ARCHANGEL: Trusted Archives of Digital Public Documents

J. Collomosse, T. Bui, A. Brown
University of Surrey
Guildford, Surrey, UK
j.collomosse@surrey.ac.uk

J. Sheridan, A. Green, M. Bell
The National Archives
Kew, London, UK
jsherdian@nationalarchives.gov.uk

J. Fawcett, J. Higgins
O. Thereaux
Open Data Institute (ODI), UK
jamie.fawcett@theodi.org

ABSTRACT

We present ARCHANGEL; a de-centralised platform for ensuring the long-term integrity of digital documents stored within public archives. Document integrity is fundamental to public trust in archives. Yet currently that trust is built upon institutional reputation — trust at face value in a centralised authority, like a national government archive or University. ARCHANGEL proposes a shift to a technological underscoring of that trust, using distributed ledger technology (DLT) to cryptographically guarantee the provenance, immutability and so the integrity of archived documents. We describe the ARCHANGEL architecture, and report on a prototype of that architecture build over the Ethereum infrastructure. We report early evaluation and feedback of ARCHANGEL from stakeholders in the research data archives space.

CCS CONCEPTS

• **Information systems** → **Digital libraries and archives**; • **Networks** → *Peer-to-peer protocols*;

KEYWORDS

Distributed Ledger Technology (DLT), Blockchain, Trusted Archives, Document Provenance, Content Integrity and Verification.

ACM Reference Format:
J. Collomosse, T. Bui, A. Brown, J. Sheridan, A. Green, M. Bell, and J. Fawcett, J. Higgins O. Thereaux . 2010. ARCHANGEL: Trusted Archives of Digital Public Documents. In *Proceedings of ACM Document Engineering (ACM DocEng'18).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Archives and Memory Institutions (AMIs) are the lens through which future generations will perceive today. AMIs are founded upon the principles of public trust — of being neutral and completely trustworthy. The immutability and integrity of the documents they hold are essential to maintaining their objectivity; be they government documents in National Archives, or research documents held by University archives. Yet, today's digital age presents urgent, new challenges to this trust and immutability. Digital documents are ephemeral, and produced in great volume. Their intangibility leaves them open to modification — not only to tampering but also to degradation over longitudinal time periods (e. g. decades). For example, file formats become obsolete and documents are transcoded. How can we ensure that meaning is not lost?

In this paper we describe ARCHANGEL; a de-centralised architecture for ensuring the integrity of digital documents within public archives. ARCHANGEL utilises blockchains — a form of secure decentralised ledger technology (DLT) — as a basis for ensuring the provenance and integrity of documents. Blockchains store chronologically ordered transactional data, permanently preserving that data through peer-to-peer distribution and consensus checking without the need for a trusted third party. Although best known for underpinning digital exchanges of cryptocurrencies e. g. BitCoin [11], their transaction based model has also been applied to transmit data payloads for internet domain name management (Namecoin) [8] and interpersonal messaging (Bitmessage) [2]. Recently there has been significant interest in applications of DLT to public services by government [17] including to record-keeping [9]. ARCHANGEL breaks new ground by proposing the use of a blockchain payload to record digital signatures (*content evidence*) derived from either scanned physical, or born-digital, document to ensure their integrity over decade- or century-long timespans.

We have implemented a prototype of ARCHANGEL based on the Ethereum DLT infrastructure and exposed this to archivist stakeholders in the University research data management space for feedback. We document this prototype in Sec. 3, and reflect upon this early evaluation in Sec. 4. We discuss future directions for development of the ARCHANGEL platform in Sec. 5.

2 RELATED WORK

AMIs are struggling to keep pace with the exponential rate of digital transformation in society [16, 17]. Today, the majority of content is born-digital and there is inexorable end-user demand for digitisation of existing content e.g. for open access and operational efficiencies [7]. Recent work addressing this trend focuses on developing theories of record keeping, such as the Records Continuum Model [10], or standards and technologies for describing, cataloguing or searching archives such as Discovery [3], the Archives Portal Europe project [4] or the Records in Context initiative [14]. Some work has been done looking at analytics of archival data enabled by big data approaches such as Traces Through Time [1]. Rather than exploring novel ways to index and annotate documents with metadata, ARCHANGEL explores the orthogonal challenge of ensuring the long-term integrity and sustainability of digital archival content, proposing a platform for verifying the integrity and provenance of digital documents whilst entrusted to the archive (curation) and upon document release (presentation).

