


Article

Design of Chained Document HTML Generation Technique Based on Blockchain for Trusted Document Communication

Hyun-Cheon Hwang¹ and Woo-Je Kim^{2,*} 

¹ Graduate School of Public Policy and Information Technology, Seoul National University of Science and Technology, Seoul 01811, Korea; a.hwang@seoultech.ac.kr

² Department of Industrial Engineering, Seoul National University of Science and Technology, Seoul 01811, Korea

* Correspondence: wjkim@seoultech.ac.kr

Abstract: Digital document communication between an enterprise and a customer is becoming a primary form of communication rather than the traditional physical document communication. A PDF document, the most popular document format, provides an identical document layout regardless of OS or device and has a content integrity verification feature with a digital signature. However, it has a bad user experience, such as low readability on a mobile device. On the other hand, an HTML document has a weakness in verifying the content integrity even though it is the primary document format and provides a good user experience on mobile devices. There are certified document services using blockchain technology, but it is still vulnerable to verifying content integrity. Furthermore, research on the document HTML has proposed the trusted document generation technique by HTML conformance and digital signature; however, this research does not provide content delivery verification, and there is a file size overhead. In this paper, we have developed the chained document HTML by defining HTML conformance, digital signature, and blockchain technology. First, the chained document HTML has to embed all resources and does not allow loading content on-demand. Second, the file is signed by a digital signature, and the signature value is added in the file header. Lastly, the metadata to verify the content integrity is inserted in a blockchain node. We have created the chained document HTML generation and verification experiment environment by Ethereum and Python. We have confirmed that the chained document HTML provides content and delivery integrity verification in the research. We expect the chained document HTML will be widely used in document communication between an enterprise and a customer, especially if the document has sensitive personal information that might have a legal dispute.

Keywords: chained document HTML; digital document; blockchain; document communication



Citation: Hwang, H.-C.; Kim, W.-J. Design of Chained Document HTML Generation Technique Based on Blockchain for Trusted Document Communication. *Electronics* **2022**, *11*, 1006. <https://doi.org/10.3390/electronics11071006>

Academic Editor: Christos J. Bouras

Received: 24 February 2022

Accepted: 22 March 2022

Published: 24 March 2022

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



Copyright: © 2022 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/>).

1. Introduction

The definition of an electronic document is described as “recorded information structured for human consumption” [1]. This definition also accommodates various documents such as contracts, blueprints, email, and video clips. The definition of the electronic document in Korean law is the information that is sent, received, or archived, which is created or converted electronically by the information processing system [2]. It means the electronic document is not just the traditional paper layout document but all kinds of digital information that can be recorded and reviewed, including video and audio. We can distinguish an electronic document into those that focus on representing the content with human readability and those that focus on exchanging information between systems. A document format such as an XML (Extensible Markup Language) is used to exchange information between information systems, whereas a PDF (Portable Document Format) is used to present as a traditional paper document.

Digital document communication by an electronic document and digital delivery channel between an enterprise and a customer has been a significant communication

channel since the digital era. There are various digital document formats for the content representation with human readability, such as MS Word, HWP, PDF, PostScript, Jpeg, Tiff, and HTML. A PDF document is the most popular digital document format because it has the same layout as a physical document, independent format against devices and OS, and a content integrity verification feature by the digital signature. However, a PDF document provides a weak readability feature on a mobile device, and there is no way to track the document delivery. An HTML document has been getting more critical than a PDF document in recent years, along with the mobile era. An HTML document is the primary front-end language on a mobile device. An HTML document provides a good customer experience, such as being easy to navigate and read, and responsive interaction. However, an HTML document has weak verification features for content integrity compared to a PDF document as there is no standard content integrity verification feature. Besides, it also does not provide to track the document delivery like a PDF document.

There are researches and attempts regarding digital document communication to ensure content integrity. The IT service company DocuSign provides a digital content signature and delivery service [3], but it is based on PDF. The Korea Internet and Security Agency provides the Certified Electronic Document Intermediary service by using blockchain technology to deliver digital documents such as PDF or HTML [4]. This service stores content and delivery metadata in a blockchain node to verify content and delivery integrity for various dispute situations. However, the service has the vulnerability to verify content integrity in cases where a digital document consists of multiple separated resource files. Another research has proposed embedding linked resources and adding a digital signature for content integrity to overcome an HTML document weakness [5]. The document HTML from the research provides strong content integrity like a PDF document and a good user experience on a mobile device. However, this research only solves the content integrity problem in terms of the document content perspective, even though the content delivery verification is also a critical section. Moreover, the document HTML has some file size overheads due to embedding all metadata to verify the content integrity.

In this paper, we have researched blockchain technology and designed the chained document HTML specification can have content integrity. The chained document HTML is the restricted HTML specification to ensure content integrity along with the blockchain technology. The chained document HTML has been proposed to embed a related resource and prohibit loading content on demand. Moreover, the chained document HTML does digital signature for the content integrity, and digital signature value and document delivery tracking information are stored in a blockchain node. The chained document HTML based on the blockchain can verify both content and delivery integrity by a decentralized mechanism. We did the related research for the digital document and the blockchain in Chapter Two, designed the chained document HTML based on the blockchain in Chapter Three, validated the chained document HTML by the experiment in Chapter Four, and compared the chained document HTML to the current digital document formats in Chapter Five. Finally, we have concluded the value of this research in Chapter Six.

2. Related Research

2.1. PDF and HTML

PDF is the ISO 32000 standard specification and developed by Adobe [6]. PDF document is an independent format against devices and OS, and it includes all necessary resources inside of the file to represent the content. A PDF document provides identical representation on any device and any OS. PDF documents are the de facto standard electronic document, and Adobe has estimated that there were 2.5 trillion PDF documents in circulation [7]. Furthermore, PDF has the specification to validate the content integrity by using the digital signature [8]. The private key can sign the content based on the PKI certificate, and the signed value and the public key can be added to the PDF document. The signed PDF document can be verified by a document consumer using the standard PDF

document viewer software. Lots of the documents, which contain personal information from an enterprise to a customer, use the PDF document because of these characteristics.

