

Article

Research and Development of Blockchain Recordkeeping at the National Archives of Korea

Hosung Wang¹ and Dongmin Yang^{2,*} ¹ National Archives of Korea, Daejeon 35208, Korea; kinghosung@gmail.com² Graduate School of Archives and Records Management, Jeonbuk National University, Jeonju 54896, Korea

* Correspondence: dmyang@jbnu.ac.kr

Abstract: In 2019, the National Archives of Korea (NAK) developed a blockchain recordkeeping platform to conduct R&D on recordkeeping approaches. This paper introduces two types of R&D studies that have been conducted thus far. The first is the use of blockchain transaction audit trail technology to ensure the authenticity of audiovisual archives, i.e., the application of blockchain to a new system. The second uses blockchain technology to verify whether the datasets of numerous information systems built by government agencies are managed without forgery or tampering, i.e., the application of blockchain to an existing system. Government work environments globally are rapidly shifting from paper records to digital. However, the traditional recordkeeping methodology has not adequately kept up with these digital changes. Despite the importance of responding to digital changes by incorporating innovative technologies such as blockchain in recordkeeping practices, it is not easy for most archives to invest funds in experiments on future technologies. Owing to the Korean government's policy of investing in digital transformation, NAK's blockchain recordkeeping platform has been developed, and several R&D tasks are underway. Hopefully, the findings of this study will be shared with archivists around the world who are focusing on the future of recordkeeping.

**Citation:** Wang, H.; Yang, D.Research and Development of
Blockchain Recordkeeping at the
National Archives of Korea.*Computers* **2021**, *10*, 90. <https://doi.org/10.3390/computers10080090>

Academic Editor: Hossain Shahriar

Received: 30 May 2021

Accepted: 20 July 2021

Published: 21 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: blockchain; recordkeeping; audiovisual records; authenticity; integrity; recordkeeping transaction; trustworthiness; business system; dataset

1. Introduction

This study summarizes some of the results of the blockchain recordkeeping R&D project of the National Archives of Korea (NAK), which was conducted from 2019 to 2021. In 2019, NAK developed a blockchain recordkeeping platform using hyperledger fabric. In 2020, R&D was conducted to apply several recordkeeping practices using the platform, and as of 2021, more research has been conducted to verify the integrity of datasets produced by numerous government business systems in Korea.

To understand the rationale for introducing blockchain technology, we first need to comprehend the digital environment of the Korean government. In 2001, the Korean government enacted the Electronic Government Act and began to electronically convert government affairs that were carried out using paper documents. Various paper documents such as government tax, criminal justice, social insurance, and financial documents were converted electronically, and a number of information systems were established to manage them [1]. Since the early 2000s, the Korean government has provided a significant budget to establish various business systems for converting large quantities of paper records into digital versions.

Owing to these efforts, there are now approximately 16,000 information systems used by government agencies engaged in digital conversion of paper records; the types of electronic records produced by these systems are significantly increasing in number [2]. NAK archivists, who are responsible for protecting public records, experienced increasing

difficulty in effectively ensuring the authenticity of these new types of records with traditional record management methods. Therefore, NAK has become interested in innovative technologies, including blockchain technology.

It is the responsibility and duty of the archives to maintain authenticity and deliver authentic records to future generations, regardless of the type of public record. Fixity is the minimum requirement for all electronically produced records to maintain authenticity from production to the end of the life cycle. More specifically, the bitstream of the records must not be altered, except in cases where it is judged that it is no longer possible to maintain the current state of the records and they should be legally converted into another format. Exceptional cases include changes in the preservation format owing to a suspension of the related SW policy, repackaging caused by a change in the packaging method of information, a change in information itself owing to a change in the authenticity verification method, and other factors. NAK has adopted a long-term preservation strategy that combines migration and encapsulation to ensure the authenticity of digital records. The preservation strategy aims to convert digital record objects into PDF and encapsulate them by adding electronic signatures and recording metadata. NAK labels these encapsulation objects as NAK encapsulated objects (NEOs) [3]. Although the conversion of most electronic document types into an NEO is generally not difficult, the conversion of large objects, such as audiovisual records, and a number of objects such as datasets in information systems, is not easy. Thus, NAK has been unable to apply authenticity strategies, such as NEO, to audiovisual records or datasets. In 2020, NAK decided to use the audit trail technology of blockchain to solve the problem of authenticity in audiovisual records. In 2021, NAK decided to verify measures to ensure the integrity by using the audit trail technology of blockchain to monitor transactions in datasets produced in these systems. Based on this decision, NAK has promoted R&D projects that have developed a blockchain system for audiovisual records and a dataset for authenticity.

In this paper, we introduce two R&D cases that have been conducted thus far. The first case is the use of blockchain transaction audit trail technology to ensure the authenticity of audiovisual archives. The second is a case study that uses blockchain technology to verify whether the datasets of numerous information systems built by government agencies are managed without forgery or tampering. In addition to the existing electronic documents, the former proposes a method of maintaining the authenticity suitable for a new type of record, called an audiovisual record, and the latter datasets. Aside from the fact that each R&D case deals with a different type of record, the most different aspect of note is the location and environment in which each record physically exists. Audiovisual records are transferred into and managed by the central archives management system (CAMS) and multimedia asset management (MAM), which are operated by NAK. By contrast, datasets are managed in two major ways according to the appraisal of the records. The first is a method of transferring records to NAK, such as audiovisual records, and the second is a method of self-managing the records by the institution that first produced them. When datasets are transferred to NAK, a newly designed and built system can manage them without considering the system before transferring. Therefore, this study focuses on self-managed datasets. In the method of transferring into NAK, a new system can be designed and built freely. However, in the method of self-management by the origin institution without transferring into NAK, there is a restriction in that it should not affect the information system, including the current datasets, and it should be applicable to various information systems.

The remainder of this paper is organized as follows. Blockchain-based methods for the authenticity of audiovisual records and datasets are presented in Sections 2 and 3, respectively. A discussion on the directions of blockchain technology for records and archive management is presented in Section 4. Finally, Section 5 provides some concluding remarks.

2. Method for Authenticating Audiovisual Records

2.1. Motivation

For the long-term preservation of electronic records, NAK has defined the technological specifications for such formats; it manages these records by converting them into NEOs, which have a format for long-term preservation. NEOs allow the long-term preservation and verification of the authenticity of records by implementing the requirements related to the records, metadata, and digital signatures of files in a single package. However, at present, the target number of records for conversion into a format for long-term preservation is more than 40 million cases annually in terms of standard electronic documents. This has resulted in problems such as the high incidence of conversion errors, including the omission of the essential items of NEO metadata and others, and the requirement of a periodic reconversion. Moreover, it takes a lot of time and requires many computing resources to convert large objects, such as audiovisual record types, into the set format for long-term preservation. NAK decided to link its archive with a blockchain platform and MAM system to verify the authenticity of records without converting them into NEOs by tracking the integrity of audiovisual records.

2.2. Application Concept

ISO 15489 establishes the core concepts and principles for the management of records. According to ISO 15489, records should maintain the characteristics of authenticity, integrity, reliability, and useability throughout the entire management process. To prove that these four characteristics have been maintained in record management processes, metadata used in records are managed along with the records themselves [4]. In [5], the author distinguished three distinct stages in the development process of the blockchain record-keeping concept. In the first stage, the mirror type signifies the method for ensuring the integrity of authentic records by separately establishing a blockchain platform as a cryptographic mirror system relative to the records retained within the existing work system that is currently under operation. The digital record type in the second stage enables the tasks to be carried out together with the management of records within a single system by developing and implementing the new work system itself as a blockchain platform. In the last stage, the tokenized type guarantees the integrity of non-electronic records by attaching to the records and tracing the specific tokens that can be recognized by the blockchain platform [5].

Two concepts were verified in this study. First, we verified the concept of ISO 15489 as a blockchain, which only traces the audiovisual records management history and manages the records with metadata without conducting conversion tasks such as NEOs. Second, we verified that it is possible to configure the blockchain as a cryptographic mirror system of the existing MAM system. This corresponds to the first stage, as determined by Lemieux [5].