AMIs now exist within an age where people are increasingly questioning institutions and their legitimacy. Historically, an archives' word was authoritative, in effect vouching for the integrity of documents through their reputation. ARCHANGEL advocates the use

ACM DocEng'18, August 2018, Halifax, NS, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

of a de-centralised model (DLT) rather than — say — a secondary database of document signatures or certificate key authorities maintained centrally, to evidence the integrity and provenance of those documents [14]. In doing so, ARCHANGEL enables a shift from an institutional underscoring of trust, to a technological underscoring of trust. Some recent work has explored distributed assurance of provenance through embedding document signatures within versioning information e. g. in Microsoft Word documents [5, 15]. ARCHANGEL goes further, creating a de-centralised repository of content evidence independent of the document itself that is collectively maintained across multiple participating archives through consensus checking on the Blockchain.

3 TRUSTED DIGITAL DOCUMENTS

ARCHANGEL utilises a permissioned blockchain model, in which operators or automatic processes authorised to add content to the AMI commit blocks into the chain encoding content evidence (and if that evidence is derived in a non-standard way, a hash of associated code for deriving that evidence). The security of a blockchain is afforded by the immutability of data with the blocks, delivered by the compounding effect of each new block being hashed to include the hashes of previous blocks. Thus as content is committed into the blockchain, the security of the content is reinforced. The Blockchain remains publicly readable to enable open verification of documents released from the archive at any time.

ARCHANGEL proposes a cross-AMI model in which a single DLT is contributed to by multiple AMIs, potentially across different disciplines and nations, mitigating any risk of distortion of an archive by its operating AMI. In this way multiple archives potentially across international borders, can mutually reinforce the integrity of each others' documents without necessarily being party to the content of those documents until release.

3.1 Illustrative Scenarios

Documentary evidence gathered for public inquiries (e. g. into the UK 7/7 terrorist attacks, or the Chilcot Inquiry) are kept by the UK National Archives and can remain closed for the best-part of a century. ARCHANGEL enables such archives to lodge content evidence within the Blockchain at time of document deposition, enabling the public to verify the integrity and provenance of those documents upon release. The verification is open and may be performed by anyone through reapplying the hashing algorithms used to extract content evidence at deposition, and comparing those hashes to those lodged within the Blockchain also at the time of deposition. If a non-standard hashing algorithm (for example a machine learning based content parser) was used, then the hash of that algorithm code may also be included in the Blockchain.

Similarly, consider a University publishing a study on climate change and including a DOI to supporting data released openly. That data could be hashed into content evidence, and lodged in the ARCHANGEL DLT alongside a hash of the code necessary to recreate that hash at time of publication. A decade passes, and years later the research integrity of the paper is called into question. The University can evidence that the research data it provides matches that at time of deposition.

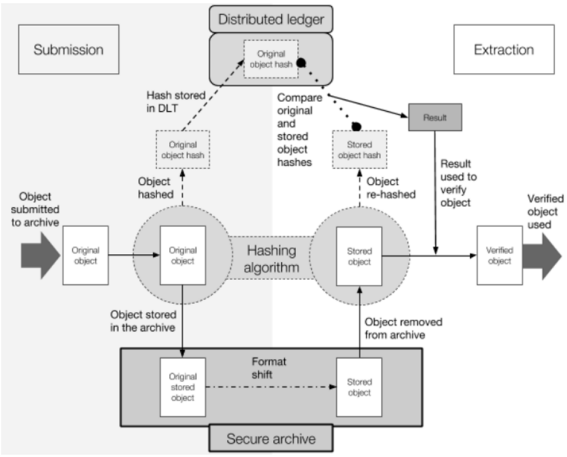

Figure 1: Architecture of the proposed ARCHANGEL platform. Documents are processed to extract *content evidence* which is stored immutably within a Blockchain alongside metadata identifying both that content, and the algorithm used to extract the evidence. A document's integrity and provenance can be checked at any time by re-extracting and comparing the content evidence to that in the Blockchain.

3.2 ARCHANGEL Architecture

Figure 1 illustrates the ARCHANGEL architecture. ARCHANGEL does not propose a distributed filesystem or similar for the storage of documents (the AMI is assumed to provide this solution) rather we propose the decentralised storage of compact hashes derived from documents on a Blockchain alongside metadata to assist in future identification and verification of those documents.