Even though PDF is the de facto standard document format, it is challenging in the mobile environment. PDF is born to replace the traditional paper, and it shows the best value when a user reads a PDF document on a big screen device such as a laptop. However, PDF does not support dynamic re-flow content, unlike HTML, and it can provide a low readability experience on a mobile device. Moreover, most users expect two-way communication on the mobile device, like to submit an inquiry for the next step after reading content. However, PDF has a limited interactive communication feature, whereas HTML has an interactive communication feature.

HTML is the markup language for the web, and the latest version is HTML5.2 [9]. HTML5 supports multimedia content without third-party software, and the responsive HTML technique provides the best user experience on various devices. As a single HTML can cover various devices, the hybrid application development methodology, which uses HTML content as a part of the primary content representation in mobile applications, is widely used. However, the cons of HTML in terms of a digital document is that HTML requires related external resources such as an image, and there is no standard feature to validate the content integrity. Because of that, digital document based on HTML has the vulnerability of the long-term archive and to be a trusted document. Hence, many enterprises are servicing both HTML digital documents for user experience on mobile devices and PDF digital documents for content integrity.

2.2. Document HTML

The document-HTML is the proposal to eliminate the vulnerabilities of an HTML documents to act as a trusted electronic document [5]. It is hard to treat an HTML document as a single document because the HTML file and the resource file exist separately. PDF documents provide a content integrity verification method by a digital signature, whereas HTML documents do not provide a standard protocol for content integrity verification. The Document-HTML solves the weakness of an HTML document as a trusted document through the following rules.

- (a) The encoding of HTML documents should be UTF-8.
- (b) All related resources such as an image for an HTML document must be embedded and do not compose the document by linking external resources. HTML documents have to be single files like PDF and do not allow linking to external resources.
- (c) Do not use <audio>, <video> multimedia tags. Audio and video tags are difficult to be included in a document due to the file size problem, and these tags are not essential as a document. Therefore, the <audio> and <video> multimedia tags, which are multimedia tags provided by HTML5, are not used.
- (d) Do not use external resources container tags which are <object>, <iframe>, <embed>, <param>. The external resource container tags are hard to be embedded in a single document, and the function of the tags can be shown depending on the device. Moreover, these tags are not an essential element as a document. Therefore, <object>, <iframe>, <embed>, <param> tags, which are external resources container tags, are not used.
- (e) Do not allow asynchronous data loading. After the HTML document is opened, it is not allowed to load data by an asynchronous call method using a scripting language such as JavaScript jQuery library. Data itself is not included with asynchronous data loading in the HTML document, and HTML document content can be different based on data loading. The asynchronous data loading can be a vulnerability to the content integrity of the HTML document.
- (f) The content of the HTML document must be digital signed. The digital signed hash value and the digital signature public key must also be included so that they can be verified.

The document HTML structure is as shown below in Figure 1. The header of the document HTML document contains a digital signature value to verify the authenticity of the document itself, and the linked resource is included in the document HTML document using the data URL Scheme.

```
<!--BEGIN DOCUMENT-HTML-TYPE
DocHTML5 v0.1
END DOCUMENT-HTML-TYPE-->
<!--BEGIN DOCUMENT-HTML-CONTENT-DIGEST
HEXDIGEST_VALUE
END DOCUMENT-HTML-CONTENT-DIGEST-->
<!--BEGIN DOCUMENT-HTML-CONTENT-SIGNED-DIGEST
HEXDIGEST_VALUE
END DOCUMENT-HTML-CONTENT-SIGNED-DIGEST-->
<!--BEGIN DOCUMENT-HTML-CONTENT-VALIDATION-KEY
HEXDIGEST_VALUE
END DOCUMENT-HTML-CONTENT-VALIDATION-KEY-->
<!DOCTYPE HTML>
<html>
<head>
  <title>Document-HTML Sample</title>
</head>
<body>
  <div>
    
  </div>
</body>
</html>
```

Figure 1. Document HTML structure with extended tag.

This document HTML provides the best user experience for web and mobile, and it provides content integrity as much as the PDF document. In addition, it is suitable for the long-term archive as an HTML document is a single document with all embedded resources. However, it requires the central management system to generate the document HTML and validate the document HTML. In addition, it is also a critical point to prove of delivery of the document HTML from the producer to the consumer. However, there is no way to validate the delivery of the document HTML.

2.3. Blockchain

Blockchain technology was developed by Satoshi Nakamoto for bitcoin in 2008 [10]. The first fundamental of the blockchain is that data is stored in a block with the previous block hash value, which is calculated by using the data in the previous block. Therefore, blocks are linked to the previous block, and any block cannot be altered once a block is added in the blockchain node unless to change all blocks from the genesis block, the first block. Blockchain technology provides immutability with this characteristic. The second fundamental is decentralization. Anyone can download the full dataset of the blockchain, and any user can contribute to adding the new block into the blockchain. Blockchain has a consensus algorithm to avoid conflict from the many users who simultaneously try to add a new block. Blockchain has been started for cryptography, which is bitcoin. In addition, blockchain is becoming used widely in various domains that require reliable transaction archival because of its immutability and decentralization. Current cyber risk concerns are founded on intelligent systems capable of human–computer interactions [11] including an electronic document system; the blockchain mitigates the current cyber risk by the decentralized immutable distributed ledger.

Bitcoin, the most famous cryptography by blockchain, allows any anonymous user to attend the bitcoin blockchain network, and anyone can be a peer. This is called the public blockchain or permissionless blockchain [12]. The public blockchain allows joining the blockchain network by anonymous, and it gives a robust decentralization methodology as anyone can be a peer. However, it requires more processing time to run the algorithm

for consensus to append a new block into the blockchain network. It could be impossible to identify who has attended the network. On the other hand, private blockchain allows joining the blockchain network to whoever has authorization [12]. It provides partial decentralization, but it is valuable if only authorized entities are needed to be members of the blockchain network. It can remove the potential risk by anonymity and has a faster processing time to run the consensus algorithm. In a specific domain such as finance, a private blockchain is used more than a public blockchain. There are different blockchain technologies, as shown in Table 1; Ethereum is one of the public blockchains and Hyperledger is one of the private blockchains.