2.3. Related Studies: ARCHANGEL Project

The ARCHANGEL project [6] was a research project that was developed in the UK to ensure the integrity of digital records with the participation of three institutions under the initiatives of The National Archives (TNA) in the UK. This project commenced in July 2017, and the final report was published in August 2018. The aim of the project was to assure the integrity, authenticity, and trustworthiness of the records created electronically by utilizing distributed ledger technology (DLT) and machine learning as type of artificial intelligence technology.

As a backdrop to this research, the authors in [6] stated that the trust conferred by the general public to the Archives and Memory Institutions (AMIs) had eroded due to the ease by which forgery and unauthorized modifications to electronic records were conducted owing to advances in technology and the generation of numerous types of composited content. To recover public trust in AMIs and guarantee the integrity of the records, a method utilizing the existing databases as well as a method to utilize a Merkle tree or durable storage provided by private corporations have been taken into consideration;

however, it was concluded that they are difficult to apply owing to several limitations. Unlike in the past, when archives have relied on the products and technologies of certain companies, blockchain presents a completely different paradigm from openness and expandability. The blockchain awards permission to write records in a distributed ledger to only authorized institutions, whereas permission to view the recorded contents after accessing the distributed ledger is granted to every node participating in the blockchain. Moreover, scalability allows the utilization of various tools provided by open-source codes and enables the integrity of records to be assured by multiple parties through a consensus mechanism rather than by a single centralized institution. A scheme used to guarantee the integrity of records based on the blockchain proposed for the ARCHANGEL project is shown in Figure 1.

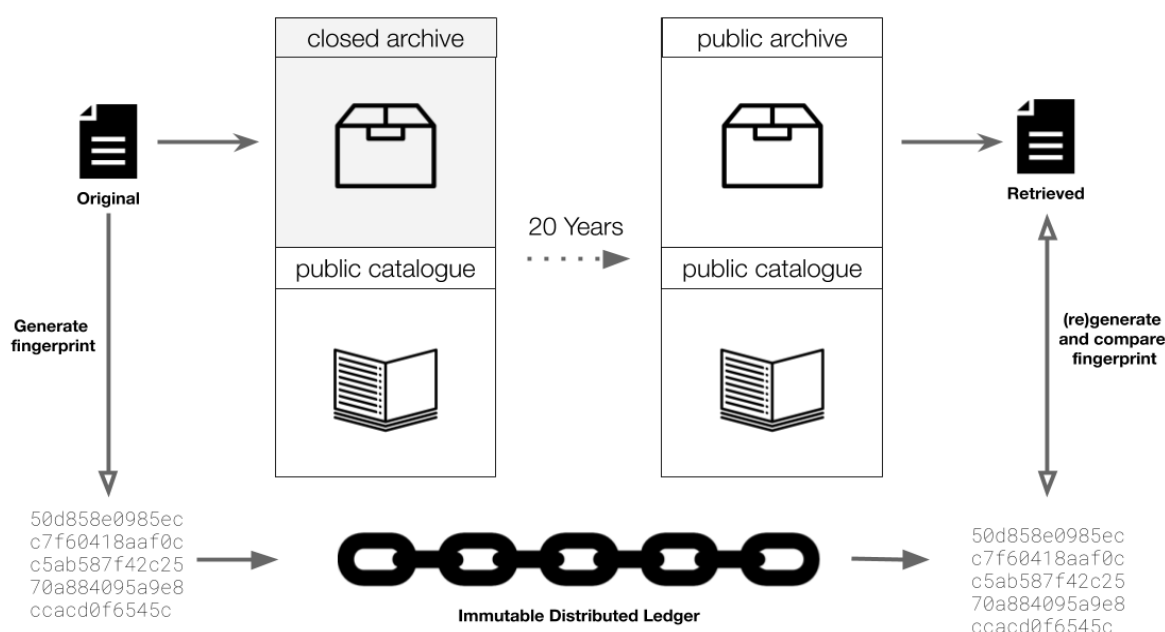


Figure 1. Repository of metadata of undisclosed records within the distributed ledger [6].

The scheme shown in Figure 1 saves the hash value within the blockchain, which can be used to verify the integrity of undisclosed records transferred to TNA; even after 20 years, it can be verified that the record was not forged or modified in an unauthorized manner by comparing the hash values. Within the ARCHANGEL project, a combination of blockchain and artificial intelligence was attempted in a bid to ensure the integrity of audiovisual records. Audiovisual records require a continuous conversion of formats for utilization, and the converted record has a different hash value even with the same content as the original copy if the format is changed. To resolve this problem, the ARCHANGEL project developed the temporal content hash (TCH) by applying the deep neural network as a kind of artificial intelligence technique [7] (p. 1). In this research, the machine learning algorithm was executed with the aim of extracting the content characteristics of audiovisual records irrespective of whether the format was converted. As the unique value is extracted by capturing the content characteristics regardless of the format conversion, TCH was the concept suggested for the first time by the ARCHANGEL project [7] (p. 3). This research proposed the application of TCH for the replacement of SHA 256 hash values as a means to guarantee the integrity of audiovisual records.

Figure 2 shows the machine-learning method used for the extraction of TCH. To extract TCH, a machine learning algorithm was executed by utilizing convolutional neural networks (CNNs) and recurrent neural networks (RNNs) on video data. The materials used for learning included the records of the courts in the UK with only limited motions included and Olympic video clips with numerous motions included [7] (pp. 5–6). The

machine learning algorithm was purposefully designed to enable the extraction of the hash values after checking the content characteristics of the records by learning three targets; namely, a clip of the original record, the same clip as the original but in a different format, and a clip different from the original but in the same format, as indicated in Figure 2 [7] (p. 5). The TCH extracted through machine learning was stored in the blockchain together with the unique identifier (UID) of the record, whereas the original record, the UID, the encoder used in the extraction of TCH, and the machine learning model were stored in the record management system after being bundled into a submission information package (SIP) [7] (p. 2).

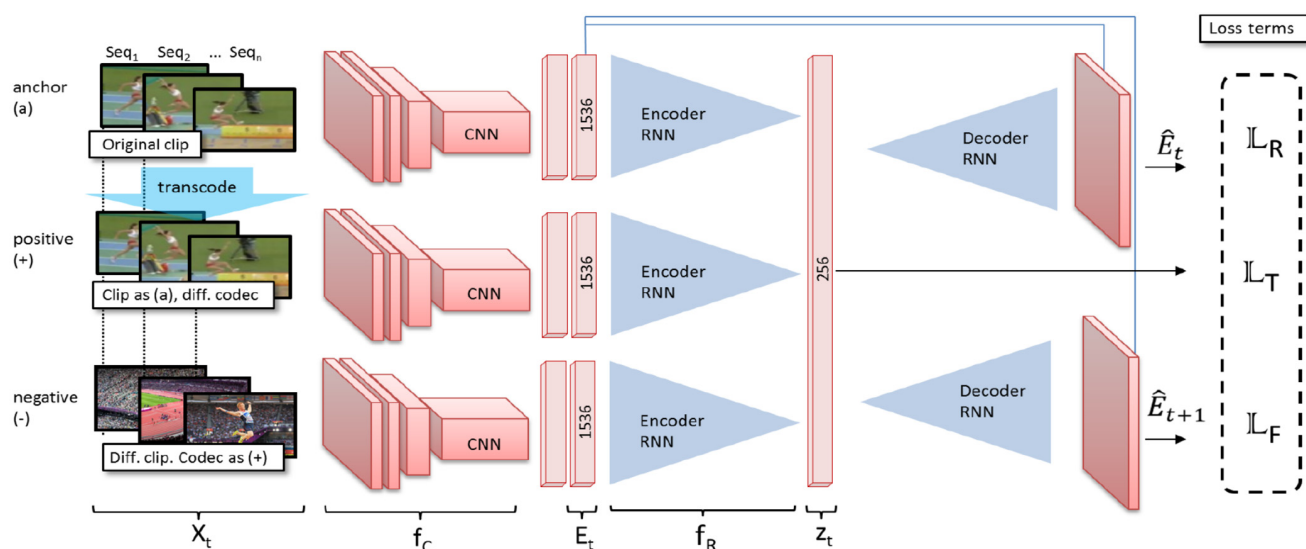


Figure 2. Method of machine learning for extraction of TCH [7] (p. 3).