Upon deposition of a document, a file format identification tool determines the content type of the document (e. g. PDF, Microsoft Word) by performing classification upon the binary information within the file irrespective of its accompanying metadata e. g. filename. Content evidence is then extracted from the document in a format-dependent manner via a content hashing algorithm. In the simplest form this content hashing might be a classical binary hashing algorithm (e. g. SHA-256) applicable to all formats, however we consider that bespoke content hashing processes might be applicable to specific formats. For example, a digital image of a scanned physical document might employ a deep neural network (DNN) to extract robust visual features from visual content that are invariant to appearance properties (e. g. illumination, ageing) of that document. Having extracted the content evidence, a filename or other global unique identifier (GUID), the content hash, and an unique identifier signifying the content hashing process are stored alongside supplemental metadata with the Blockchain. This supplemental metadata might include archivist's notes, date of deposition, versioning information, and if the content hashing process is bespoke potentially a hash (using a binary standard hash such as SHA-256) of the code or equivalently DNN model used to extract the features. In the latter case, the code or model would also be secured within the archive.

To store this block of new data, it is appended to the end of a linked list of blocks in a distributed data structure (the 'Blockchain').

Multiple nodes within the DLT infrastructure replicate this operation and consensus check the result per the protocol of the underlying DLT infrastructure. We have chosen to implement ARCHANGEL in Ethereum DLT infrastructure (c.f. Sec. 3.3), which is currently a proof-of-work based Blockchain. In practice this means that multiple nodes in the network must solve a cryptographic puzzle in order to append to the list while other nodes must be in consensus as to the validity of the operation [11]. In our architecture we propose two modes of consensus checking, both predicated upon a permissioned DLT model:

(1) The Blockchain is maintained via proof of work across a private set of nodes, which are maintained collectively by multiple AMIs each with independent governance structure e. g. national archives of different nation states. As such an unprecedented level of collusion would be required to corrupt the Blockchain.
(2) The Blockchain is maintained via proof of work across a public Blockchain maintained globally. In such case a program embedded within the Blockchain (a 'smart contract') with sole permission to write to the Blockchain is invoked in order to the append data. Access to the smart contract end-point is granted via secret key. In this case corruption would require more than half of the public DLT infrastructure miners to collude, which is again unlikely e. g. on the Ethereum main network.

To verify the provenance and integrity of a document curated or released by an AMI, the content evidence must be computed from that copy of the document and compared against that immutably stored within the Blockchain. The public availability of the contents of the Blockchain enable anyone to search and identify the appropriate data block using the unique identifying information (GUID) accompanying the content hash, stored during deposition. Should the content evidence be hashed using a bespoke technique, the instance of that technique (code or network model) must too be requested from the AMI and compared to the hash computed at deposition. Assuming both content and algorithm hashes match, the document is considered to be authentic.

3.3 Prototype Implementation

We have implemented a prototype of ARCHANGEL on the Ethereum public test net (Rinkeby), adopting consensus model (2) described above. Ethereum was selected due to its global adoption and prominence as a DLT platform, and due to its technical capability to store both data and execute smart contracts (via EVM); which we implemented in the Solidity language.

We used the DROID (Digital Record Object IDentification) application developed by the UK National Archives to classify document type [6]. Without loss of generality we assume that this process is executed via a cloud-based service capable of accepting uploads and running them both through DROID and through content extraction. Currently our prototype uses a standardised binary hash (SHA-256) but ongoing work is exploring format-specific hashing using bespoke machine learning models for document feature extraction.

Figure 2 presents a screen-shot of our implementation which contains the functionality to 1) to deposit documents; 2) to search for documents e. g. based on GUID or content hash string; and 3) to verify documents (effectively running operations 1) and 2) in succession to identify prior instances of the content hash).

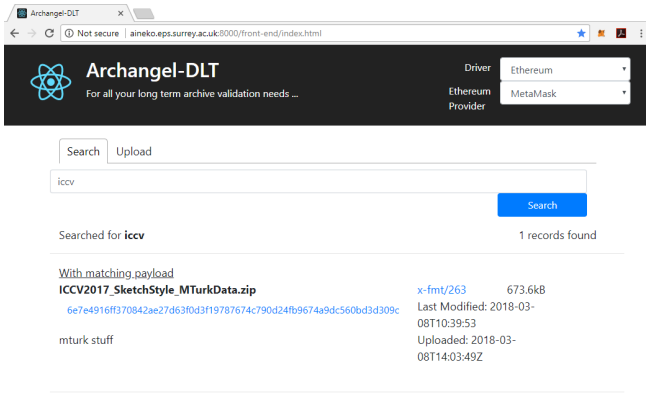

Figure 2: Screenshot of the ARCHANGEL platform running on Ethereum (search functionality shown). User may search on document global unique identifier (GUID), metadata, or content evidence (hash) to verify its integrity.

