electronics

MDPI

Article

Design of an Enhanced Web Archiving System for Preserving Content Integrity with Blockchain

Hyun Cheon Hwang [1], Jin Gon Shon [2] and Ji Su Park [3,*]

[1] Department of Industrial and Information System, Graduate School, Seoul National University of Science and Technology, Seoul 01811, Korea; a.hwang@seoultech.ac.kr
[2] Department of Computer Science, Korea National Open University, Seoul 03087, Korea; jgshon@knou.ac.kr
[3] Department of Computer Science and Engineering, Jeonju University, Jeonju 55069, Korea
* Correspondence: jisupark@jj.ac.kr

Received: 15 June 2020; Accepted: 4 August 2020; Published: 5 August 2020

Abstract: A Web archive system is a traditional subject for preserving web content for the future and the importance is getting more significant due to the explosive growth of web content. The reference model for an open archival information system (OAIS) has been advising guidance for a long-term archiving system and most organizations that archive web content follow this guidance. In addition, the web archive (WARC) ISO standard is for web content archiving. However, there is no way to secure content integrity, and it is hard to identify the original. Because of limitations, a web archive system has a weakness against the dispute of content integrity. In this paper, we proposed the blockchain linked (BCLinked) web archiving system, which uses blockchain technology and an extended WARC field to keep a web content integrity metadata into a blockchain. Furthermore, we designed the BCLinked web archiving system, and we confirmed the proposed system secures content integrity through the experiment.

Keywords: blockchain; web archive; WARC; web crawling; BCLinked; web archiving system

1. Introduction

The purpose of the web archive is to collect web content and preserve it for the long term to deliver valuable data to descendants [1]. Many personal data exist in various digital spaces such as web, email. Moreover, there are multiple types of research regarding collecting this personal digital data [2]. Nowadays, most web contents are dynamic and personalized web content rather than it used to be static web content because the web is the primary channel to deliver personalized information to people. The volume of content from the digital channel, including the web, is getting bigger, and the requirement of standardization for digital resources archiving was needed. Consultative committee for space data systems (CCSDS) has developed a reference model for an open archival information system (OAIS), which is ISO 14,721 for long-term preservation of digital records, and many web archive systems follow the standard [3]. However, the OAIS reference model does not provide an implementation method but prescribes a requirement to ensure OAIS-compliant [4]. There is the weakness of content integrity against unauthorized manner, and there is no mechanism to guarantee content integrity for the long-term, even though the OAIS reference model tells the long-term preservation of digital records. A blockchain is an emerging technology that connects a previous dataset called a block by cryptography key and decentralized system which anyone can share the ledger to review all datasets. A blockchain is getting more popular in an enterprise because it provides an enhanced trusted system with these characteristics [5]. In this paper, we proposed the extended specification of web archive (WARC) to integrate with blockchain technology and designed the enhanced web archiving system based on blockchain technology to provide better web content integrity. Finally, we implemented the

proposed enhanced web archiving system and confirmed the enhanced archiving system provides enhanced web content integrity.

2. Related Research

2.1. OAIS

OAIS reference model was developed by CCSDS in 2002 [3] as technical guidance for the long-term archiving and sharing the archiving information. This reference model is ISO 14,721 standard, and many organizations that do archive content follow this guidance model for their archive system infrastructure. It provides (a) the concept of archiving management for digital assets' long-term archiving and access; (b) the guidance for effective archiving management in a typical organization (c) the conceptual architecture for digital archiving (d) the way to compare each other different archiving strategy (e) the way to compare the data models (f) the way to identify the key factors for digital archiving. The environment model of an OAIS consists of producer, consumer and management [3]. A producer has the role which provides the information to be preserved, management has the responsibility to control of the OAIS and consumer's role is to interact with OAIS service to query and retrieve preserved information. A producer collects the data to be preserved and deliver to the OAIS as the submission information package (SIP) and OAIS manages the data as the archival information package (AIP). A consumer sends a query and gets query responses to retrieve interesting information from OAIS. There are three services guidance; common services, network services and security services in the detailed description of functional entities of OAIS. The security services advise data integrity service should secure the data which should not be altered or destroyed by an unauthorized manner, and data should be in permanent data storage [3]. Even though OAIS advises this security guidance, the current existing web archiving solution has not robust standard security specifications to secure the content integrity in terms of implementation and operation.

2.2. Web Archive

The purpose of the web archive is the long-term preservation of current information from the web for descendants [1]. The web archive is one of the essential topics due to we have a responsibility to archive the existing web content for descendants. However, the challenge for the web archive is that the amount of web data are growing significantly and the web content is getting more important because the web is the primary communication channel to deliver sensitive personal information such as a legal contract. The web archive workflow can be described as four steps of the web archiving as shown in Figure 1 [6]. The appraisal and selection step evaluates the value of records and decides whether the records should be preserved and the preserving period. The acquisition step does archive the records by using several web technologies such as a web crawler. The organization and storage step decide which kinds of resources are organized and which storage will be used. Moreover, the description and metadata step generate metadata and description from the preserved the records for postprocess after archiving such as content searching.