Table 1. Public blockchain vs. private blockchain.

Public Blockchain	Private Blockchain
Open network	Closed network
Anyone can join the network	The user who is invited can join the network
Anonymous	Identified
Bitcoin, Ethereum	HyperLegder
Anonymous	Nonymous
Bitcoin, Ethereum	Hyperledger

A service by blockchain is undertaken by the smart contract. The smart contract is the program that is run on the blockchain network. Once the smart contract is developed, the binary code of the smart contract is added in the blockchain network as blockchain data like a normal transaction, and it will be triggered for a service. As the smart contract is one of the types of data in the blockchain, it has immutability. It requires the front-end application to trigger the smart contract for a user or a system. DApp (decentralized application) is an application for blockchain, and works with the smart contract to manage the blockchain data.

3. Chained Document HTML Generation Technique

3.1. Chained Document HTML Conformance

There is no standard specification to preserve the content integrity in HTML, and an electronic document based on a digital signed PDF is the de facto standard to preserve the content integrity. We have proposed the document HTML generation technique in the previous research to preserve the content integrity in which an electronic document based on HTML can be a trusted document [5]. However, this previous research has a weakness, which is the file size overhead due to the certificate value for the content integrity verification in the header, and vulnerability of non-repudiation for the document delivery. The chained document HTML is the expanded version of the document HTML. The chained document HTML stores the certificate value in the blockchain transaction instead of in the document itself to reduce the file size overhead and the document delivery history can be proved by the blockchain transaction history which is immutable.

The chained document HTML has the same conformance as the document HTML to be a single document content as below:

- All related resources must be in an HTML document internally;
- An HTML document can be opened in a standard web browser without any additional software.

3.2. Design of the Chained Document HTML Structure

The chained document HTML is a subset definition of HTML, and it does not require any third-party software to open the file in a browser to view content. A standard HTML document requires external resources to compose content to display. For the chained document HTML, the data URI scheme, which is the standard specification as RFC 2397 [13], is used to convert and embed external resources to internal resources with BASE64 format, as shown in Figure 2.


```
data:[<mediatype>][;base64],<data>
```

Figure 2. Data URI scheme.

The chained document HTML has the extended header definition for content integrity, as shown in Figure 3. The structure of extended header information used the HTML comment tag so that it will not have any impact on the original HTML5 content rendering. The extended header DOCUMENT-HTML-TYPE indicates the document type and the version of the chained document HTML. In this research, we use the constant value, which is CDocHTML5 v0.1. The DOCUMENT-HTML-CONTENT-DIGEST has the hash algorithm name and the HTML content hash digest value with a colon (:) to distinguish the algorithm name and the HTML content hash digest value. The HTML content except the extended header definition section is calculated for the HTML content hash digest value. The DOCUMENT-HTML-CONTENT-SIGNED-DIGEST has the digital signed value of the DOCUMENT-HTML-CONTENT-DIGEST value. The document creator signs the DOCUMENT-HTML-CONTENT-DIGEST value by using a private key, and the signed value is recorded in the DOCUMENT-HTML-CONTENT-SIGNED-DIGEST. TRANSACTION-ADDRESS has the blockchain transaction address to verify the content integrity.

```
<!--BEGIN DOCUMENT-HTML-TYPE
CDocHTML5 v0.1
END DOCUMENT-HTML-TYPE-->
<!--BEGIN DOCUMENT-HTML-CONTENT-DIGEST
Hash algorithm:hash digest value
Ex) sha256:884DE890BDCABE2EE65DD5124BF2AB00286B16918AC78A77C89EDB28B6303623
END DOCUMENT-HTML-CONTENT-DIGEST-->
<!--BEGIN DOCUMENT-HTML-CONTENT-SIGNED-DIGEST
Digital signed value
Ex) CmW5TooDnBamnvchi0r3kJ88KB+/TTwxL8Mm1b4umdyVcrCQMP1CJzdIo414yrYR
EqcLIioqcGfpxnIQYdyjLFaw9rzf916jhuwSIFAjR5B9H/Zm4N01+LSkQ+KtbYI3
8bliff5SA2vkb/oWT34s1JPXxjCR3/8Z8nmjqOPkEFQPFmskLePUOoetscUC0JI5
zjD0P+1mQxmVXtyR7Y3w3MQYrmNI3qyn6SCD0k3MUHyh70LPSRgKUAhwaS5HPnV
mP+1N1am+CGg3GPeKRNXQkysipqxBRWWIk8T83X1bJKZeDBdiGTSz22PCmJU3bwz
e2jg+i7ntcAPX5kiLdqRmQ==
END DOCUMENT-HTML-CONTENT-SIGNED-DIGEST-->
<!--BEGIN TRANSACTION-ADDRESS
Blockchain transaction address
Ex) 0x091ff0ca58c349f27ade41e4dd3f019e5473d5ce1061d89060e3d7e561d067fe
END TRANSACTION-ADDRESS-->
<!DOCTYPE html>
<html>
<head></html>
<body></body>
</html>
```

Figure 3. Extended header information of the chained document HTML.

Hence, the chained document HTML structure can be expressed, as shown in Figure 4. The HTML content, in which all resources are embedded, is the target area for content integrity. The content hash digest value is stored in the DOCUMENT-HTML-CONTENT-DIGEST, and the digital signed value of the DOCUMENT-HTML-CONTENT-DIGEST is stored in the DOCUMENT-HTML-CONTENT-SIGNED-DIGEST. The validation value, including the public key, is stored in the blockchain, and the blockchain transaction address is stored in the DOCUMENT-HTML-TRANSACTION-ADDRESS.