4 STAKEHOLDER WORKSHOP

A workshop was convened with 13 expert participants drawn from AMIs in the government, commercial legal and university research document management spaces. Participants were briefed on the ARCHANGEL platform and provided the opportunity to interact with the platform prototype for one hour in a lab-based setting given a set of research documents provided by the University of Surrey. Participants explored the functionality of the platform to deposit, search and verify documents not as a usability exercise but as a provocation for a facilitated semi-structured group discussion to capture feedback on the value of ARCHANGEL to digital document preservation in AMIs. The dominant theme was the perceived value in enabling archives to engender trust both in their records and in their practices; that as a result of providing proof that the records have not changed, their authenticity can be demonstrated. Several discussion themes emerged.

1. Defenders of the record. In an era when the technologies to fake digital content are becoming increasingly pervasive, it is not surprising that the public has less trust in all things digital. At the same time there is a lowering of trust in public institutions; a result of the frequent reports of untrustworthy behaviour of politicians and other officials, the ease with which misinformation is spread via the internet and the difficulty of identifying the source and verifying its accuracy [13]. Together these create a perfect storm for the digital archive. Blockchain offers a shield which archives can use to defend the records as authentic. By enabling researchers to compare the content evidence (including the checksum) of the record to that recorded on the blockchain, they can see proof that no changes (deliberate or accidental) have been made to the record since it was preserved in the archive. Further to this, the decentralised nature of the blockchain removes the need for citizens to trust individual institutions as each is the guardian for the other guardians. It was also noted that acceptance of content evidence might eventually become similar to acceptance of DNA evidence in court, but that establishing that level of confidence would require strong public engaged to explain Blockchain in an accessible manner particularly explaining why one could trust the cryptographic assurances inherent in a DLT solution. It was noted that although this could be considered a form of digital forensics there was no

specific precedent or case law (within the UK) in relation to verification of digital documents in AMIs. There was strong potential for further socio-technical research around the consequences of ARCHANGEL as a technical platform not only in relation to law but also in the potential to evolve archival practice itself.

2. Engagement with emerging technologies. Archives are not generally viewed as digital institutions nor archivists as a digital profession. Yet, there are many areas where archives are actively engaged with digital technology, from the preservation of born-digital records to researching the uses of machine learning for appraisal, selection and access. AMI involvement in the practical application of Blockchain demonstrates an interest in and openness to new technologies. The attendees were very keen to be kept informed of the project's progress congiscent of the tidal wave of digital documents arriving in their archives at increasing rate.

3. Demonstrates our willingness to be transparent in our practices. While blockchain encourages citizens to trust the records in our custody as being the 'ground truth' it can also make our practices transparent in a way that can be verified by researchers. Archivists would be able (in an automated environment) to use the technology throughout their processing of the records. They could record the content evidence on the blockchain after each significant curatorial action, creating an audit trail of those actions and a series of hashes to verify. This further encourages trust in archives' role as custodians but also in use of best practice.

4. Demonstrates the collaborative nature of the digital archives community. A major benefit for the archival community of using blockchain technology is that it requires collaboration. The ability and willingness of archives to engage with each other as well as other heritage organisations both nationally and internationally is key to its success. Archives have a good track record of supporting each other. The value of a distributed approach to assuring trust in digital records may be most keenly felt by archives at risk. Blockchain provides a way for archives to underscore trust in each others' collections introducing an entirely new form of collective defence of the archive. For example ARCHANGEL raises the potential for AMIs to collaborate to provide the computational power and the consensus checking of the platform — potentially across international or jurisdictional boundaries. This gives the technology the huge advantage of not only preventing individual governments from tampering with the public record but also guarantees a degree of longevity due to the legislative position of public archives. The technology itself engenders trust by guaranteeing that the records have not been tampered with, and it also allows archives to make their processes transparent both of which encourage trust in their integrity as custodians of the public record.

