# ARCHANGEL: guaranteeing the integrity of digital archives

Open Data Institute

# Contents

# About

This report has been created in collaboration between the University of Surrey's Centre for Vision, Speech and Signal Processing (Surrey), the UK's National Archives (TNA), and the Open Data Institute (ODI). It was published in August 2019. It uses, summarises and refers to material created and published between 2017 and 2019 as part of the ARCHANGEL project[1].

Lead editor was Olivier Thereaux (Open Data Institute). Authors were Alex Green (TNA), Arindra Das (ODI), Daniel Cooper (Surrey), Jamie Fawcett (ODI), Jared Keller (ODI), Jez Higgins (ODI), John Collomosse (Surrey), John Sheridan (TNA), Mark Bell (TNA), Olivier Thereaux (ODI), and Tu Bui (Surrey). Other contributions and review by Caley Dewhurst, Becky Ghani, Anna Scott, Jeni Tennison, Rachel Wilson (ODI).

To share feedback or to get in touch, contact the lead editor Olivier Thereaux at ot@theodi.org.

The ARCHANGEL project was publicly funded by the UKRI/EPSRC Digital Economy Programme under grant reference EP/P03151X/1.

[1] Surrey Blockchain (2019), 'ARCHANGEL - Trusted Archives of Digital Public Records', https://blockchain.surrey.ac.uk/projects/archangel.html

# Executive summary

> *Exploring blockchain technology together with some of the world's leading archives, the ARCHANGEL project has shown, for real, how archives might combine forces to protect and assure vital digital evidence for the future.*
> *-- John Sheridan, Digital Director, The National Archives*

Records have been preserved for thousands of years, and the preservation of archives is a mature discipline. In the United Kingdom (UK), the National Archives holds over 120 miles of papers and documents, from the Domesday book to recent UK government cabinet meeting minutes[2].

Over the past two decades, society has experienced rapid technological change, which has resulted in vast quantities of information being captured and stored on media other than paper. Although practices around digital preservation have developed over the past 25 years, many of them attempt to replicate archival practices designed for paper collections.

One of the unique challenges of this shift to digital preservation is that of guaranteeing integrity of the digital records. Digital records – that are transient, easy to copy and modify, and prone to corruption in copy and storage – often need to be ported from one format to another, as technology evolves and software used to read certain formats stops being available. In a context of increasing mistrust in institutions and attacks on the notion of truth, this makes for an explosive, existential challenge for archives and memory institutions (AMIs).

The ARCHANGEL project has been exploring the possibilities offered by distributed ledger technology (commonly known as blockchain) and machine learning and how they could address the challenges around trust, integrity and authenticity that preserving born-digital material presents.

Over two years – between July 2017 and July 2019 – a team formed with members from the University of Surrey (Centre for Vision, Speech and Signal Processing), the UK's National Archives, and the Open Data Institute (ODI). The aim of this team was to collaborate in exploring, developing and prototyping these technologies to underpin trust in digital archives, and the work culminated in the pilot of a system across five countries.

The technology developed through the ARCHANGEL project, and the results of the pilot study, were incredibly promising. It showed how the appropriate use of emerging technology could change digital archiving methodologies and create new collaborations between institutions.

Beyond archives, the results of the ARCHANGEL project have potential to inform and support other domains where truth and integrity of information over time – such as journalism – are crucial to their long-term sustainability.

[2] The National Archives (2014), 'Secrets Of The National Archives', http://bookshop.nationalarchives.gov.uk/9780091943356/Secrets-Of-The-National-Archives/

# The changing role of archives and memory institutions

> *It is becoming easier and easier to manipulate digital records, which makes it crucially important for the institutions who take care of those records to be able to demonstrate their trustworthiness.*
> *-- Jeni Tennison, CEO, Open Data Institute*

## The challenges of digital preservation

To fulfil their role, archives and memory institutions (AMIs) need to be both trustworthy and trusted. They need to do everything they can to prevent the corruption of historical records and be seen to be doing everything they can to achieve this.

One of the important challenges AMIs face is the shift from primarily physical objects to primarily digital objects.