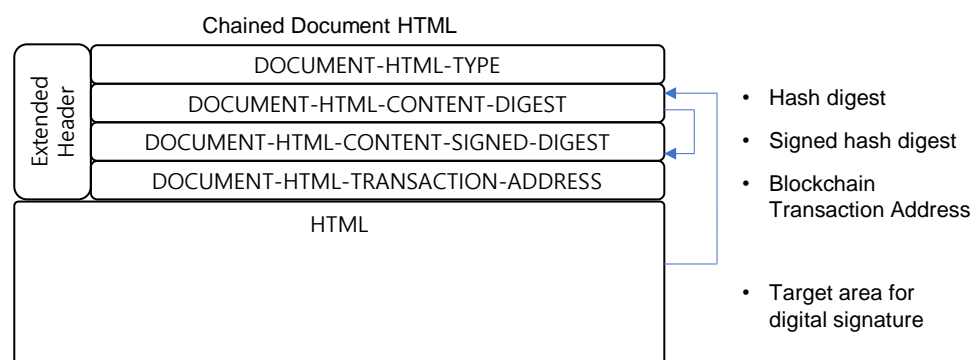


Figure 4. The Chained Document HTML Structure.

3.3. Design of the Chained Document Blockchain Structure

A user who opens the chained document HTML needs to validate the fact that the content has not been altered after it was created. So, the validation value for the content integrity is required for the chained document HTML. The chained document HTML uses the blockchain to store the validated value for content integrity. The validation values are shown in Table 2. Timestamp shows the document creation date and time, Creator is the document creator, and the value of the DOCUMENT-HTML-CONTENT-DIGEST from the chained document HTML file is also stored in the blockchain in the same key. Also, the public key and the certificate information are stored in the Certificate.

Table 2. The validation values in the chained document Blockchain.

Key	Description
TIMESTAMP	Timestamp for the document creation
CREATOR	Document creator
DOCUMENT-HTML-CONTENT-DIGEST	Hash digest value of the HTML content
CERTIFICATE	Public key for the document-HTML-content-signed-digest

Each chained document HTML is linked to the transaction of the chained document Blockchain, and the validation values can be found in the transaction value, as shown in Figure 5. First, the chained document HTML and the transaction of the chained document Blockchain must have the identical DOCUMENT-HTML-CONTENT-DIGEST to identify the content has not been altered. Second, CERTIFICATE value is used for the verification of the DOCUMENT-HTML-CONTENT-SIGNED-DIGEST. CERTIFICATE has the digital signature certificate information and the public key to verify the DOCUMENT-HTML-CONTENT-SIGNED-DIGEST. The unsigned value from DOCUMENT-HTML-CONTENT-SIGNED-DIGEST should be identical to DOCUMENT-HTML-CONTENT-DIGEST, which is in the chained document HTML header.

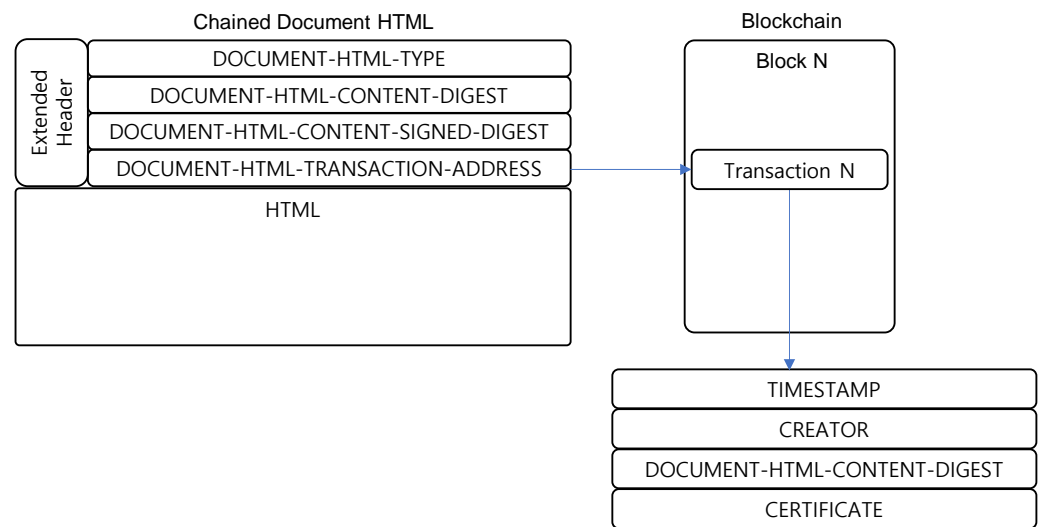


Figure 5. The chained document HTML Structure with Blockchain.

3.4. Chained Document HTML Generation and Verification Process

The chained document HTML system has two major processes: the chained document HTML generation process from a standard HTML and the chained document HTML validation process. It has four major sub-sequence processes for the chained document HTML generation process, as shown in Figure 6. First, a standard HTML file will be validated whether it has incompatible HTML objects for the chained document HTML. Second, all external resources in the HTML document will be downloaded and converted to internal resources using the Data URI scheme. Third, the encrypted hash digest value to be used for the document validation will be added to the blockchain as blockchain transaction data. Lastly, the metadata, including the blockchain transaction address, will be added in the chained document HTML metadata. The finalized chained document HTML will be returned to a user.

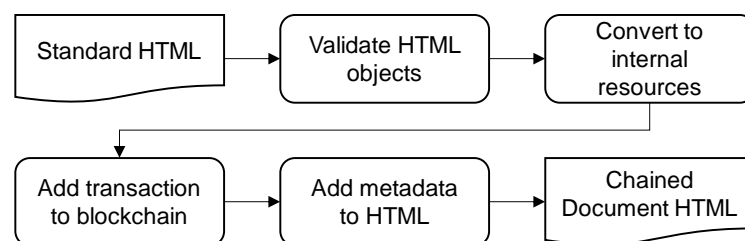


Figure 6. The chained document HTML generation process.