A user survey was conducted using the ARCHANGEL prototype predicated on Ethereum, which was the outcome of the research conducted by The National Archives (TNA) in the UK, National Archives of Australia (NAA) in Australia, NARA in the United States, the National Archive in Norway, and the National Archive in Estonia [6]. In the user survey, most of the participants responded that they would affirmatively consider the adoption of the system at the level of the institution if the authenticity and integrity of the record are warranted. In addition, there was a common opinion that there were difficulties in use because blockchain technology is not easy to understand [6].

The ARCHANGEL project developed a prototype that can be applied for the management of public records by combining artificial intelligence with blockchain and conducting user research to enhance the quality of the prototype. Moreover, it aimed at providing a resolution for the long-term preservation of electronic records by extracting hash values that do not change despite the conversion of formats by applying a deep neural network. Developers of the ARCHANGEL project stressed the fact that the participation of diverse institutions was necessary to utilize blockchain for record management and that the collaboration of different AMIs is necessary for utilization of the trust structure of the blockchain.

At present, NAK is creating and managing audiovisual records in three separate formats, including the original, preservation, and utilization copies, using the MAM system. In the future, it will be necessary to identify authentic audiovisual records by distinguishing the characteristics of each format, and a TCH study as per the ARCHANGEL project is expected to be useful as a reference.

2.4. Model Designed for Authenticating Audiovisual Records

At present, audiovisual records are being managed and preserved in a distributed manner using CAMS and MAM systems. Information such as the title, retention period, and permission status for disclosure is used for storage and management of copies con-

verted into the format for preservation and utilization, including the original copies of the audiovisual records and the system attribute information related to the files. With respect to the processing method, if the information for record management is first registered in CAMS, the related information is delivered to the MAM in fixed time intervals. Management is difficult owing to the structure of duplicate layers and the omission of procedures for confirmation of the record authenticity; hence, there is room for improvement. This study proposes a method applying blockchain to guarantee the authenticity and long-term preservation of audiovisual records. Figure 3 shows business tasks carried out in each system based on the linkage between the CAMS, MAM, and blockchain systems [8].

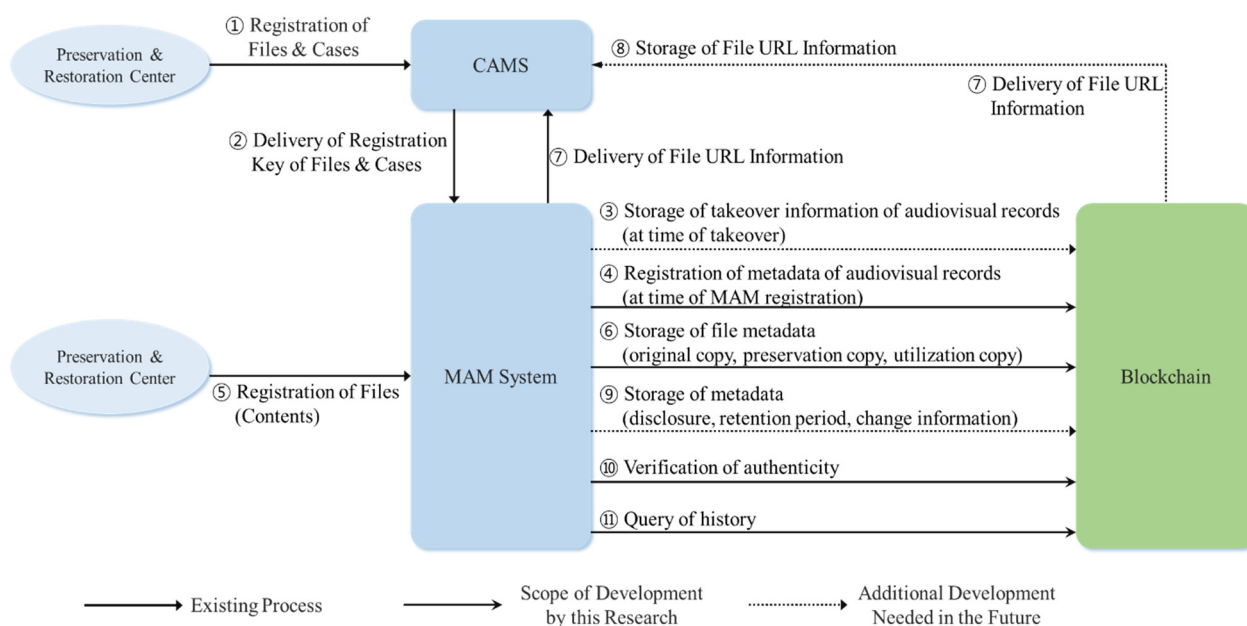


Figure 3. Proposed model for linkage between MAM and blockchain systems.

Within CAMS, the metadata of the records are registered in a hierarchical manner in terms of the files and cases as the current basic structure of the records, and the keys related to them are delivered to the MAM system. The metadata and other types of audiovisual record data are stored in the blockchain linked to the MAM system. After uploading the records to the MAM system, the staff in charge of managing the audiovisual records update the corresponding information to the blockchain at fixed time intervals. Subsequently, if the authenticity of the record needs to be confirmed, the verification can be carried out using the metadata and hash value stored in the blockchain as a reference, along with confirmation of the history of the record management as per the current design. Figure 3 shows the process of implementing the linkage between MAM and the blockchain systems.

Figure 4 shows the scheme for linking the blockchain with the MAM system. The MAM system was designed to be interfaced with the blockchain by way of a database that is linked with the MAM system as well as the linkage agent at the time when the registration key is delivered from CAMS and when the original, preservation, and utilization copies are stored in the repository. If the metadata of the audiovisual record and the metadata of the file that is distinguished into the original, preservation, and utilization copies are stored in the linked database by way of the MAM system, they are also recorded in the blockchain through the linkage agent linking the database with the blockchain and the smart contract. If any errors occur during this process, the linkage agent can be invoked to attempt the recording again, as per the design. If the metadata have been recorded in the blockchain, the processing status is transferred back to the database to synchronize the status between the blockchain and the database. The linkage agent system keeps the blockchain and MAM system synchronized by periodically checking whether the data in

the database have been updated. The model designed using this method was validated through a development process.

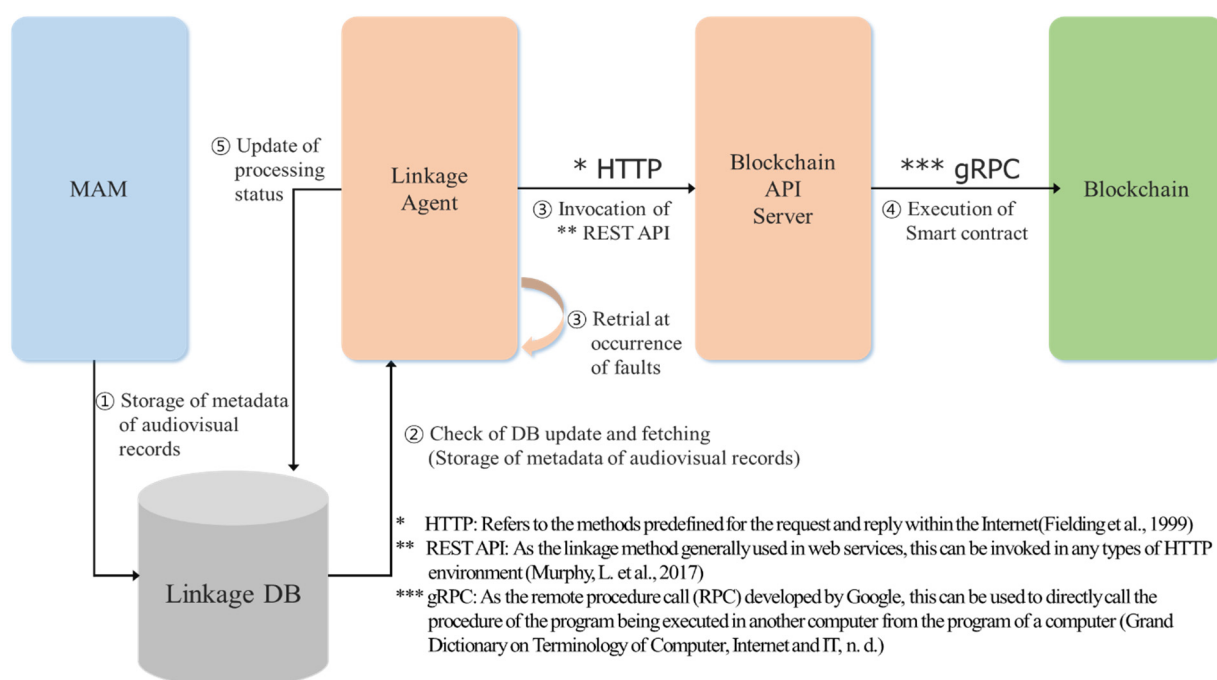


Figure 4. Model for linkage between MAM and blockchain systems.