Organisational practices, in government and beyond, are increasingly shifting from physical to digital – from paper memos to emails, printed reports to PDFs, overhead projector transparencies to digital presentation slides.

One specific focus of digital-preservation researchers is guaranteeing the integrity of these born-digital objects – ie that they remain unaltered while stored in the archive. While changing physical objects without obvious evidence of tampering is difficult, digital objects by their nature are relatively easy to change. This raises an important question: how can AMIs guarantee that a stored document is the same document that was originally archived?

In theory, public scrutiny can help: members of the public can compare information in one record against other sources, or in an earlier copy. However, this is made particularly difficult for sensitive closed records that may not be released for decades – records that are nonetheless vital to historical scrutiny. For example, records that could compromise government operations if published contemporaneously are securely stored until a predetermined amount of time passes, at which point they become available to the wider public. In the UK, this is known as the '20-year rule'. [3]

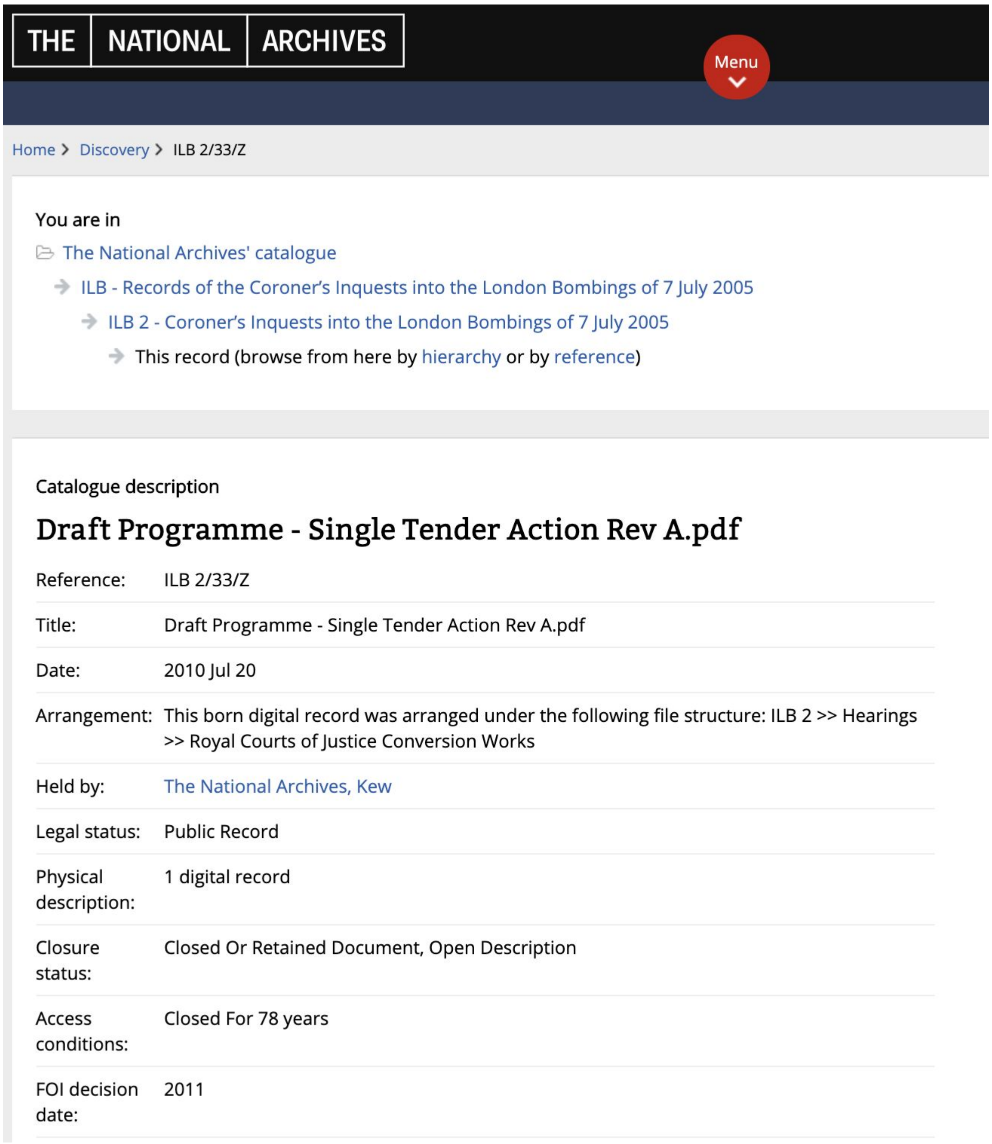
## Guaranteeing the integrity of closed records

While objects are being stored, only archivists with appropriate permission can access them to ensure they are properly preserved.