A user who receives the chained document HTML can open and view the document using standard HTML viewer software, as the chained document HTML uses standard HTML specification. If the user wants to verify the content integrity of the chained document HTML, the user can request the document verification. The verification process has four major sub-sequence processes, as shown in Figure 7. The verification process reads the metadata from the chained document HTML and compares the content digest from the HTML body and the content digest from the metadata. This is the first verification process by the content digest value in the file. Furthermore, the transaction data from the blockchain is retrieved by using the transaction address in the metadata to find the public key and content digest. In addition, the signed-content-digest will be verified by using the public key. This part is the second verification process.

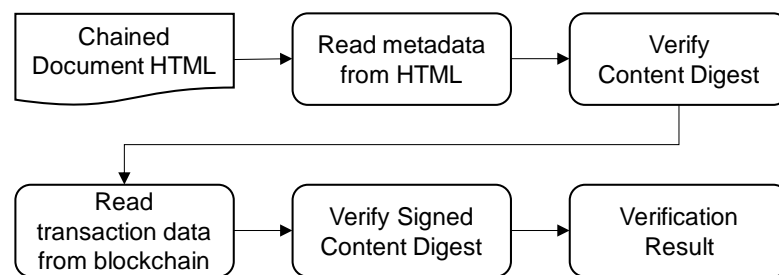


Figure 7. The chained document HTML verification process.

4. Experimental Verification

4.1. Experiment Environment

We used the python script and Ethereum blockchain network in Linux for the experiment, as shown in Table 3. We have created the python script to generate the chained document HTML and verify the chained document HTML. The certificate from the Let's Encrypt [14] is used to sign the content digest by using the OpenSSL. The Ethereum local network is used for the blockchain to save the blockchain metadata of the chained document HTML; the blockchain network was run by the Geth console.

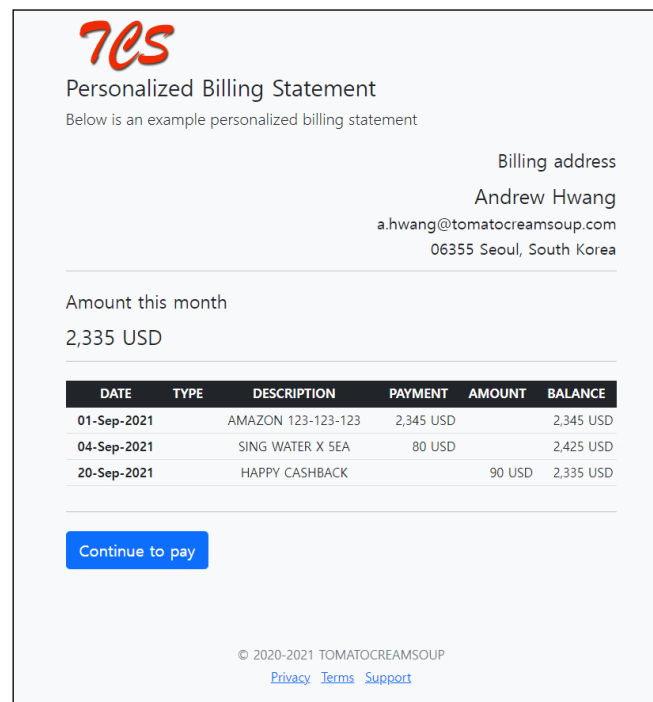
Table 3. Experiment environment.

OS	Ubuntu 20.04.2 LTS 64bit
Geth	1.10.4-stable
OpenSSL	OpenSSL 1.1.1f
Certificate	Let's Encrypt SSL Certificate 2048bit key
Python	Python 3.8.8
Web3	

The Ethereum local network was running in the backend, and the python script could connect to the network by using the Web3 library. Furthermore, the python script could execute the OpenSSL to sign the content digest or validate the signature. So, the python script was the entry point for the experiment; it can read the sample HTML document and generate the chained document HTML. After that, the python script also verified the content integrity of the chained document HTML.

4.2. Sample HTML Document for Experiment

We have researched the most common HTML tags and external resources from financial enterprises. We also prepared the sample HTML document based on the research, as shown in Figure 8. The sample HTML document used HTML tags which were found from the research, and most tags were mainly for the layout display and linked as shown in Table 4. In addition, the sample HTML document had external resources such as JavaScript, as shown in Table 5. The sample HTML document contained a personal customer address area, amount to pay area, detailed transaction list area, and the interaction button area to simulate a billing statement in a financial enterprise.



TCS

Personalized Billing Statement

Below is an example personalized billing statement

Billing address
 Andrew Hwang
 a.hwang@tomatocreamsoup.com
 06355 Seoul, South Korea

Amount this month
 2,335 USD

DATE	TYPE	DESCRIPTION	PAYMENT	AMOUNT	BALANCE
01-Sep-2021		AMAZON 123-123-123	2,345 USD		2,345 USD
04-Sep-2021		SING WATER X SEA	80 USD		2,425 USD
20-Sep-2021		HAPPY CASHBACK		90 USD	2,335 USD

[Continue to pay](#)

© 2020-2021 TOMATOCREAMSOUP
[Privacy](#) [Terms](#) [Support](#)

Figure 8. Sample HTML.

Table 4. The tags in the sample HTML.

Tag	Count	Tag	Count	Tag	Count	Tag	Count
td	15	li	3	form	1	h2	1
th	9	hr	3	table	1	thead	1
div	9	h4	2	head	1	tbody	1
p	6	script	1	style	1	html	1
meta	5	title	1	footer	1		
tr	4	img	1	body	1		
a	4	link	1	ul	1		

Table 5. The external resources in the sample HTML.

Type	Description
Image	PNG logo
Style Sheets	Bootstrap CSS
JavaScript	jQuery JS library

4.3. Generate Chained Document HTML

The chained document HTML was generated from the sample HTML document, as shown in Figure 9. All resources are embedded by data URI format, as shown in Figure 9-(4), so the chained document did not require to open an external resource to display the content. The hash digest value, which was calculated by the content, including internal resources, can be found in the DOCUMENT-HTML-CONTENT-DIGEST metadata, as shown in Figure 9-(1). Furthermore, the digital signed value, which was signed with the DOCUMENT-HTML-CONTENT-DIGEST as an input by a private key, can be found in the DOCUMENT-HTML-CONTENT-SIGNED-DIGEST metadata, as shown in Figure 9-(2). The file size of the sample HTML document was 303,775 bytes, including external resources,

and the file size of the chained document HTML is 405,064 bytes because of adding metadata and converting resources to data URI format.