2.5. Development of Blockchain System of Audiovisual Record Authenticity

We implemented a prototype to verify whether the model proposed in Section 2 operates properly. In this section, we describe the implementation of the proposed prototype. We used hyperledger fabric as our blockchain platform, and the prototype code can be found in [9]. This code is provided only when the sharing request is approved by the security policy, and permission is necessary when internal information is disclosed.

2.5.1. Transaction Audit Trail Interface

As a result of R&D, we developed a distributed application that audits and tracks the transactions of audiovisual records in MAM and stores them in the blockchain. Figure 5 shows the interface of distributed applications that track the transactions of audiovisual records.

Through this interface, we can search for and find the contents of the blockchain capturing information from the original audiovisual file registered in the MAM and store it in a block. MAM id is the identifier for audiovisual files granted by MAM, and the creation date is the time registered in MAM. The modification date is the time at which the transaction information of the record is stored in a block, and the transaction id is given as follows. The name, format, size, and hash algorithm used can be found using MD5. Further information can be found in the field of detailed information.

Media Asset Original File			
mam id	OI677553423455	media type	video
creation date	2020-11-27 01:15:25	modification date	2020-11-27 01:23:24
status	block	transaction id	7595341f74ccbc73f13f5b70b4b29ceaa0a04f522ff20c057fbc3e3547fe3cce
file name	oriental hospital	file format	MP4
volume	1930000	hash algorithm	MD5
hash	D71F373789A2602B34B6F67E843CA0DC		
error			
detail information			
video codec	FFV1	audio codec	FLAC
bit rate	55,747	frame per second	4,567
width	768	height	1,024
audio channel	25788	audio sampling rate	47,454
frame	97,553	playing time	00:01:15

Figure 5. Original file information in blockchain.

Original audiovisual records may be converted to other file formats for use or preservation purposes. When a transaction occurs, the blockchain captures this transaction and stores it in the block. Figure 6 shows the information conversion of audiovisual records captured by the blockchain as a transaction.

Media Asset Archival File

MAM id	OI677553423455	media type	video
creation date	2020-12-11 12:31:02	modification date	2020-12-11 12:32:05
status	block	transaction id	003893ecc23e4e2ba2948177605ee53434fa99e8325983481b262f06523c4ced
file name	oriental hospital	file format	MP4
volume	15652100	hash algorithm	MD5
hash	21B4D9A11BA348A0177865128106251F		
error			

detail information

video codec	FFV1	audio codec	FLAC
bit rate	55,747	frame per second	4,567
width	800	height	600
audio channel	25788	audio sampling rate	47,454
frame	97,553	playing time	00:01:15

Figure 6. Transaction information conversion in blockchain.

Figure 6 shows the conversion time, new hash value, and change in file size. Audiovisual files that have been changed according to a conversion transaction are given new hash values, which can be stored in blocks to verify their authenticity.

2.5.2. Authentic Verification Service

The archives should be able to provide part of an entire record and in different formats, depending on the needs of the consumer and service policies. Audiovisual records are changed into various formats as needed, and thus blockchain should be able to audit and track all change transactions and provide authentic verification services. Figures 7 and 8 show the interface of our developed authentic verification service.

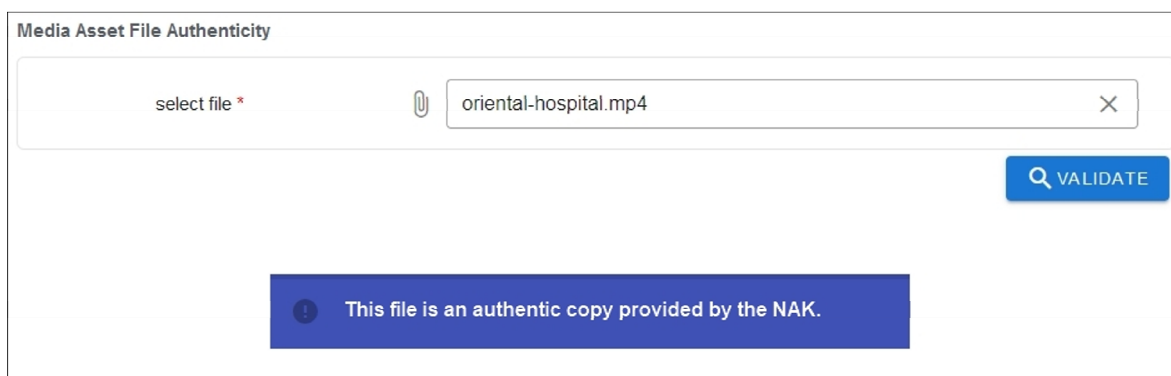


Figure 7. Authentic copy verification message.

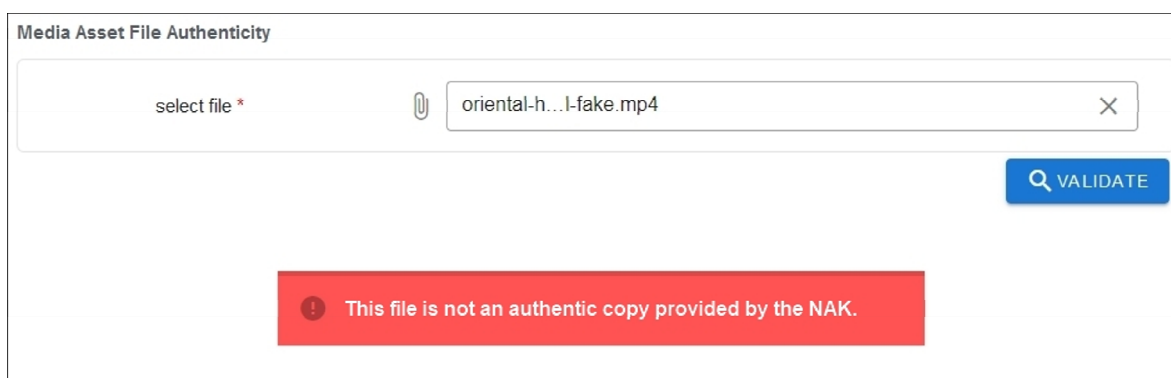


Figure 8. Forgery copy verification message.

Consumers who use records can check online whether the audiovisual file records they received have been authenticated through the service interface provided by the NAK. They can immediately verify the authenticity by uploading their audiovisual record files to the authentic verification web page, as shown in Figure 7. If the uploaded file is not an authentic copy owing to a forgery, a message is provided to confirm it, as shown in Figure 8.

The integrity of the records can be guaranteed by storing the metadata of the audiovisual records in the block, which cannot be forged or modified. The two systems that used to be managed in a discrete fashion can be interfaced organically by synchronizing the blockchain, CAMS, and MAM; therefore, the management of audiovisual records can be facilitated. Because the authenticity of the record can be confirmed by verifying the hash values of the record metadata stored in the blockchain, the authenticity of the record can be easily confirmed when audiovisual records are utilized. It is also possible to trace the process of the task performance as well as the history of the records because the update history of the record metadata stored in the blocks is immutable given the characteristics of the blockchain. By developing the interfaces in advance for those parts that also need to be incorporated in the future as the MAM system is enhanced, it becomes easy to develop and link the corresponding functions after a system enhancement. Based on the linkage between the blockchain and the MAM system, a scheme for the management of audiovisual records

was established with the pivotal role assumed by the MAM system, thereby eliminating the duplicate management structure as well as the burden of redundant tasks. By combining these fundamental systems with artificial intelligence in the future, it will be possible to devise diverse methods for creating unique values for the identification of audiovisual records and for long-term preservation. However, this may warrant additional research.

Many records provided by NAK are currently used in court trials. Only authenticated records must be provided because the records required by the courts are adopted as legal evidence. Figure 9 shows the process of exchanging records between the NAK and the court.