While the practice of keeping records closed for decades is not new, in the past it would have been relatively straightforward to know whether a paper-based record, once opened, had been redacted or doctored. Much less so for digital

[3] The National Archives (2015), 'The 20-year rule', 'http://www.nationalarchives.gov.uk/about/our-role/transparency/20-year-rule/

records: deleting a problematic paragraph can be done in a keystroke; and forensically examining electronic documents to find evidence of tampering is still very much a complex art.



You are in

The National Archives' catalogue

ILB - Records of the Coroner's Inquests into the London Bombings of 7 July 2005

ILB 2 - Coroner's Inquests into the London Bombings of 7 July 2005

This record (browse from here by hierarchy or by reference)

Catalogue description

Draft Programme - Single Tender Action Rev A.pdf

| | |
|---|---|
| Reference: | ILB 2/33/Z |
| Title: | Draft Programme - Single Tender Action Rev A.pdf |
| Date: | 2010 Jul 20 |
| Arrangement: | This born digital record was arranged under the following file structure: ILB 2 >> Hearings >> Royal Courts of Justice Conversion Works |
| Held by: | The National Archives, Kew |
| Legal status: | Public Record |
| Physical description: | 1 digital record |
| Closure status: | Closed Or Retained Document, Open Description |
| Access conditions: | Closed For 78 years |
| FOI decision date: | 2011 |

*Example of a closed record description in the catalogue of The National Archives*

## The additional challenge of format shifting

The repositories at The National Archives contain around 200km of shelving that holds millions of original paper, parchment and photographic documents. These documents are kept in very precise atmospheric conditions, with tightly controlled temperature and humidity.

Similarly, digital documents are held in conditions conducive to long-term preservation, minimising the degradation of tapes and spinning disks. Digital documents differ from paper in that they are not really the original: they have, at some point in their lifetime, been copied from one medium to another.

Further backup copies are made by the archive to de-risk the preservation process. In medieval times, scribes would make copies of documents and only carefully comparing the copy and the original could verify a faithful copy. In the digital world, we use methods originating in cryptography to automatically verify that not a single bit is out of place in a copied file.

Digital files are made up of electronic 'bits' which encode their contents together with formatting instructions for the software that will be used to display – or render – them. We can use mathematical techniques (as used in cryptography) to reduce these millions of bits down to a unique, short, alphanumeric string of characters for every file. If even one bit of the file were to change, the unique string would change. This unique string is sometimes called a cryptographic hash, a fingerprint, or a checksum.

When a file is received by the archive they generate the file's hash and store it in a database. Regular recomputation of file hashes are made and compared with the original hash to proactively identify corrupted files. If corruption is detected by this process, the file can be replaced with an uncorrupted backup copy.

When archives present a user with a digital file, there are two options available: they could download an exact copy of the original file, or download the file in an alternative format. One reason for the second option being presented is because as time goes by software becomes obsolete. It is replaced with updated versions or entirely new software, and the file formats change with them.

For example, WordStar was a very popular word processor in the 1980s but there is no longer a version that runs on modern computers,[4] although emulators, created by enthusiasts, are available. A WordStar file may be opened in a modern version of Microsoft Word after first installing a conversion add-in. This keeps the format alive and usable for now, but can we guarantee that these files will load on a standard computer in 20 or 50 years? Even if it can read a file, a modern word processor is not necessarily faithfully rendering the original.

In the interests of long-term preservation, and for the convenience of users, the archive may create copies of these WordStar files and then convert them to an open format which is more likely to still be readable decades from now. Similar actions may also be taken for formats such as high-definition video, converting them to a compressed format to reduce the download time, again for user convenience.