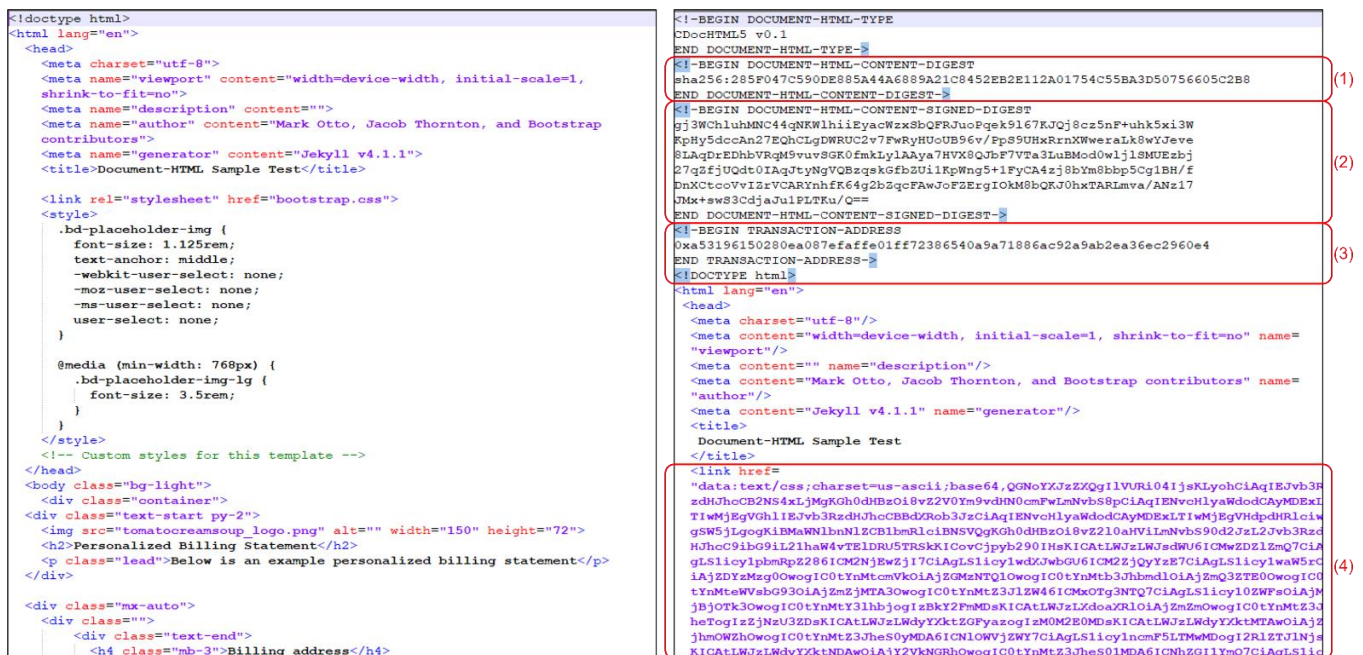


Figure 9. HTML vs. chained document HTML.

The value of TRANSACTION-ADDRESS metadata is indicated to the transaction of Ethereum as shown in Figure 9-(3), and the transaction value in the transaction address can be found as shown in Figure 10. Chained document HTML blockchain metadata is in the input key section as binary, and the binary value contains DOCUMENT-HTML-CONTENT-DIGEST and digital signature certificate as JSON format, as shown in Figure 11.



Figure 10. Chained document HTML metadata in Ethereum.


```
{ 'DOCUMENT-HTML-CONTENT-DIGEST':
  '285F047C590DE885A44A6889A21C8452EB2E112A01754C55BA3D50756605C2B8',
  'CERTIFICATE': '-----BEGIN CERTIFICATE-----
\nMIIFZDCCBEygAwIBAgISAzopABY5pWN1/8+IkGo+g9CnMA0GCSqGSIb3DQEBCwUA\nMDIxClZAJBgNVBAYTA1
VTMRYwFAYDVQQKEw1MZXQncyBFbmNyeXB0MQswCQYDVQQD\nnEwJSMzAeFw0yMTAyMDMwNzZmZdaFw0yMTA1MD
QwNzZmZdaMBwGjAYBgNVBAMT\nnEwLzLnF1YWVpZmV0LmNvLmtyMIIIBIjANBgkqhkiG9w0BAQEFAAOCAQ8AMI
IBC
...
r/ytc5q3F04Q==\n-----END CERTIFICATE-----\n'}
```

Figure 11. Chained document HTML structure in Ethereum.

DOCUMENT-HTML-CONTENT-DIGEST is the identical value to DOCUMENT-HTML-CONTENT-DIGEST in chained document HTML metadata, and CERTIFICATE contains digital signature information and a public key to verify DOCUMENT-HTML-CONTENT-SIGNED-DIGEST.

4.4. Verify Chained Document HTML

We verified the chained document HTML by using the verification python script. The python script can generate the verification report after finishing to verify the chained document HTML integrity. We cloned the chained document HTML from the generation experiment step, and we harmed the file to simulate altering the file in an unauthorized manner. The verification reports for the original chained document HTML and the harmed chained document HTML can be seen, as shown in Figure 12. The verification report from the original one showed all of the content digest and the signed content digest are valid. However, the verification report from the harmed file showed the content was not identical and it was invalid.