5 CONCLUSIONS AND NEXT STEPS

We have described ARCHANGEL — a platform for verifying the provenance and integrity of digital documents within public archives. Uniquely, ARCHANGEL combines content hashing with distributed ledger technology (DLT) to create a de-centralised trust model. We have proposed an architecture and reported both its implementation on the Ethereum infrastructure, and feedback from an early trial at a focus group of archivists drawn from a research data management background. Although fully functional, ARCHANGEL is a prototype and several promising directions exist for future extension. Currently our content hashing is performed using standard binary hashing (e. g. SHA-256) and we would like to specialise hashing to particular document types such as PDF or even images and video. The latter presents the unique possibility of extracting content-aware hashes of scanned documents that are sensitive to tampering or degradation but invariant to factors such as illumination or imaging device. We are also considering the integration of the W3C proposed PROV standard [12] for document versioning given that blocks may only be added to supercede older content (and not deleted) from a Blockchain. Currently our implementation uses smart contracts as a gateway for writing to the Blockchain, but not for search or verification. We might also explore the use of smart contracts for the latter, exploring new business models to encourage sustainability. For example, the maintenance of the DLT (in terms of computational effort for mining) might be facilitated by users who seek document verification 'paying' for that service via contribution of mining effort to maintain the DLT. Our initial feedback from stakeholders is based upon a workshop environment and not a platform deployed in practice. We believe that ARCHANGEL has the position to disrupt the professional practice of archivists, as a technology tool that enables robust curation of digital documents that can easily suffer damage e. g. through format shifting. We feel ARCHANGEL shows significant promise as a means for ensuring the future sustainability of archives as they undergo the challenges of digital transformation.

ACKNOWLEDGEMENT

ARCHANGEL is funded by EPSRC Grant Ref: EP/P03151X/1 under the UKRI Digital Economy Programme.

REFERENCES

[1] M. Bell. 2015. Traces through time: a case-study of applying statistical methods to refine algorithms for linking biographical data. In *Proc. CEUR Workshop*, Vol. 1399.
[2] V. Buterin. 2012. BitMessage: A Model For A New Web 2.0? *Bitcoin Magazine* (2012).
[3] J. Cates. 2014. Developing a National Archives' Catalogue. In *Proc. Intl. Conf. on Digital and Traditional Manuscripts.*
[4] Archives Portal Europe. [n. d.]. http://www.archivesportaleuropefoundation.eu. In *Accessed: 2018-03-27.*
[5] A. Filho, E. Munson, and C. Thao. 2017. Improving Version-Aware Word Documents. In *Proc. Intl. Conf. on Document Engineering (ACM DocEng).*
[6] DROID (Digital Record Object IDentification) User Guide. [n. d.]. http://www.nationalarchives.gov.uk/documents/information-management/droid-user-guide.pdf. ([n. d.]).
[7] V. Johnson and D. Thomas. 2013. Interfaces with the Past...Present and Future? Scale and Scope: The Implications of Size and Structure for the Digital Archive of Tomorrow. In *Proc. Digital Heritage Conf.*
[8] H. Kalodner. 2005. An empirical study of Namecoin and lessons for decentralized namespace design. In *Proc. WEIS.*
[9] V. L. Lemieux. 2016. *Blockchain Technology for Recordkeeping: Help or Hype?* Technical Report. U. British Columbia.
[10] S. McKemmish. 2010. *Records Continuum Model.* Ency. of Library and Info. Sci. 4447–4448 pages.
[11] A. Narayanan, J. Bonneau, E. Felten, A. Miller, and S. Goldfeder. 2016. *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction.* Princeton Univ. Press.
[12] W3C: An Overview of the PROV Family of Documents. [n. d.]. ([n. d.]).
[13] O. O'Neill. 2002. *A Question of Trust âĂŞ The BBC Reith Lectures 2002.* Cambridge Univ. Press.
[14] D. Pitti, B. Stockting, and F. Clavaud. 2016. Records in Contexts - Conceptual Model. In *Proc. Intl. Council on Archives.*
[15] A. Shatnawi, E. Munson, and C. Thao. 2017. Maintaining Integrity and Non-Repudiation in Secure Offline Documents. In *Proc. Intl. Conf. on Document Engineering (ACM DocEng).*
[16] J. Sheridan. 2014. *The Digital Landscape in Government.* Technical Report. The National Archives.
[17] M. Walport. 2015. *Distributed Ledger: Beyond Blockchain.* Technical Report. UK Government.