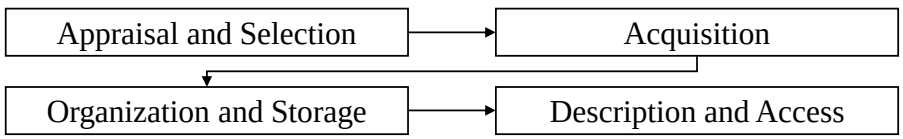

Figure 1. Four steps of web archiving.

In the acquisition step, many web content collection strategies have been developing by using web crawlers for the capture and archiving stage. An HTML is the standard web content markup language to display a web content and the HTML has been version up from HTML 1.0 to HTML 5.X until now. An HTML file consists of many linked resources to display web content such as an

external CSS (cascading style sheets) file, images, fonts, etc. An HTML has different characteristics like mentioned above, unlike another digital document standard format, which is PDF. In the case of PDF, it can embed all related resources into a single file [7]. Hence, storing an HTML file and related resources into a repository have vulnerary because there is no way to confirm whether all related resources are stored. Moreover, it needs to synchronize the version, such as saving the HTML and resources with the same timestamp.

IIPC has discussed WARC (web archive) file format, which is standard ISO 28500. WARC provides a standard way to structure, manage and store web content, which is collected from the web and elsewhere [8]. A WARC file format is used in most of the digital content archiving organizations such as the national library of Korea [9]. Many of archiving software such as Heritrix and wget provide web archiving feature to archive as a WARC file. A WARC file format has more than one WARC records and additional metadata such as header as shown Figure 2. In general, the first WARC record has the description for the WARC file and the other records after the first record have the result of crawling from a source such as HTML, images and CSS files. There are the defined-fields in WARC record format to describe each record's properties as shown in Figure 3. There are two defined-fields for content integrity, which are WARC-block-digest and WARC-payload-digest [8].

```
warc-file    = 1*warc-record
warc-record  = header CRLF
                   block CRLF CRLF
header       = version warc-fields
version      = "WARC/1.0" CRLF
warc-fields  = *named-field CRLF
block        = *OCTET
```

Figure 2. A web archive (WARC) file structure.

```
defined-field = WARC-Type
              | WARC-Record-ID
              | WARC-Date
              | Content-Length
              | Content-Type
              | WARC-Concurrent-To
              | WARC-Block-Digest
              | WARC-Payload-Digest
              | WARC-IP-Address
              | WARC-Refers-To
              | WARC-Target-URI
              | WARC-Truncated
              | WARC-Warcinfo-ID
              | WARC-Filename                ; warcinfo only
              | WARC-Profile                 ; revisit only
              | WARC-Identified-Payload-Type
              | WARC-Segment-Origin-ID       ; continuation only
              | WARC-Segment-Number
              | WARC-Segment-Total-Length    ; continuation only
```

Figure 3. Defined-fields in the WARC.

WARC-block-digest has a digest value of the full block of the record and WARC-payload-digest has a digest value of the payload referred to or contained by the record by using a hash algorithm such as an SHA-1. These two fields can be used for the proof of the content of WARC record integrity as shown in Figure 4. Figure 4 shows WARC-block-digest and WARC-block-digest has a digest value for the record, and the integrity of the WARC content can be confirmed by using these two fields. However, these two fields are optional defined-fields in the WARC specification, and there is no way to prevent to secure a WARC file against unauthorized modification manner. It tells a WARC file can be altered the digest value and the content together. It means there is no way to guarantee that a WARC file is the same as original and the digest value and record value are not revised. Hence, a WARC file format has a vulnerability against unauthorized content modification.

```
WARC/1.0
WARC-Type: response
WARC-Record-ID: <urn:uuid:a08fc169-aae6-487b-ab78-19ead02d1feb>
WARC-Warcinfo-ID: <urn:uuid:54ff6dde-e1fd-440c-9cae-ee50c078b8d2>
WARC-Concurrent-To: <urn:uuid:c161c896-4990-40de-8c34-ba1d527caae8>
WARC-Target-URI: http://www.naver.com/
WARC-Date: 2019-07-13T11:32:21Z
WARC-IP-Address: 210.89.164.90
WARC-Block-Digest: sha1:EQJJBPT6SS3GYE63UVJS6HYSYISOWVTA
WARC-Payload-Digest: sha1:4FBKOFIMXJPCK4TIQPGYLLB2A5A3SBRN
Content-Type: application/http;msgtype=response
Content-Length: 175773

HTTP/1.1 200 OK
Server: NWS
Date: Sat, 13 Jul 2019 11:32:20 GMT
Content-Type: text/html; charset=UTF-8
Transfer-Encoding: chunked
Connection: keep-alive
Set-Cookie:
PM CK loc = c652dd1f42634579826392464f2ced28e83ea967ad241309894b26cd29ff16f2;
Expires = Sun, 14 Jul 2019 11:32:20 GMT; Path = /; HttpOnly
Cache-Control: no-cache, no-store, must-revalidate
    <continue>
```

Figure 4. Example of WARC file.