Changing formats in this way introduces a problem: mathematically, a converted file is different to the original – it contains different formatting bits even if the contents are unchanged – and so the system of comparing checksums breaks down and identifies these as non-identical files.

Software providers have used checksums for years to allow customers to verify that they are downloading a genuine copy, and the AMI can use them to verify born-digital files in the same way. By changing the format, however, we are offering a cryptographically different file to the one which was originally deposited.

How can we assure the integrity of the file during the conversion process, especially in cases where there may no longer be software available to render the original? Is there a way of demonstrating that two files in different formats are still the same without comparing them side by side? Or, when applied to video material, without resorting to a painstaking task of comparing two long videos frame by frame?

## AMIs in a post-truth culture

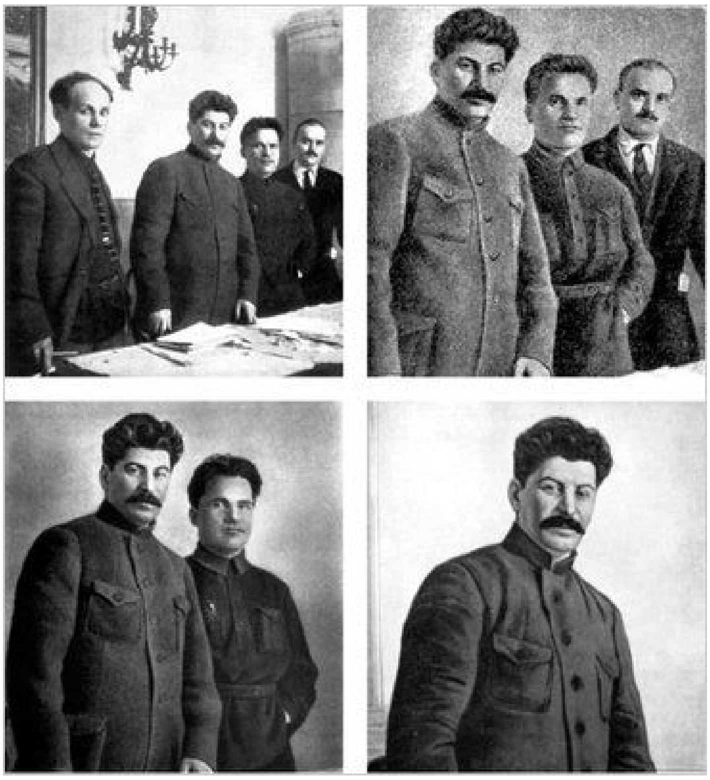
AMIs are continually developing ways of preventing errors, data corruption and other issues. The digital preservation domain had adopted and built on industry best practice in this area, and is offering increasingly effective solutions to guard against error, data corruption and other effects of degeneration.

But the existential challenge to AMIs is not only technological – it is also societal,

[4] Jenny Mitcham, *What are the significant properties of a WordStar file?* https://digital-archiving.blogspot.com/2018/08/what-are-significant-properties-of_85.html

and cultural. Many of the technological solutions to detect and prevent errors or decay do nothing to prevent the deliberate modification of records. And while there is no army of hackers intent on covertly corrupting the nation's record,[5] the attacks on archives are rather more indirect.

Records could be changed from within, for example by order of new governments keen to rewrite history in their favour – as the 20th century has shown time and time again.



*Nikolai Antipow, Sergej Kirow and Nikolai Schwernik edited, over time, from the record for propaganda purposes. From* https://commons.wikimedia.org/wiki/Category:Altered_Soviet_photographs

At the time of writing, the public discourse frequently undermines societal trust in public institutions. Heads of state label media organisations as 'fake news', and raise doubts about the competence or neutrality of intelligence and law enforcement services on matters of national security; politicians label judges as 'enemies of the people'; a UK minister states in his resignation letter that he believes the civil service is providing misleading briefings.[6] We also live in a time when easily-created synthetic content is capable of generating believable videos, putting words into politicians' mouths.

In this atmosphere, AMIs do not need to be directly attacked. They are damaged simply by the miasma of institutional mistrust. Indeed, AMIs are, perhaps, particularly at risk. Their age and importance puts them at the heart of 'the establishment'. Many of the records they hold are not immediately available for public view. By policy or by law, many documents are redacted or even entirely withheld from the public for a period of time.