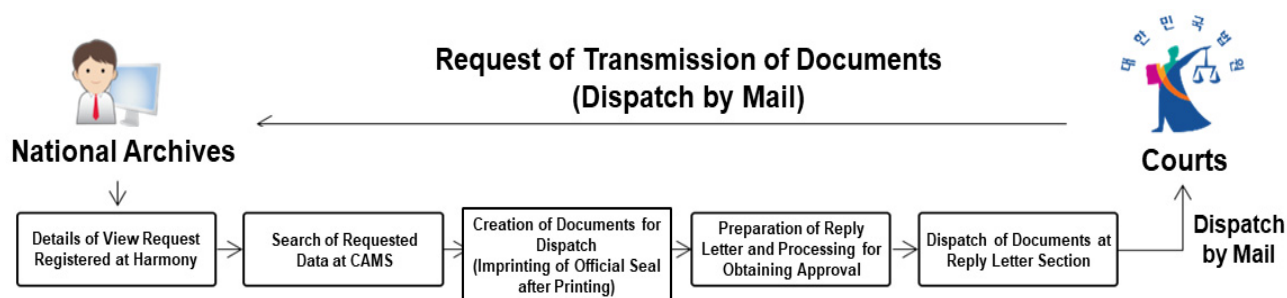


Figure 9. Record exchange process between NAK and the court.

The courts still request records by mail, and the NAK replies to the courts with watermarked and stamped records. This administrative process occurs repeatedly by stakeholders such as lawyers, related agencies, and defendants in litigation. As such, the NAK consumes considerable time and resources to repeatedly verify the authenticity of the same records. The blockchain verification service developed by NAK can improve such resource-consuming administrative activities through technology. From 2024, the courts of Korea will extend the evidential value of analog records only to digital records.

3. Method for Authenticating Information Dataset

3.1. Overview

In general, a dataset is defined as a collection of related information, particularly information formatted for analysis by a computer [10]. In Korea, this is similarly defined. Citing the original text, a dataset of administration information is an “administrative information dataset referring to characters, numbers, figures, images, and other data combined for production, collection, processing, storage, search, provision, transmission, and reception in the information system” [11]. A dataset that is legally subject to record management is produced by and stored in an information system operated by an administrative agency, called an “administrative information system”. The scope of the dataset is specified as a dataset in the administrative information system, and this dataset is referred to as an “administrative information dataset”.

Korea has introduced administrative information systems; that is, administrative information systems used throughout the national administration, to achieve a fast and efficient e-government with the world’s best information technology. As administrative work is gradually carried out through administrative information systems, most of the data produced are managed and stored in a database. The dataset is an important record not only for its value in terms of evidence and historical documents, but also for its value in analysis and use. It is the data source of public electronic documents and certificates issued by administrative agencies, the essential data for reproducing the functions of the administrative information system, and the input data for analysis and utilization. Because informatization and digitization will accelerate further, and the number of systems and administrative information datasets will increase enormously, it is necessary to study record management methods for them. According to [12], as of 2019, there are a total of 16,228 administrative information systems operated by 1095 public institutions, where a

total of 19,533 DBMSs have been installed. There are various DBMS vendors, including Oracle, SQL Server, PostgreSQL, Tibero, CUBRID, and DB/2. In addition, an average of 14,226 (approximately US\$ 3 billion) informatization projects were ordered per year between 2015 and 2019, and this trend will continue.

NAK has an obligation to establish a policy for the standardization of record management such that all records can be managed and efficiently and uniformly utilized. Therefore, since 2005, NAK has endeavored to conduct R&D projects and create laws, guidelines, and regulations to establish a record management system for administrative information datasets. In 2020, NAK revised the enforcement ordinance on the management of public records, established public standards for the record-keeping criteria for datasets [11,13], and conducted pilot projects and briefing sessions. Thus, a record management methodology for an administrative information dataset is systematically being developed.

Among the administrative information datasets, the dataset stored in the database of the administrative information system must be dealt with most urgently. To summarize the record management method for a dataset, NAK does not accept all datasets. It only selects and accepts some valuable datasets considering the efficiency and administrative independence of the system. The method is divided into (1) transferring datasets into NAK and (2) self-management by public institutions. The first method creates datasets in a dump file format that can guarantee the preservation of the structures and their data (e.g., SIARD-KR [13,14] and SIARD 2.0 [15]) and transfer them to the CAMS. In the second method, each public institution manages the datasets. In the case of transfer to NAK, a method used to verify the integrity of the datasets can be freely designed and implemented regardless of the administrative information system that stores them before transfer. By contrast, the self-management method should have little effect on the current administrative information system where the datasets are stored. In addition, integrity verification can be executed in a consistent manner for various DBMSs. In this paper, we propose a method for verifying the integrity of a dataset self-managed by a public institution based on blockchain technology.

3.2. Range and Type of Dataset

The dataset dealt with in this paper is an administrative information dataset stored in the database of the administrative information system, which is self-managed by public institutions without being transferred to NAK. Most databases that contain administrative information datasets are relational databases. Although NoSQL databases have recently emerged, there are more relational databases. The different relational databases available include Oracle, MS SQL Server, MySQL, Tibero, Sybase, and Cubrid. Databases of domestic companies, such as Tibero, Sybase, and Cubrid, are also used. In summary, the target dataset of this study is an administrative information dataset stored in a relational database of an administrative information system managed by public institutions. A plan that can be applied to various database management systems (DBMSs) should be developed.

3.3. Application Direction of Blockchain Technology

As a representative record and archive management agency in Korea, NAK has established and improved fundamental policies for records and archive management. In particular, NAK should establish a standardization policy for record and archive management and develop standards that allow public institutions affiliated with NAK to manage and utilize the archives in an efficient and uniform manner. Therefore, a method for verifying the integrity of an administrative information dataset based on blockchain technology should be designed and proposed to accommodate various conditions and situations.

As mentioned in Section 3.2, we studied datasets that were not transferred to NAK but were self-managed by each public institution and stored in several relational databases of various companies. As mentioned earlier, blockchain recordkeeping solutions can be classified into three types [5]. The method proposed in this study corresponds to the first of these three types. In the second and third types, it is assumed that actual records must be created and managed in the blockchain system. However, most administrative information

systems in Korea were designed and built for each public institution before the advent of blockchain technology or without such consideration. It is impossible to introduce blockchain technology in a consistent manner for an enormous DBMS. Moreover, it is almost impossible to replace various existing DBMSs with blockchain databases. Therefore, it is most suitable to verify the integrity of the databases using a blockchain as a mirroring system that can be applied independently to various DBMS products.

3.4. Model Designed to Verify Dataset Integrity

3.4.1. Typology of Blockchain Recordkeeping Solutions

Because existing administrative information systems do not consider, or are not built to accommodate, blockchain databases, it is necessary to connect the blockchain in the form of a mirror system to verify the integrity of the dataset accumulated in a relational database.

Considering the blockchain from the viewpoint of storage, we can consider it as a database where, once stored, the data can never be changed. The data stored in the blockchain have numerous copies as nodes that comprise the blockchain network. Thus, the greater the number of nodes, the higher the security and stability, but the lower the performance. Owing to the tremendous trend of increasing information resources and data, storing large amounts of datasets as they are in the blockchain will lead to a worst-case situation in which the performance is degraded and costs increase. It is appropriate to store information of a relatively small size that is representative of the associated large dataset. Therefore, the type of method introduced in this case study can be seen as the first type, a mirror system, among the three types presented in [5].

The metadata for the datasets stored in the blockchain can be configured as shown in Table 1. The following information must be included: integrity information that can verify the integrity of the dataset (e.g., the hash value and hash function type), server connection information that refers to the location of the database in which the dataset is stored (e.g., DBMS software, IP address, port number, and account (ID/password)), data extraction information that specifies how to create a file for the dataset (SQL query or export command), and data range information that specifies the range of the dataset for verification (e.g., schema information, table information, PK range, and PK list). The content of the data range information may differ according to the data extraction information. In addition, system overview and data type information may be included in the metadata.

Table 1. Information stored in the blockchain.