2.3. Blockchain

Satoshi Nakamoto developed blockchain technology for bitcoin in 2008 [10]. It uses a dataset called a block, and these blocks are linked with the previous block by using cryptography as shown in Figure 5 [10]. This linked block set is called a blockchain. Each block contains the cryptography hash value of the previous block, a timestamp and a transaction dataset. A blockchain dataset stores across a peer-to-peer network to avoid the centralization of holding datasets, and it called distributed ledger. Anyone who attends the blockchain network can access and download the distributed ledger and can access each block data. A blockchain is widely used for cryptocurrency, customer contract storing in FSI (financial services industry) and many other enterprise area due to the characteristics of blockchain which are (1) all blockchain will be broken in case anyone block is damaged and (2) anyone who in peer-to-peer network can have the distributed ledger.

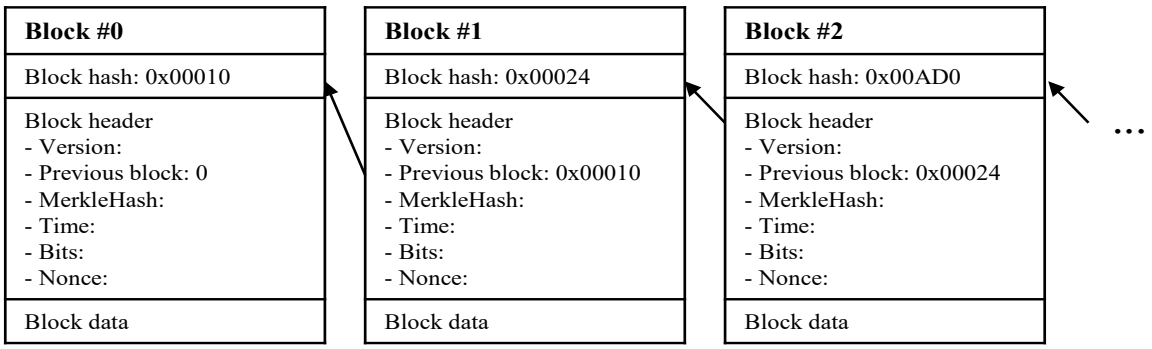

Figure 5. Blockchain with linked blocks.

Compared to a conventional centralized system, a blockchain has several advantages, including advantages in efficiency, security, resilience and transparency [11]. Most security solutions consider blockchain as a viable, robust and sustainable cybersecurity solution because of these characteristics [12]. Because of these characteristics, many enterprises consider introducing blockchain technology for their internal legacy system. For example, the e-document integrated support center and KISA,

which are one of Korea's government centers, use blockchain technology for an e-document delivery certificate [13].

Blockchain is a decentralized system, and there is the possibility that any user in peer-to-peer networks can create a new block, and it can be a conflict with another user. blockchain has the mechanism "proof of work" [14] to solve this possibility. This can be done by the Proof of Work algorithm, which finds the target value by using nonce value. Moreover, it has "difficulty" to adjust the number of a block can be created per minute [14]. The "difficulty" is increasing along with the number of nodes. For example, the Bitcoin can be created only one block per more than 10 min because of the "difficulty" [14].

2.4. Public Blockchain vs. Private Blockchain

A public blockchain is a permission-less blockchain. Anyone can add the data or read the data without any permission. Bitcoin is the most popular public blockchain, and anyone can download the Bitcoin blockchain to read or add data. A public blockchain such as the Bitcoin is decentralized and does not need to have a centralized entity that controls the network. Data on a public blockchain are secure as it is not possible to alter data once the data added into the blockchain. A private blockchain is a permissioned blockchain [15]. A private blockchain is a compromised solution between a public blockchain and traditional centralized server architecture. A private blockchain runs by the owners of the network who has controls, and they can accept a new candidate on the Internet. In a private blockchain, only the accepted participants can add a block and see the transaction, whereas the others cannot access it. Hyperledger Fabric of Linux Foundation [16,17] is one of the famous examples of a private blockchain. Various types of a private blockchain can be designed such as (a) anyone can read, but permissioned users only can write or (b) only permissioned users can read and write. Conversely, all customers can write the transaction information, whereas business users and authorization organizations can have only read permission to monitor all transactions in a financial enterprise.

A public blockchain and a private blockchain has a different characteristic like mentioned and as shown in Table 1. A public blockchain is more preferred for an enterprise that wants to do decentralizing. Moreover, private blockchain is more preferred for an enterprise that keeps controlling its security assets.

Table 1. Private and public blockchain.

Public Blockchain	Private Blockchain
Open, anyone can join the network	Restricted and permissioned
Low speed of transaction accomplishment	Fast speed of transaction accomplishment
Long transaction approval frequency	Short transaction approval frequency
High cost of each transaction	Comparatively cheap cost of each transaction
Anonymous	Known users
Large energy consumption	Low energy consumption

3. Design of BCLinked Web Archiving System

3.1. Web Archiving Method Based on Blockchain

One of the main challenges in the current web archive in the acquisition step is there is no integrated managed web content collection system, and there is no way to prove content integrity. Even though many organizations try to run a global standard web archive system for descendants, a user who uses the archived content cannot be sure their archived content is the same as the original. Moreover, there is no preventing system against unauthorized modification manner. A web archive system which has the content integrity proof can provide the trusted web archive.