<p>Verification Result</p> <p>Verified OK</p> <p>Chained Document HTML Content Digest</p> <p>285F047C590DE885A44A6889A21C8452EB2E112A01754C55BA3D50756605C2B8</p> <p>Message</p> <p>Verified OK</p> <p>Certificate Message</p> <p>Certificate: Data: Version: 3 (0x2) Serial Number: 03:3a:29:00:16:39:a5:63:65:ff:cf:88:90:6a:3e:83:d0:a7 Signature Algorithm: sha256WithRSAEncryption Issuer: C = US, O = Let's Encrypt, CN = R3 Validity Not Before: Feb 3 07:33:37 2021 GMT Not After : May 4 07:33:37 2021 GMT Subject: CN = is.quadient.co.kr Subject Public Key Info: Public Key Algorithm: rsaEncryption RSA Public-Key: (2048 bit) Modulus: 00:be:77:14:80:d5:2d:98:10:0a:74:f0:cf:62:e9: 4e:bd:a3:e9:4f:c7:45:34:bf:57:65:4a:b6:39:59: ec:05:32:0c:5c:91:98:d2:9d:4a:15:8b:cf:eb:fa: aa:69:ae:01:28:a7:75:a2:80:08:29:49:a2:21:dc: 1a:b6:c8:15:a2:27:99:58:c9:51:05:cf:c5:9a:42: ea:97:8e:4a:91:4d:cc:dd:64:0f:ba:7e:e6:a4:f3: bd:bf:52:04:71:07:b7:49:af:2e:64:05:d4:c8:b0: 07:6d:cf:a9:68:80:24:1f:3e:59:f2:46:b8:7f:0d: b7:Ba:87:15:64:4c:d1:bf:1b:1a:70:9d:d8:10:a6: c3:a0:83:d0:ac:61:54:1b:46:fe:08:59:df:61:93: 5a:80:0d:b5:7a:5c:0e:bd:52:66:2b:16:0f:d4:4c: f7:01:a1:c3:51:04:51:f7:7a:7a:dd:21:ad:94:4a: 63:c8:05:61:f1:6a:f1:19:ed:57:10:10:32:29:3e: c7:db:b4:e8:55:d3:ee:8e:d3:22:c9:1b:83:52:1f:</p>	<p>Verification Result</p> <p>Verification is failed</p> <p>Chained Document HTML Content Digest</p> <p>Content Digest in Header: 285F047C590DE885A44A6889A21C8452EB2E112A01754C55BA3D50756605C2B8 Current Digest in Document: 2A51D72FD76338DA1EBE18375E175F503766BAFBF27D41AF9286119989FAFA39</p> <p>Message</p> <p>Hash digests are different</p> <p>Certificate Message</p>
--	---

Figure 12. The verification result report.

The content digest value can be easy to be altered in an unauthorized manner. However, the signed content digest, DOCUMENT-HTML-CONTENT-SIGNED-DIGEST, was generated by a private key in a digital signature. It takes around 2100 MIPS years to find a

private key for unauthorized access in case RSA-140bit, and takes 3423 years for the time estimation for factoring RSA-1024 bit [15]. In the experiment, the digital signature with RSA-2048bit length was used, and we can say that it is impossible to alter the content in an unauthorized modification manner as the time estimation for factoring RSA-2048 bit is 4.00633×10^{12} years [15]. In addition, the metadata in the transaction in the blockchain is immutable, so it also shows that computing to find a private key is useless as it cannot alter the metadata in the transaction in the blockchain.

5. Comparison

5.1. PDF vs. HTML vs. Document HTML vs. Chained Document HTML

The chained document HTML is the enhanced version of the document HTML based on HTML for the trust document communication. The chained document HTML has the advantage against a PDF, a standard HTML, and the document HTML, as shown in Table 6. A PDF document does not provide a good user experience on a mobile device as well as a content delivery verification. A user experience is the most important factor in the interaction between an enterprise and a customer, so an HTML document is more widely used as it provides a good user experience. The chained document HTML and the document HTML use related resources as an embedded internal format, whilst the standard HTML allows external and internal resources. In addition, the chained document and the document HTML do not allow to use of a specific HTML tag, such as iframe, to avoid changing the content on the fly. Therefore, the chained document HTML and the document HTML provide strong content integrity with this characteristic. The chained document HTML and the document HTML have the metadata for the content integrity verification. The checksum and the signed checksum in the metadata are used for document verification. In addition, the signed checksum by using a private key protect to modify the content against unauthorized manner. The document HTML needs to embed all metadata, including a public key and a certificate, to verify the content integrity with the signed checksum. It provides the content integrity feature as well as the chained document HTML. However, there is no way to trace the document delivery status between an enterprise and a customer. The chained document HTML contains the minimum metadata, and the metadata for the content integrity and content delivery verification is stored in the blockchain node. As a result, the chained document HTML minimizes overhead files and verifies content integrity and delivery based on the blockchain.

Table 6. Comparison among HTML vs. document HTML vs. chained document HTML.

	PDF	HTML	Document HTML	Chained Document HTML
HTML compatibility	Incompatible	Compatible	Compatible	Compatible
Structure	Single file with internal resources	Multiple files with external resources	Single file with internal resources	Single file with internal resources
Resources format	Any resources on PDF specification	Any resources on HTML specification	Any resources with Data URI scheme	Any resources with Data URI scheme
HTML tags	Nonsupport	Any HTML tags under HTML specification	Don't allow to use iframe, object, video, audio	Don't allow to use iframe, object, video, audio
Data loading by scripting	Allow	Allow	Don't Allow	Don't Allow
Metadata in the document	XMP, Pieceinfo	-	Content digest, Signed digest, Certificate	Content digest, Signed digest, Transaction address

Table 6. Cont.