Information Type	Content	Ex.
Integrity	hash value, hash function type, etc.	d4a97779039f735c765c9179496a49de3442fcb46a6c909d82478dfa72bc437, SHA256
Server Connection	DBMS software, IP address, Port Number, Account, etc.	Oracle Database 12c Enterprise Edition Release 12.1.0.2.0—64bit, 113.198.49.87, 1521, user1
Data Extraction	SQL query or export command, etc.	SELECT * FROM Table 1 WHERE PK1 ≥ 1 AND PK1 ≤ 10000 or SELECT * FROM Table 1 WHERE PK1 IN (3,59,101,92)
Data Range	Schema information, Table information, PK range or PK list, etc.	Schema 1, Table 1, [1, 10,000] or Schema 1, Table 1, (3, 59, 101, 92)

3.4.2. Configuration of Blockchain Network

Because the information of the national administrative agency may require confidentiality and security, a private blockchain would be suitable. Regardless of whether it is private or public, blockchain requires multiple nodes to participate. To construct a blockchain network of governmental administrative agencies, it is necessary to develop a policy method rather than a technical method. For example, administrative agencies should be encouraged to participate as nodes in the blockchain network through enforcement and incentive strategies.

Administrative agencies in Korea are organized into hierarchical structures. It is necessary to enforce institutions that perform similar administrative tasks (e.g., local offices of education, local governments, and national universities) to participate as blockchain network nodes in a policy manner; for example, when budgeting or evaluating institutions, additional incentives can be provided to institutions participating in the blockchain network.

In the administrative information system, a large number of datasets are placed in DBMSs in a variety of formats and structures and are utilized in many different domains. The procedure used to create a blockchain transaction for integrity verification that can be applied to various DBMSs is as follows. The first step is to extract a dataset into a file using functions provided by the DBMS, such as SQL queries, dumps, exports, and downloads. The second step is to generate integrity information for the extracted file and create metadata composed of integrity and other types of information. The third process is to create and store blockchain transactions using metadata. It is assumed that the tasks related to the dataset subject to integrity verification have already been terminated and will not be updated or deleted until the preservation period expires. If the dataset needs to be updated or deleted, another blockchain transaction must be created and stored.

The procedure used for creating and storing blockchain transactions is illustrated in Figure 10 and is described as follows. When extracting a dataset into a file, specific criteria for extraction are required. The dataset is extracted in table units, and the appropriate size of the extracted file is determined depending on the computing or storage resources of the institution. Using this fixed size as a standard unit, the dataset is separated by the number of rows or the value of the timestamp, and one file for each separated dataset is extracted. A hash value for each file is generated, and a blockchain transaction is created with metadata consisting of integrity information, server connection information, data extraction information, and data range information, as indicated in Table 1.

One of the following methods can be selected for extracting the dataset.

- SQL query:

The SQL grammar may differ for each DBMS. The range of the extraction dataset is customized. The data are stored without the structure and additional information on the schema and tables.

- Export function in DBMS:

Every DBMS provides an export function. Because additional information related to the SQL query is included along with the actual data, the file size may increase. A table is used as the dataset extraction unit. When upgrading, the DBMS may change the format of the dump file, and the blockchain transaction must be re-created.

- SIARD_KR [13,14] and SIARD 2.0 [15]:

It is advantageous to extract the dataset into files with a standard tool used for most DBMSs. However, doing so is time-consuming because the file is created through steps such as XML parsing and XML validation. As the dataset extraction unit, a schema is used in SIARD and a table is applied in SIARD_KR. Because they are open sources, it is possible to change the source code to improve the tool performance and add functionality.

- Commercial DB Software:

As a significant advantage, files can be extracted in various ways, including through an SQL query and export functions, using a single tool for several different DBMSs. To use the software, a subscription is required or the software can be purchased outright. TOAD and DBeaver are representative types of commercial DB software.

After the file is extracted, integrity information with the cryptographic hash function (e.g., SHA-1, SHA-256, and SHA-512) and the hash value generated by the function is created. A blockchain transaction consisting of integrity information, server access information, data range information, and data extraction information, among other information types, is generated.

If the DBMS is inevitably changed owing to system obsolescence and SW support interruption, or if the data-encoding format changes after the upgrade of the DBMS, another blockchain transaction must be created.

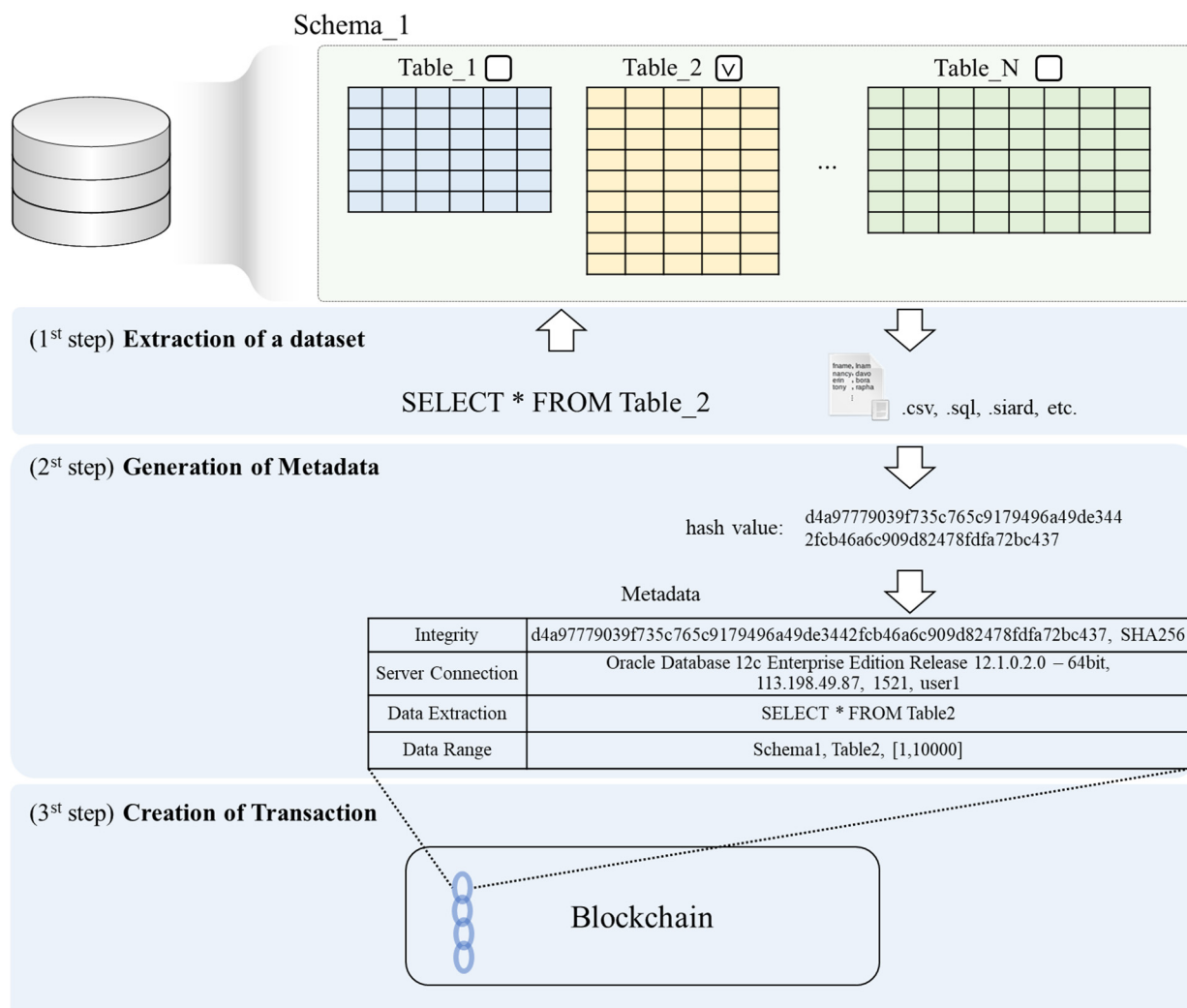


Figure 10. Generation of blockchain transactions.

3.4.3. Integrity Verification of Datasets

Figure 11 illustrates the process of verifying the integrity of the datasets. After selecting the datasets for verification, the metadata of the blockchain transactions are searched to find the desired transaction. The first step is to connect the DBMS using the IP address, port number, and account information of the server from the server connection information for the transaction. The second step is to extract the dataset into a file by executing an SQL query made with the schema and table information, PK range, and PK list, and to generate a hash value of the file using the cryptographic hash function from the integrity information. The final step is to compare the hash value stored in the integrity information with the newly generated hash value. If two hash values are identical, the integrity of the dataset is guaranteed; if they are different, the dataset changes.

