verifies them.

In the process of provenance restoration, storage nodes randomly generate new random matrix G and encode$C_{n+i} + DG_{n+i} (1 \leq i \leq n)$. G_j and $C_j$combine and form a new data blocks $B_j = \{G_j, C_j\} (j > n + 1)$, it replaces the tampered data. And then the hash values of data are computed. The new hash values are sent to storage nodes and update old hash values.

6 Conclusion

In this article, we summarize the Development of Provenance and introduce five traceability models that provide ideas for our approach. Unlike previous solutions, we creatively use blockchain technology to construct traceability of data. Our program fully considers the security requirements for electronic record source information. Then, combined with the advantages of blockchain and combined coding theory, a distributed security origin guarantee electronic record technology is constructed to ensure the authenticity, integrity, confidentiality, irreparable modification and traceability of the origin information.

Funding Statement: This work is supported by the NSFC (61772280, 61672290), and the PAPD fund from NUIST. Yongjun Ren is the corresponding author.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Bao, Z.; Cohen-Boulakia, S.; Davidson, S. B.; Eyal, A.; Khanna, S. (2009): Differencing provenance in scientific workflows. *IEEE 25th International Conference on Data Engineering*, pp. 808-819.

Castro, P. C.; Pistoia, M.; Ponzo, J. (2016): Transparently tracking provenance information in distributed data systems. *U.S. Patent.*

Dai, C. F.; Wang, T.; Zhang, P. C. (2010): Survey of data provenance technique. *Jisuanji Yingyong Yanjiu*, vol. 27, no. 9, pp. 3215-3221.

Das, A. K.; Zeadally, S.; He, D. (2018): Taxonomy and analysis of security protocols for Internet of Things. *Future Generation Computer Systems*, vol. 89, pp. 110- 125.

Factor, M.; Henis, E.; Naor, D.; Rabinovici-Cohen, S.; Reshef, P. et al. (2009): Authenticity and provenance in long term digital preservation: modeling and implementation in preservation aware storage. *Workshop on the Theory and Practice of Provenance.*

Huang, X.; Chen, X.; Li, J.; Xiang, Y.; Xu, L. (2014) Further observations on smart-card-based password-authenticated key agreement in distributed systems. *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 7, pp. 1767-1775.

Karvounarakis, G.; Ives, Z. G.; Tannen, V. (2010): Querying data provenance. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 951-962.

Lehman, D. W.; O'Connor, K.; Kovacs, B.; Newman, G. (2019): Authenticity. *Academy of Management Annals*, vol. 13, no. 1, pp. 1-42.

Li, X; Liu, S; Wu, F; Kumari, S; Rodrigues, J. (2018): Privacy preserving data aggregation scheme for mobile edge computing assisted IoT applications. *IEEE Internet of Things Journal*.

Liang, X.; Shetty, S. S.; Tosh, D.; Njilla, L; Kamhoua, C. A.; et al. (2019): ProvChain: blockchain-based cloud data pvenance. *Blockchain for Distributed Systems Security*, pp. 67-94.

Lin, C.; He, D.; Huang, X.; Khan, M. K.; Choo, K. R. (2018): A new transitively closed undirected graph authentication scheme for blockchain-based identity management systems. *IEEE Access*, vol. 6, pp. 28203-28212.

Montecchi, M.; Plangger, K.; Etter, M. (2019): It's real, trust me! Establishing supply chain provenance using blockchain. *Business Horizons*, vol. 62, no. 3, pp. 283-293.

Moreau, L.; Clifford, B.; Freire, J.; Futrelle, J.; Gil, Y. et al. (2011): The open provenance model core specification. *Future Generation Computer Systems*, vol. 27, no. 6, pp. 743-756.

Ren, Y.; Leng, Y.; Cheng, Y.; Wang, J. (2019): Secure data storage based on blockchain and coding in edge computing. *Mathematical Biosciences and Engineering*, vol. 16, no. 4, pp. 1874-1892.

Ren, Y.; Qi, J.; Cheng, Y.; Wang, J.; Alfarraj, O. (2020): Digital continuity guarantee approach of electronic record based on data quality theory. *Computers, Materials & Continua*, vol. 63, no. 3 pp. 1471-1483.

Ren, Y.; Zhu, F.; Sharma, P.; Wang, T.; Wang, J. et al. (2020): Data query mechanism based on hash computing power of blockchain in Internet of Things. *Sensors*, vol. 20, no. 1, pp. 207.

Sahoo, S. S.; Barga, R. S.; Goldstein, J.; Sheth, J. (2018): Provenance algebra and materialized view-based provenance management. *Second International Provenance and Annotation Workshop*, pp. 531-540.

Sahoo, S. S.; Sheth, A. P. (2009): Provenir ontology: Towards a framework for escience provenance management.

https://www.researchgate.net/publication/228611325_Provenir_ontology_Towards_a_Framework_for_eScience_Provenance_Management.

Simmhan, Y. L.; Plale, B.; Gannon, D. (2008): Karma2: Provenance management for data-driven workflows. *International Journal of Web Services Research*, vol. 5, no. 2, pp. 1-22.

Šumilo, D.; Nichols, L.; Ryan, R.; Marshall, T. (2019): Incidence of indications for tonsillectomy and frequency of evidence-based surgery: a 12-year retrospective cohort study of primary care electronic records. *British Journal of General Practice*, vol. 69, no. 678, pp. e33-e41.

Syalim, A.; Nishide, T.; Sakurai, K. (2010): Preserving integrity and confidentiality of a directed acyclic graph model of provenance. *IFIP Annual Conference on Data and Applications Security and Privacy*, pp. 311-318.

Teke, I.; Tarhan, C. (2019): Impacts of electronic record management system on business processes: Manisa Celal Bayar University case. *5th International Management Information Systems Conference.*

Wang, J.; Cao, Y.; Li, B.; Kim, H.; Lee, S. (2017): Particle swarm optimization based clustering algorithm with mobile sink for WSNs. *Future Generation Computer Systems*, vol. 76, pp. 452-457.

Wang, L.; Peng, Z.; Luo, M.; Ji, W.; Huang, Z. (2006): A scientific workflow framework integrated with object deputy model for data provenance. *International Conference on Web-Age Information Management*, pp. 569-580.

Wang, M.; Blount, M.; Davis, J.; Misra, A.; Sow, D. (2007): A time-and-value centric provenance model and architecture for medical event streams. *Proceedings of the 1st ACM SIGMOBILE International Workshop on Systems and Networking Support for Healthcare and Assisted Living Environments*, pp. 95-100.

Wu, C. H. K.; Luk, S. M.; Holder, R. L.; Rodrigues, Z.; Ahmed, F. et al. (2019): Correction: How do paper and electronic records compare for completeness? A three centre study. *Eye*, vol. 32, no. 7, pp. 1232-1236.

Xie, Y.; Dan, F.; Tan, Z.; Chen, L.; Zhou, J. (2013): Experiences building a provenance-based reconstruction system. *Digest APMRC*, pp. 1-6.

Zhou, W.; Cronin, E.; Loo, B. T. (2007): Provenance-aware declarative secure networks. *University of Pennsylvania Department of Computer and Information Science Technical Report*, no. MS-CIS-07-27.