We propose the new web archiving method based on blockchain to solve this challenge, and it records the web archiving activity and the content into the blockchain dataset. Once recording this information into the blockchain, the data cannot be deleted or altered. In general, the web archive system can collect the content from the various web site and collect the same page periodically to collect the latest content. Hence, the web archive system needs to care of web archive transaction for each domain. However, there is no relationship between each domain. It should be enough that the web archive system can recognize the life cycle of each domain. Hence, we define the BCLinked (blockchain linked) web archiving system, which has two levels of blockchain for the web archive as shown in Figure 6.

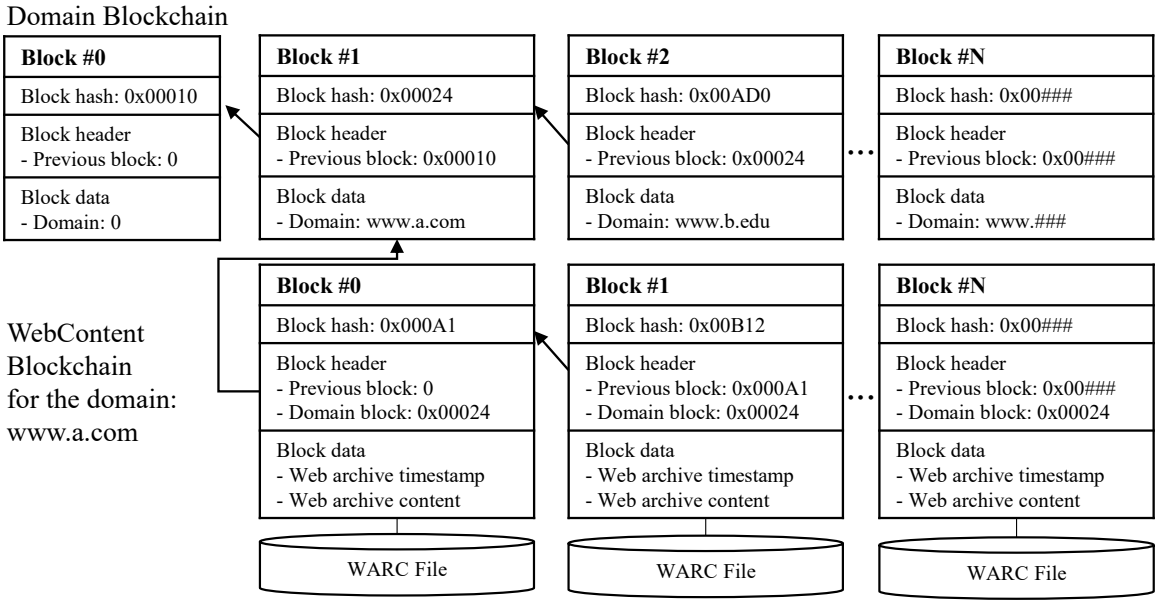

Figure 6. Blockchain architecture used in the blockchain linked (BCLinked) web archiving system.

The blockchain block in the first level which named domain blockchain contains a domain name, and a new block will be added in case a new domain is selected for web archive. The domain blockchain can be used for a web domain that exists on the web. The blockchain block in the second level which named webcontent blockchain, contains the linked information to the block in the domain blockchain and the web archive retrieval information. The number of webcontent blockchain will be the same as the number of blocks at the first level. All content of a WARC file can be too big to put into the blockchain block in the domain blockchain as data due to the limitation of the block size. Therefore, only the digest value which can identify the content will be added into the domain blockchain, and the original WARC file will be stored in a file repository. This web archive method provides a way to record the web archive process activity and the content information into the blockchain to secure content integrity.

3.2. Extension of WARC Record Format

We define fields for blockchain block data and a WARC record to integrate with blockchain and a WARC. The domain blockchain block will have each domain description, and the webcontent blockchain block will have each web retrieval result for each domain. Moreover, blockchain block has a specific block size, and all content of a WARC file can be too big to put into the blockchain block. So, we define the blockchain block data fields as shown in Table 2. Domain blockchain has WARC-domain and WARC-domain-creationtime fields. In case the web archiving system selects a new web domain to archive, then the information of the web domain will be added into a new block in the domain-creationtime blockchain. A new webcontent blockchain block will be added into the

webcontent blockchain after a WARC file is created. A webcontent blockchain block does not have all content of a WARC file, but only the fields to identify the WARC as shown in Table 2. A WARC file has multiple WARC records and these records per a WARC will be added into one webcontent blockchain node. It can be hard to synchronize a WARC file across participants who attend the web archive blockchain due to several issues such as file size, whereas the web archive blockchain can be synchronized. Hence, WARC-filename and WARC-location are used to keep the WARC file physical location and it will be added once into one block. WARC-record-ID, WARC-block-digest and WARC-payload-digest are used for each WARC record and it will be added multiple times because the number of a WARC record is greater than zero. These fields will be added as JSON format as shown in Figure 7.

Table 2. Definition of the domain blockchain and webcontent blockchain block data.