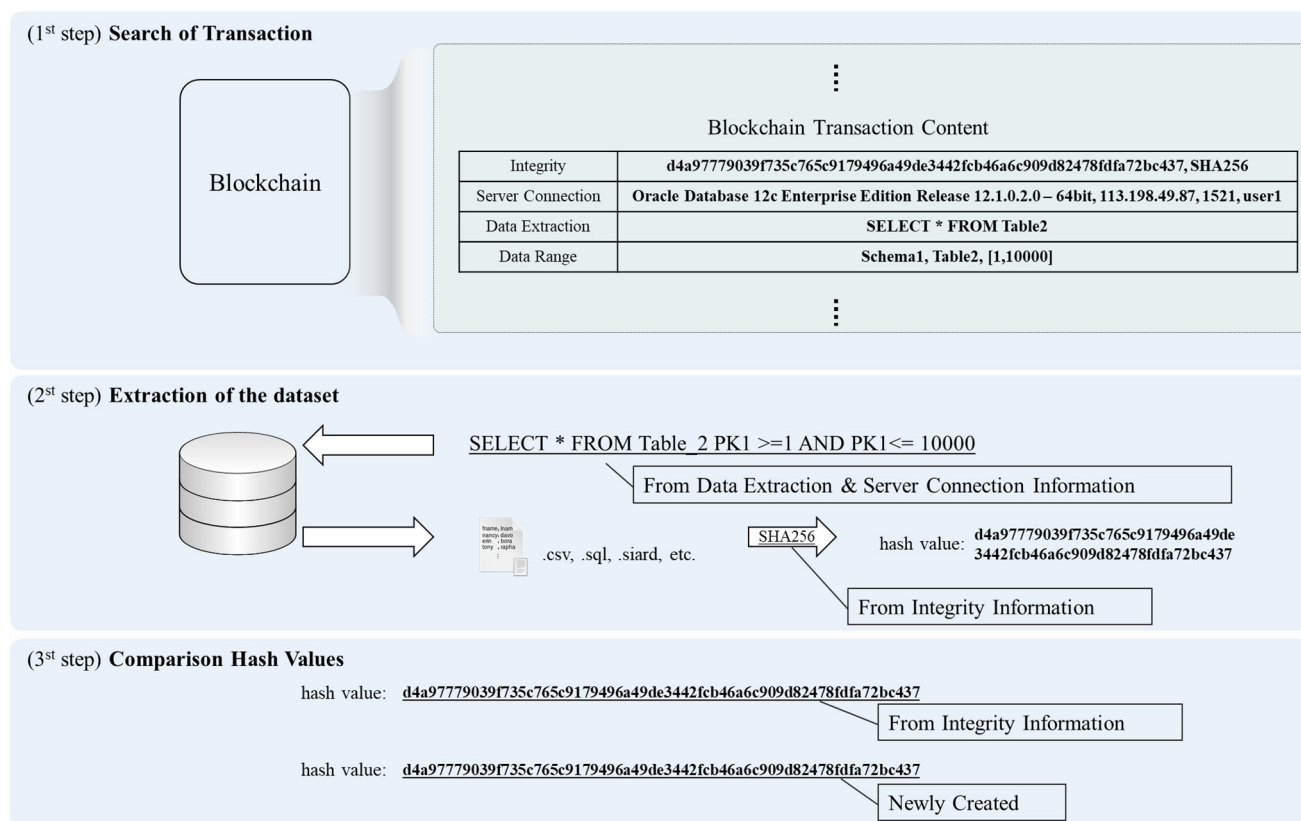



Figure 11. Integrity verification of administrative information dataset.

3.4.4. Implementation

We implemented a prototype to verify the effectiveness of the proposed model. We used LOOPCHAIN for the blockchain platform [16]. The prototype code can be found in [17]. Figure 12 shows the GUI (for storing a dataset in blockchain) providing the execution of an SQL query on the selected table, generating the desired contents as a dump file in .csv format, and storing the hash value and metadata of the file. In addition, Figure 13 shows the GUI (for verifying the integrity of a dataset) providing the selection of one of the generated dump files, executing a SQL query with an IN operator and a set of PKs stored in blockchain, sending the results into a dump file, generating a hash value for the file, and comparing the hash value stored in the blockchain with the generated hash value.


Integrity verification of datasets based on blockchain
 Verify the integrity of a dataset which is defined by a PK list stored in blockchain


INTEGRITY VERIFICATION OF DATASETS BASED ON BLOCKCHAIN
 Storing a Dataset
Verifying Integrity

Database:
 Table:
 PK column:

Metadata to be Stored into Blockchain

SQL Query for initial Retrieval	select * from address order by address_id
Start of PK	1
End of PK	605
Dataset ID	dataset.address.address_id
Hash Value(SHA256) from results by query	bed87e74bd1cc6b4a5709d535282e370f22806b21fe93c40c0153b331e7ea839

Figure 12. Prototype GUI for storing a dataset in blockchain.


Integrity verification of datasets based on blockchain
 Verify the integrity of a dataset which is defined by a PK list stored in blockchain

INTEGRITY VERIFICATION OF DATASETS BASED ON BLOCKCHAIN
 Storing a Dataset
Verifying Integrity

List of dataset IDs

Dataset ID	dataset.actor.actor_id	dataset.film.film_id
	dataset.film_actor.actor_id	dataset.city.city_id
	dataset.country.country_id	dataset.address.address_id
	dataset.category.category_id	dataset.film_text.film_id

Dataset ID	dataset.address.address_id
SQL Query for Verifying(A)	select * from address where address_id between 1 and 605 order by address_id ASC

Hash Value(SHA256) from results by query(A)	bed87e74bd1cc6b4a5709d535282e370f22806b21fe93c40c0153b331e7ea839
Hash Value(SHA256) stored in blockchain(B)	bed87e74bd1cc6b4a5709d535282e370f22806b21fe93c40c0153b331e7ea839
Verification result(A==B)	identical

Figure 13. Prototype GUI for verifying the integrity of a dataset.

3.5. Implications

3.5.1. Blockchain Participation

Because both the level of security and stability improve with an increased number of nodes making up the blockchain, a public blockchain is the most ideal model. A private blockchain has been evaluated as having slightly less stability than a public blockchain. However, it is difficult to introduce a public blockchain into the public domain owing to the existence of confidential or personal information that should not be exposed to the public sector. Even if there is a method by which confidential or personal information can be protected, it is difficult to obtain a positive response from people sensitive to the disclosure of personal information. Hence, a private blockchain is appropriate for government administration, involving several participating institutions or authorized groups. Because a larger number of participating nodes results in more security and stability, the node participation of a private blockchain should be increased by establishing control of government administrative organizations through close cooperation. To increase the level of participation, incentives such as adding related items to evaluate the public institutions should be provided.

3.5.2. Beyond Mirroring

Administrative information systems have been established since the E-Government Act was enacted in 2001 and the e-government support project was promoted. This paper proposes a method for verifying the integrity of administrative datasets. When administrative information systems were first established, there was no blockchain technology, and therefore such technology was not considered. For various types of DBMSs, an integrity verification method must be proposed in a consistent and unified manner. It is most appropriate to adopt a mirror system method with a small amount of integrity information and metadata, such as hash values for real datasets.

However, it is best to create and manage actual data in the blockchain. When establishing an administrative information system in the future, it should be designed for storing datasets in the blockchain. The create, read, update, and delete (CRUD) transaction in the database should not mirror the actual data by interlocking with the blockchain through a smart contract, but should set the final goal in the direction in which the actual data itself is stored in the blockchain.

4. Directions of Blockchain for Records and Archive Management

4.1. What Is a Blockchain Used in Record and Archive Management?

A blockchain is a data storage platform that, once stored, can never be changed. This function can be achieved by distributing authority to and monitoring by all participating members, rather than using a central strong control. From such an intrinsic property of decentralization, it has important characteristics of antitrust and openness. Antitrust is consistent with the most important value in record and archive management, and must be neutral from all external powers. In addition, openness is a prerequisite for preserving the four main characteristics of authoritative records. Blockchain is an extremely attractive technology in the domain of record and archive management.