	PDF	HTML	Document HTML	Chained Document HTML
Metadata in the blockchain	-	-	-	Timestamp Creator Content digest Certificate
User experience on a mobile	Bad readability, No interaction	Good readability, Responsive	Good readability, Responsive	Good readability, Responsive
Verification content integrity	Verification by a digital signature	Nonsupport	Verification by document HTML metadata	Verification by chained document HTML metadata
Verification content delivery	Nonsupport	Nonsupport	Nonsupport	Verification by chained document HTML blockchain metadata

5.2. File Size Overhead

There is a file size overhead in the document HTML and the chained document HTML because these need to include all resources and necessary information. However, the chained document HTML contains the minimum metadata, and the metadata for document verification is stored in the blockchain node. As a result, the chained document HTML minimizes overhead files and verifies content integrity and delivery based on the blockchain. The file size overhead from the chained document HTML and the document HTML shows different overhead sizes as shown in Figure 13. The file size overhead of the document HTML is the sum of the content digest, signed content digest, content verification key, and each resource's size, increasing by data URI conversion. The file size overhead of the chained document HTML is the sum of the content digest, signed content digest, transaction address, and each resource's size increases by data URI conversion. The size of the transaction hash address is about 66 bytes as string format with Ethereum, whilst the size [16] of the content verification key has no limitation of the file size. The chained document HTML can reduce more than 2000 bytes of overhead file size as the content digest certificate is more than 2000 bytes in general.

$$S_d = M_d + M_s + M_v + \sum_{i=0}^n f(R_i)$$

$$S_c = M_d + M_s + M_t + \sum_{i=0}^n f(R_i)$$

S_d : Document HTML overhead file size	S_c : Chained document HTML overhead file size
M_d : Content digest metadata	M_s : Signed content digest metadata
M_v : Content verification key metadata	M_t : Transaction address
R : Resource	$f(R)$: Data URI conversion

Figure 13. The file size overhead in the document HTML and the chained document HTML.

6. Conclusions

In this paper, we researched and designed the technique of chained document HTML, which provides strong content integrity by using blockchain technology. The chained document HTML treats a traditional HTML document as a single document like a PDF document. Hence, a user can store and review the chained document HTML without breaking a document due to external resources regardless of network connection status. Moreover,

the chained document HTML validates the content integrity by using a digital signature and the metadata for the content integrity is treated as the extended header information. The extended header information is stored in the chained document Blockchain, so it is impossible to modify once the information is recorded. As a result, a user who receives a personal electronic document based on the chained document HTML can ensure content integrity and expect an enhanced user experience in a mobile environment. The electronic document based on the chained document HTML could be a new de facto electronic document standard format on the mobile environment beyond the electronic document based on PDF as the chained document HTML has both mobile-friendly user experience and content integrity verification feature. Therefore, we expect the chained document HTML technique can be widely used in an enterprise such as a financial enterprise that needs to deliver sensitive personal information via a mobile channel. In future research, we will research to optimize the data capacity of the metadata strong for the chained document content integrity verification in the chained document block for more efficient blockchain block size processing.

Author Contributions: Conceptualization, H.-C.H. and W.-J.K.; methodology, H.-C.H. and W.-J.K.; software, H.-C.H.; validation, H.-C.H.; writing—original draft preparation, H.-C.H.; writing—review and editing, H.-C.H. and W.-J.K.; supervision, W.-J.K.; funding acquisition, W.-J.K. All authors have read and agreed to the published version of the manuscript.

Funding: Seoul National University of Science and Technology:1.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Acknowledgments: This study was supported by the Research Program funded by SeoulTech (Seoul National University of Science and Technology).

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

References

1. Sprague, R.H., Jr. Electronic document management: Challenges and opportunities for information systems managers. *MIS Q.* **1995**, *19*, 29–49. [CrossRef]
2. Kim, H.C. Issues and Subjects of the Framework Act on Electronic Document and Electronic Commerce. *Law Stud.* **2012**, *15*, 293–322.
3. DocuSign. Available online: <https://www.docusign.com/> (accessed on 20 November 2021).
4. E-Document Integrated Support Center. Available online: <https://www.npost.kr/> (accessed on 20 November 2021).
5. Hwang, H.C.; Kim, W.J. Design of Document-HTML Generation Technique for Authorized Electronic Document Communication. *J. Soc. Korea Ind. Syst. Eng.* **2021**, *44*, 51–59. [CrossRef]
6. Warnock, J.E.; Geschke, C. Founding and Growing Adobe Systems, Inc. *IEEE Ann. Hist. Comput.* **2019**, *41*, 24–34. [CrossRef]
7. Staar, P.W.; Dolfi, M.; Auer, C.; Bekas, C. Corpus conversion service: A machine learning platform to ingest documents at scale. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 774–782.
8. Rohlmann, S.; Mladenov, V.; Mainka, C.; Schwenk, J. Breaking the Specification: PDF Certification. In Proceedings of the 2021 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 24–27 May 2021; pp. 1485–1501.
9. W3C. HTML 5.2. Available online: <https://www.w3.org/TR/html52/> (accessed on 20 November 2021).
10. Kim, S.K.; Huh, J.H. Artificial Neural Network Blockchain Techniques for Healthcare System: Focusing on the Personal Health Records. *Electronics* **2020**, *9*, 763. [CrossRef]
11. Radanliev, P.; De Roure, D.; Burnap, P.; Santos, O. Epistemological equation for analysing uncontrollable states in complex systems: Quantifying cyber risks from the internet of things. *Rev. Socionetw. Strateg.* **2021**, *15*, 381–411. [CrossRef]
12. Abdi, A.I.; Eassa, F.E.; Jambi, K.; Almarhabi, K.; Khemakhem, M.; Basuhail, A.; Yamin, M. Hierarchical Blockchain-Based Multi-Chaincode Access Control for Securing IoT Systems. *Electronics* **2022**, *11*, 711. [CrossRef]
13. Masinter, L. The “data” URL scheme. In RFC 2397; The Internet Society: Reston, VA, USA, 1998.
14. LetsEncrypt. Certificate Authority Providing TLS Certificates. Available online: <https://letsencrypt.org/> (accessed on 20 November 2021).
15. Dasso, A.; Funes, A.; Riesco, D.; Montejano, G. Computing Power, Key Length and Cryptanalysis. An Unending Battle? *arXiv* **2020**, arXiv:2011.00985.
16. Etherscan. Available online: <https://etherscan.io> (accessed on 20 November 2021).