	Field	Value
Domain blockchain	WARC-domain	Web site domain name for archive
	WARC-domain-creationtime	Creation time of the node
Webcontent blockchain	WARC-filename	WARC filename
	WARC-location	WARC file location
	WARC-record-ID	block record UUID
	WARC-block-digest	block digest-algorithm ":" digest-value
	WARC-payload-digest	Payload digest-algorithm ":" digest-value

```
{
 "WARC-Filename":"at.warc",
 "WARC-Location":"server01/repository/20200113",
 "WARC-Digest-Values":,
 [{"WARC-Record-ID":"<urn:uuid:54ff6dde-e1fd-440c-9cae-ee50c078b8d2>",
   "WARC-Block-Digest":"sha1:W7EC4J2ZZFAUCYZCYMU5M3QTUPXO5QKU",
   "WARC-Payload-Digest":""},
  {"WARC-Record-ID":"<urn:uuid:a08fc169-aae6-487b-ab78-19ead02d1feb>",
   "WARC-Block-Digest":"sha1:EQJJBPT6SS3GYE63UVJS6HYSYISOWVTA",
   "WARC-Payload-Digest":"sha1:4FBKOFIMXJPCK4TIQPGYLLB2A5A3SBRN"}]
}
```

Figure 7. Data sample in webcontent blockchain block.

Once adding the information for the WARC file to the blockchain node, the new block is created. The description of the node will be added to the WARC file. We define the extended defined-fields for the WARC record to describe the blockchain node as shown in Table 3. These fields will be added to the top part of a WARC file.

Table 3. Extended defined-field for WARC record.

Field	Value
WARC-Added-Chain	True/False
WARC-domain-block	block hash of the domain blockchain
WARC-webcontent-block	block hash of the webcontent blockchain

3.3. Process of BCLinked Web Archiving System

The process of the BCLinked web archiving system is shown as Figure 8. A web crawler collects web content and creates a WARC file in the content crawling stage. All WARC records are sent to

BCLinked web archiving. BCLinked web archiving system checks whether the target domain exists in the domain blockchain with the WARC records. The existing blockchain node data will be used if the domain already exists or a new blockchain block will be added in the domain information storing stage. Then all WARC records are added to the webcontent blockchain block in the content information storing stage, and the block hash is added into WARC records for cross-validation in the blockchain information storing stage.

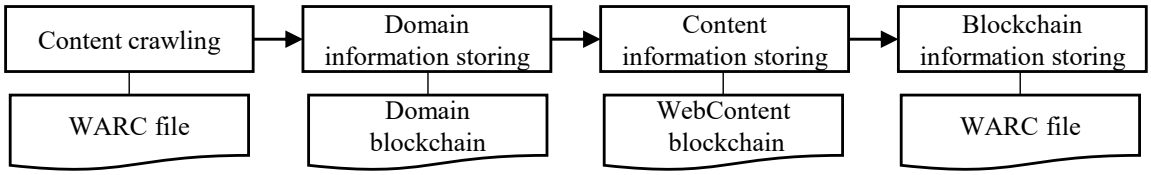

Figure 8. Web data archival process in the BCLinked web archiving method.

3.4. The System Architecture

BCLinked web archiving system consists of archive manager and blockchain manager as shown in Figure 9. The subcomponents in the archive manager and the blockchain manager can be classified into following below cases:

1. The crawling processor, it is for web content crawling;
2. The WARC handler, it is for adding the extended WARC fields and managing a WARC file storing in the repository;
3. The blockchain requester, it is for communicating with blockchain manager to request to add the new web content information into the blockchain;
4. In the blockchain manager, the blockchain requester handler is the interface to communicate with the archive manager;
5. The blockchain node handler, it is for the blockchain data processing.

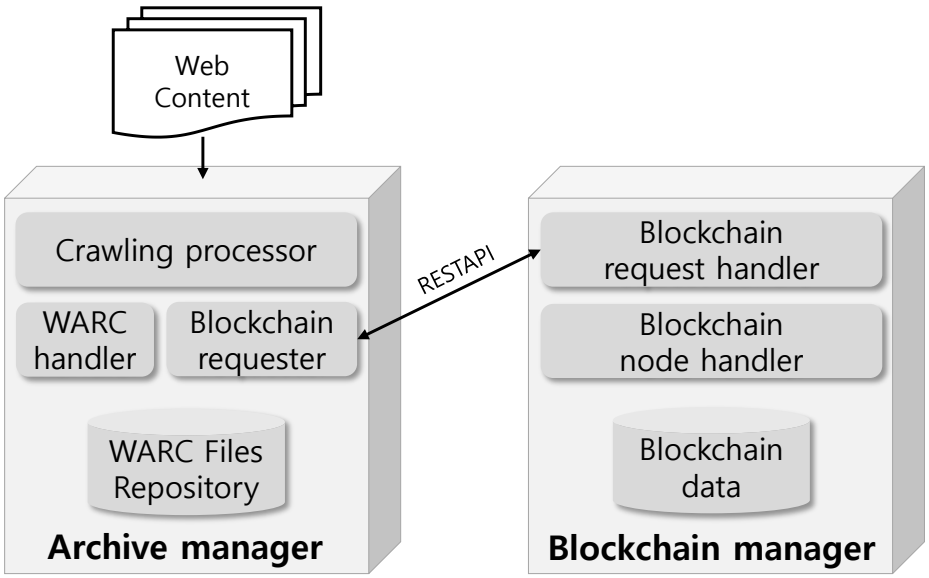

Figure 9. System architecture of the BCLinked web archiving system.