We believe that blockchain is the most outstanding of the technologies proposed thus far for ensuring data integrity. This belief is supported by the fact that huge amounts of capital are being traded on virtual currency platforms such as Bitcoin or Ethereum, and no one has yet questioned the integrity of blockchain technology.

Authenticity is the most essential attribute of record and archive management. In addition, integrity is a necessary condition for maintaining authenticity. Therefore, although there are many practical limitations (e.g., budget, technology, and time), it is necessary for national record and archive institutions to apply blockchain to their record and archive management systems and to establish a long-term plan to achieve such application.

4.2. Blockchain Directions for Record and Archive Management

Korea has built 16,000 information systems to achieve an e-government system, which is a government that efficiently conducts administrative tasks and quickly provides administrative services to the public by digitalizing the work of public institutions with information technology. Moreover, it will continue to migrate to existing information systems or build new information systems. Various types of electronic records, such as electronic documents, audiovisual records, and administrative information datasets, are produced and managed in such systems. Although blockchain technology should be applied to manage these records, it is impossible to introduce into the numerous information systems all concurrently available technologies.

A long-term plan to apply a blockchain in a step-by-step manner is required to realistically consider the resources and manpower according to the limited budget. A long-term plan is required by considering the resources and manpower required within a limited budget and applying them in stages. We present practical criteria for applying blockchain based on the experience from two case studies.

First, it is necessary to distinguish between existing and new information systems. If the system is new, blockchain can be introduced without any restrictions to the implementation. Otherwise, the introduction of blockchain should proceed without affecting the existing information system. A new system based on blockchain will ensure authenticity by storing and tracing the contents and metadata of records over the entire life cycle. By contrast, in the case of an existing information system, it is possible to verify the authenticity of the records by using hash values in conjunction with the blockchain platform.

Second, the existing information system should be divided into two categories: (1) records that pass through their life cycle in the production system of the origin institution and (2) records that are transferred to and preserved in the record and archive management system. In general, records are transferred to the record and archive management systems and are either discarded upon expiration of the retention period or preserved permanently. The administrative information dataset produced may have an exceptional life cycle. In addition, the administrative information datasets in the DBMS can continue to be managed in the corresponding system without transfer to NAK. Although these records are physically kept in the institution of origin, the management authority is logically transferred to the record and archive management system and must be continuously monitored by NAK. If the records are physically transferred to the record and archive management system, the authenticity can be easily verified using a mirror type because they are in the form of one or more digital objects (e.g., files) packaged in an information package. However, if records are continuously stored in the production system and directly linked to the blockchain platform, the structure of the existing system will change in a complicated manner, and the performance will inevitably decrease. Therefore, after accessing the system and extracting the contents, it is possible to indirectly link with the blockchain and verify its authenticity.

Finally, it is necessary to divide it into two categories according to the type of final information to be recorded and managed: (1) an evidence producing system, such as for reports on various tasks or certificates identifying persons or objects, and (2) a system that manages information (e.g., contracts and transactions) occurring in a process between two elements. In the former case, most of the evidence is managed as a digital object, such as a file. A mirror blockchain is suitable for this type of system. In the latter case, information generated in the process between the two elements (e.g., sender and receiver, and contractors) will be stored and managed as metadata in the form of text. In this case, a digital record blockchain is desirable for building the system.

5. Conclusions

In this paper, we introduced two R&D cases, the first of which is the use of blockchain transaction audit trail technology to ensure the authenticity of audiovisual archives. The second uses blockchain technology to verify whether numerous datasets created by govern-

ment agencies are authentic. Based on the lessons learned in our case study, we discussed the blockchain directions for record and archive management.

The first case study presented a transaction audit trail interface and authentic communication service, which stores hash values for audiovisual records, such as audio and video files, into the blockchain and checks the information of the transaction. We also modeled a service that allows users to check whether their audiovisual records are authentic and develop actual implementations. In the second case study, we presented a blockchain-based model that can verify whether datasets stored in relational databases have been forged. This can be applied to several DBMS types and operates on a large dataset. From the two case studies, we confirmed the authenticity of the records using blockchain, such as providing record verification services to courts for electronic records already stored in the blockchain. In addition, it is expected that a high security and stability of the record and archive management system can be guaranteed if transactions consisting of the CRUD of the databases are stored in the blockchain when establishing an administrative information system.

The two case studies described in this paper correspond to a mirror type, the first among three distinct stages. The actual record itself is not stored, but a hash value as a digital fingerprint is anchored to the blockchain. This is a simple and effective method that can apply blockchain without affecting the existing system. As the next step, we aim to study a recordkeeping solution of a digital record type, where all contents of the records can be stored in the blockchain over the entire life cycle. In the near future, we will design and implement a model for the management of electronic documents into a digital record type. In addition, we will endeavor to provide useful services to the public with record and archive management systems based on blockchain.

Author Contributions: The authors contributed to this work by collaboration. Conceptualization, H.W. and D.Y.; methodology, H.W. and D.Y.; software, H.W. and D.Y.; writing—original draft preparation of the first case study, H.W.; writing—original draft preparation of the second case study, D.Y.; and writing—review and editing, D.Y.; supervision, D.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a fund from the Archive Management Research Program of the National Archives, Korea.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jung, C.-S. *The Theory of Electronic Government*, 1st ed.; Seoul Economic Management Publishing Company: Seoul, Korea, 2007; pp. 99–128.
2. National Archives of Korea. *Government Business Datasets Records Management Development Planning*; National Archives of Korea: Daejeon, Korea, 2017.
3. National Archives of Korea. *Technical Specification for Long-Term Preservation Format ver. 2.1. Standard*; NAK 31:2017(v2.1); NAK: Daejeon, Korea, 2017.
4. ISO 15489-1. Information and documentation-records management part 1. In *Concepts and Principles*; ISO: Geneva, Swiss, 2016; pp. 4–6.
5. Lemieux, V.L. A typology of blockchain recordkeeping solutions and some reflections on their implications for the future of archival preservation. In Proceedings of the 2017 IEEE International Conference on Big Data, Boston, MA, USA, 11–14 December 2017; pp. 2271–2278. [[CrossRef](#)]
6. Green, A.; Das, A.; Cooper, D.; Fawcett, J.; Keller, J.; Higgins, J.; Bui, T. *ARCHANGEL: Guaranteeing the Integrity of Digital Archives*; Open Data Institute, The National Archives, University of Surrey: London, UK, 2019.
7. Bui, T.C.; Collomosse, D.; Bell, J.; Gree, M.; Sheridan, A.; Brown, J.A. *ARCHANGEL: Tamper-proofing video archives using temporal content hashes on the blockchain*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.

8. Wang, H.; Moon, S.; Han, N. A Study on the Applications of Blockchain Transactions and Smart Contracts in Recordkeeping. *J. Korean Soc. Arch. Rec. Manag.* **2020**, *20*, 80–105.
9. NAK-DLT. Available online: <https://github.com/Hosung-wang/NAK-DLT> (accessed on 12 July 2021).
10. Pearce-Moses, R. *A Glossary of Archival and Records Terminology*. Society of American Archivists; The Society of American Archivists: Chicago, IL, USA, 2013.
11. Ministry of the Interior and Safety. *Enforcement Decree of the Management of Public Records Act*; Article 2, No. 11; NAK: Daejeon, Korea, 2021.
12. Ministry of the Interior and Safety. *Statistical Report on 2020 GEAP (Government-Wide Enterprise Architecture)-Based Public Sector Information Resource Status*; Ministry of the Interior and Safety: Sejong City, Korea, 2019.
13. National Archives of Korea. *Record Keeping Criteria for Dataset (Composition of Dataset Management Reference Table & Exchange of Dataset)*; Standard, NAK 35:2020(v1.0); NAK: Daejeon, Korea, 2020.
14. SIARD_KR. Available online: https://github.com/nakdataset/SIARD_KR (accessed on 12 July 2021).
15. ECH. *eCH-0165 SIARD Format Specification*; Standard, eCH-0165(v2.0); Verein eCH: Zurich, Swiss, 2016.
16. LOOPCHAIN. Available online: <https://www.iconloop.com/en/loopchain> (accessed on 21 July 2021).
17. Prototype for Verifying Integrity of Dataset using Blockchain. Available online: <https://github.com/likeba/VID> (accessed on 21 July 2021).