In general, the web archiving process happens in an enterprise or government which needs to keep the historical record. Hence, this BCLinked web archiving system does not need to give all

permission to a user, but only giving read permission to a user is sufficient. Therefore, the BCLinked web archiving system can be run in a private blockchain. The council of organizations can run the BCLinked web archiving system, and the blockchain data will be synchronized via the network. However, the WARC files in the archive manager does not need to be synchronized because it requires heavy network traffic. Hence, we designed each organization in the council to keep their WARC files into their WARC repository, and only the metadata of the WARC file will be added and synchronized via the blockchain manager.

Both the archive manager and the block manager runs on each independent web server and it has the RESTful API to start a new archive process and communicate with each other as shown in Table 4. These two the archive manager and the blockchain manager can be run in separate physical enjoinment as well as in a single physical environment. The archive manager can be triggered by/crawl API. The archive manager crawls web content and returns the unique job ID. Then the/archive API will be triggered to add the information into the blockchain. The/archive API will execute the/blocks/add API and return the result. Then the/update API will record the blockchain hash information into the WARC file, and the process is finished. A user can check all blockchain data by the/blocks/list API.

Table 4. RESTful APIs in the BCLinked web archiving system.

	API Name	Description
The archive manager	/crawl	collect new web content :param target_url: target web page URL :return: the object JOD ID
	/archive	archive new web content to the blockchain :param jobid: the object JOD ID :return: WARC-domain-block WARC-webcontent-block
	/update	update the WARC :param jobid: the object JOD ID domain_block: WARC-domain-block wcontent_block: WARC-webcontent-block :return: Boolean
The blockchain manager	/blocks/add	add new block :param location: the WARC file location target_url: target web page URL digest: UUID and digest values dictionary :return: block hash information dictionary
	/blocks/list	list blocks :return: blocks list

4. Experiment and Analysis

4.1. Experiment Environment

We verified the BCLinked web archiving system described in this paper through an experiment. We develop the BCLinked archiving system by using open-source software as shown in Table 5. We used the well-known web crawler, which is wget for the crawling processor. A wget supports writing to a WARC file like Heritrix and other archiving tools since version 1.14 [18]. We developed the archive manager and the blockchain manager by using python language.

The BCLinked web archiving system has the PoW (Proof of Work) algorithm for avoiding conflict to create a new block at the same by several requests. We set the difficulty is as shown in Figure 10. This difficulty depends on various situations. For example, the difficulty of the Bitcoin is adjusted periodically depends on how much hashing power has been distributed on the network. In this experiment, we decided this difficulty based on previous research experience.

Table 5. Experiment environment.

Hardware	Intel i7 CPU/16 Gbyte RAM
OS	Microsoft Windows 10 Professional
Linux emulation	Linux Cygwin emulation
Web crawler	wget
Development language	Python
Database	MySQL

```
CONST difficulty_value
INIT current_block_trial_proof
length_difficulty_value = length (difficulty_value)

WHILE
    guess_value = COMBINE (previous_block_proof, current_block_trial_proof)
    guess_value_hash = hash-sha256 (guess_value)
    IF last_digit (guess_value_hash, length_difficulty_value) = difficulty_value
        RETURN TRUE
    ELSE
        INCREMENT current_block_trial_proof
    END IF
END WHILE
```

Figure 10. Algorithm of PoW (Proof of Work).

4.2. Experiment Results

We experimented with archiving web content by the BCLinked web archiving system and the traditional web archiving system. We had two different experiments with different difficulty values; the result summary is shown in Table 6.

In the first experiment, we set "0000" as the difficulty value, and we selected the most popular 81 websites global and archived the index page. We archived these index pages 100 times to simulate web pages are archived periodically. The total processing time is 3575.08 s and 78.67 s of the total processing time is for the blockchain processing time. The total WARC file size is 120,459,510 bytes, and the blockchain data file size is 1,142,594 bytes with the BCLinked web archiving system. Meanwhile, the blockchain processing time and the blockchain data set is 0 with the traditional web archiving because the traditional web archiving system does not use the blockchain technology, but only use the web crawler. The minor difference regarding total WARC files size between the BCLinked web archiving system and the traditional web archiving system is because of the updating a web content from the source web site. In the second experiment, we set "00000" as the difficulty value, which is more difficult to find the proof value. The total processing time is 5049.47 s and 1011.34 s of the total processing time is for the blockchain processing time. The total WARC file size is 121,408,987 bytes, and the blockchain data file size is 1,143,229 bytes. Meanwhile, the result of the traditional web archiving system shows very similar results like the first experiment because the difficulty property is used only for the BCLinked web archiving system.

We confirmed the blockchain processing time is uniform regardless of the number of blocks and the blockchain processing time only depends on the difficulty property as shown in Figure 11. Hence, there is no unexpected overhead time consumption while archiving a new web content. A web archiving solution collects a web content periodically when the content is updated and the BCLinked

web archiving system only requires a small additional blockchain processing time for adding a new content.

Table 6. Archiving experiment results.

System		BCLinked Web Archiving	Traditional Web Archiving
Number of archiving		810 (81 sites × 100 times)	810 (81 sites × 100 times)
Difficulty 0000	Total processing time	3575.08 s	3493.12 s
	Blockchain processing time	78.67 s	0 s
	Total WARC files size	120,459,510 bytes	120,375,023 bytes
	Blockchain data size	1,142,594 bytes	0 bytes
Difficulty 00000	Total processing time	5049.47 s	3380.76 s
	Blockchain processing time	1011.34 s	0 s
	Total WARC files size	121,408,987 bytes	120,946,193 bytes
	Blockchain data size	1,143,229 bytes	0 bytes

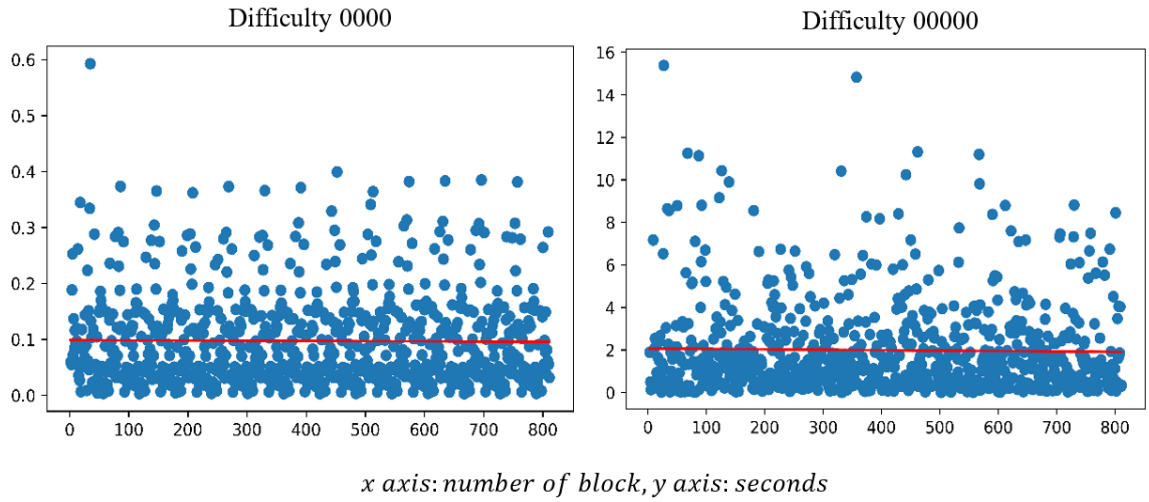

Figure 11. Processing time of blockchain.

We confirmed the all WARC files block-digest value was stored in the BCLinked web archiving system, and the block-digest in the blockchain can be used for the content-integrity experiment. In terms of the content integrity, we tried to modify the archived content by using the python script to simulate the unauthorized content modification. The python script modified a content, WARC block-digest, WARC payload-digest and rebuilding the BCLinked blockchain data in the BCLinked web archiving system. Moreover, the python script modified a content, WARC block-digest, WARC payload-digest in the traditional web archiving system without the BCLinked blockchain data. As a result, it took less than 1 s to modify the archived web content in the traditional web archiving system because there is any linked relationship to another web contents, and it can be done by modifying only the original archived web content. However, it took 81.99 s with difficulty 0000 and 998.73 s with difficulty 00000 in the BCLinked web archiving system because all blockchain node data should be rebuilt even if only one single content is modified.

The processing time of web content archiving can be express as shown below Figure 12. The overhead time in the BCLinked web archiving system is $N \times B_t$. The time complexity for both S_{a1} and S_{a2} is the same as $O(n)$ and B_t is much smaller than C_t. Therefore, this overhead time is reasonable to get the benefit from the blockchain.

The processing of web content modification can be express as shown below Figure 13. The overhead time in the BCLinked web archiving system is mostly in updating all linked blockchain data. The worst

case is that trying to modify the web content which is in the beginning position in the blockchain data. All linked blockchain data and extended defined-fields should be revised in this case, and T will be the total number of all blockchain data. The time complexity for S_{a2} is $O(n^2)$ while the time complexity for S_{a1} is $O(n)$. It shows that unauthorized content modification is almost impossible due to massive overhead processing time in the real world with the BCLinked archiving system.

$$S_{a1} = N \times C_t$$
$$S_{a2} = N \times C_t + N \times B_t$$

$S_{a1} =$ *Time to add web contents in the Traditional web archiving*
$S_{a2} =$ *Time to add web contents in the BCLinked web archiving*
$N =$ *The number of web contents to be handled*
$C_t =$ *Average time of a web content crawling*
$B_t =$ *Average time of adding blockchain data*

Figure 12. Time complexity of web content archiving.

$$S_{b1} = N \times M_t$$
$$S_{b2} = N \times (T \times M_t) + N \times (T \times B_t)$$

$S_{b1} =$ *Time to modify web contents in the Traditional web archiving*
$S_{b2} =$ *Time to modify web contents in the BCLinked web archiving*
$N =$ *The number of web contents to be handled*
$M_t =$ *Average time of a web content modification*
$T =$ *The* amount *of blockchain data from the current position to the end*

Figure 13. Time complexity of web content modification.

In conclusion, we confirmed the BCLinked web archiving system can provide have enhanced content integrity with small overhead than the traditional web archiving system.

5. Conclusions

In this paper, we proposed the BCLinked web archiving system to preserve content integrity by using blockchain technology and extending WARC specifications. blockchain technology is that connecting the previous node by using cryptography to secure all data in a blockchain, and it is the decentralization system which can be shared via peer-to-peer network. By using the BCLinked web archiving system, the block-digest value of a WARC file to identify content integrity is stored in the blockchain. The BCLinked web archiving can solve the weakness of the current web archiving in terms of securing content integrity. The BCLinked web archiving system can be used for a user who retrieves a web content from the BCLinked web archiving system and gets the confirmation of content integrity. Therefore, this system will be useful for an enterprise or a government that generated personalized web content to a customer because it can be proof against legal disputes. However, we developed the system for experiments with simple blockchain by using python. This system is enough for proof of concept, but we need to research more-with-more stable and demonstrated blockchain platforms such as Ethereum. In future research, we will research the more efficient BCLinked web archiving system in a production environment with the demonstrated blockchain platform.

Author Contributions: Formal analysis, H.C.H.; funding acquisition, J.S.P.; methodology, H.C.H. and J.G.S.; project administration, J.S.P.; resources, H.C.H. and J.S.P.; writing—original draft, H.C.H.; writing—review and editing, H.C.H., J.G.S. and J.S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

References

1. Wikipedia. Web Archiving. Available online: https://en.wikipedia.org/wiki/Web_archiving (accessed on 14 August 2019).
2. Teraoka, T. Organization and exploration of heterogeneous personal data collected in daily life. *Hum. Cent. Comput. Inf. Sci.* **2012**, *2*, 1–15. [CrossRef]
3. Consultative Committee for Space Data System. Reference Model for an Open Archival Information System (OAIS). Available online: https://public.ccsds.org/pubs/650x0m2.pdf (accessed on 14 August 2019).
4. Park, B.J.; Cha, S.J.; Lee, K.C. Data Mapping between Korea Deep Web Archiving Format and Reference Model for OAIS. In Proceedings of the Korean Information Science Society Conference; Korean Institute of Information Scientists and Engineers: Seoul, Korea, 2010; pp. 197–200.
5. Deloitte. Breaking Blockchain Open—2018 Global Blockchain Survey. Available online: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/financial-services/us-fsi-2018-global-blockchain-survey-report.pdf (accessed on 20 August 2019).
6. Jinfang, N. An Overview of Web Archiving. *D Lib Mag.* **2012**, *18*. Available online: https://scholarcommons.usf.edu/si_facpub/308/ (accessed on 20 August 2019).
7. Adobe. Adobe PDF Reference. Available online: https://www.adobe.com/content/dam/acom/en/devnet/pdf/pdf_reference_archive/pdf_reference_1-7.pdf (accessed on 21 January 2020).
8. International Internet Preservation Consortium. The WARC Format 1-1. Available online: https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1 (accessed on 1 September 2019).
9. National Library of Korea. The Website Building Guide for the OASIS Web Archiving System. Available online: http://www.oasis.go.kr/about/guide.do (accessed on 18 August 2019).
10. Lemieux, V.L. Trusting records: Is Blockchain technology the answer? *Rec. Manag. J.* **2016**, *26*, 110–139. [CrossRef]
11. Kim, H.W.; Jeong, Y.S. Secure authentication-management human-centric scheme for trusting personal resource information on mobile cloud computing with blockchain. *Hum. Cent. Comput. Inf. Sci.* **2018**, *8*, 11. [CrossRef]
12. Huh, J.H.; Seo, K. Blockchain-based mobile fingerprint verification and automatic log-in platform for future computing. *J. Supercomput.* **2019**, *75*, 3123–3139. [CrossRef]
13. Korea Internet & Security Agency. FAQ Regarding a Digital Information Delivery and a Delivery Certificate. Available online: https://www.npost.kr/pages/notice/notice_0106.jsp (accessed on 20 January 2020).
14. Nguyen, G.T.; Kim, K. A Survey about Consensus Algorithms Used in Blockchain. *J. Inf. Process. Syst.* **2018**, *14*, 101–128.
15. Cachin, C. Architecture of the Hyperledger Blockchain Fabric. *Workshop on Distributed Cryptocurrencies and Consensus Ledgers*; 2016. Available online: https://www.zurich.ibm.com/dccl/papers/cachin_dccl.pdf (accessed on 20 August 2019).
16. Jeong, W.Y.; Choi, M. Design of Recruitment Management Platform Using Digital Certificate on Blockchain. *J. Inf. Process. Syst.* **2019**, *15*, 707–716.
17. Opensource Blockchain Technology. Hyperledger. Available online: https://www.hyperledger.org (accessed on 20 January 2020).
18. GNU. Wget. Available online: https://www.gnu.org/software/wget/manual (accessed on 20 January 2020).