Those wishing to challenge the integrity of released records could plausibly suggest records are incomplete or altered in some way.. As the documents in question are, in this modern age, entirely digital – having never existed as physical objects – how could an archive demonstrate otherwise?

[5] Or rather, there probably is a motley array of hackers, but major archives are well practiced in both network and physical security. You can't tamper with a file that's  not network connected unless you're willing to engage in Mission Impossible-style shenanigans while deep underground in an otherwise disused salt mine.

[6] "Unfortunately, I do not believe the briefings you have received on these matters recently have reflected all they have achieved or the preparations our European partners have made" - Chris Heaton-Harris resignation letter, April 2019

# Using blockchains and machine learning to underpin trust in archives

> *By combining blockchain and artificial intelligence technologies, we have shown that it is possible to safeguard the integrity of archival data in the digital age.*
>
> *It essentially provides a digital fingerprint for archives, making it possible to verify their authenticity.*
> *-- Prof. John Collomosse, University of Surrey*

## A proof-of-authority distributed ledger for immutable storage

The ARCHANGEL system, created through this project, uses distributed ledger technology to guarantee that document fingerprints cannot be altered, and machine learning to create fingerprints that can withstand format shifts.

### Blockchains, or distributed ledger technology

Often considered synonymous with Bitcoin, blockchain is the technology that underpins a number of digital currencies but it has the potential for far wider application.

At its heart, it is the digital equivalent of a ledger, like a database but with three features that set it apart from standard databases.

- First, a distributed ledger technology is immutable – or ‘append only’ – meaning that data cannot be overwritten, amended or deleted; it can only be added to.
- Second, it is distributed. No central authority or organisation has sole possession of the data. Instead, a copy of the whole database is held by each member of the network and they collaborate to validate each new entry before it is written to the ledger. As a result, there is no centralised authority in control of the data and each participant has an equal status in the network: equal responsibility, equal rights and an equal stake.
- Third, it is transparent. All entries in the ledger are visible to all who have a copy.

The main feature of the ledger used for ARCHANGEL is its immutability: because data on the ledger can not be amended, it prevents tampering of the fingerprints after the fact.

ARCHANGEL uses what is called a 'permissioned blockchain' – anyone can keep a copy of the ledger, but only participating AMIs are given the ability to write operations into the chain.

Permissioned blockchains have particular advantages over typical public blockchains. The most common criticism of blockchain is that it is sensitive to '51% attacks'[7]. An operation on the blockchain is validated when a majority of nodes confirm the hash of the block to be written to the ledger. As there are typically so many members, it is deemed almost impossible to add a fake transaction, as more than 50% of the network's processing power would have to verify the transaction. However, with enough resources – ie by members combining their processing power or one member joining multiple times – any anonymous actor can take over the network by reaching 51% of stakes or 'mining power'. This is a higher risk in a public, anonymous blockchain. In a permissioned blockchain, the risk of '51% attacks' is lessened because every member has been invited, and their identity is known.

Because ARCHANGEL uses a permissioned blockchain, each member of the network gets a vote when 'sealing' the record, meaning that more than half of participating archives would have to collude to somehow falsify a record.

ARCHANGEL also uses a proof-of-authority[8] consensus mechanism. This also puts all the participating archives on equal footing in providing assurance and accountability for each other, nationally and internationally.

## What data is stored on the ARCHANGEL blockchain?

ARCHANGEL needs to prove to the public that records shown to them are the same as those received by the archives. To do so, the system stores the fingerprints of all the records in the blockchain. It does not, however, store the records in the blockchain, mainly for one reason: the information on a blockchain is visible to anyone who has a copy, therefore keeping the records on the blockchain would make it impossible to use it for closed records.

The sensitivity of public records sometimes extends to their filenames or descriptions. Adding these metadata fields to the blockchain would therefore not be appropriate. As a team, we settled on a selection of fields that included an archival reference and the record's checksum: a unique alphanumeric string generated by a mathematical algorithm that changes completely if even one bit is altered in the file.

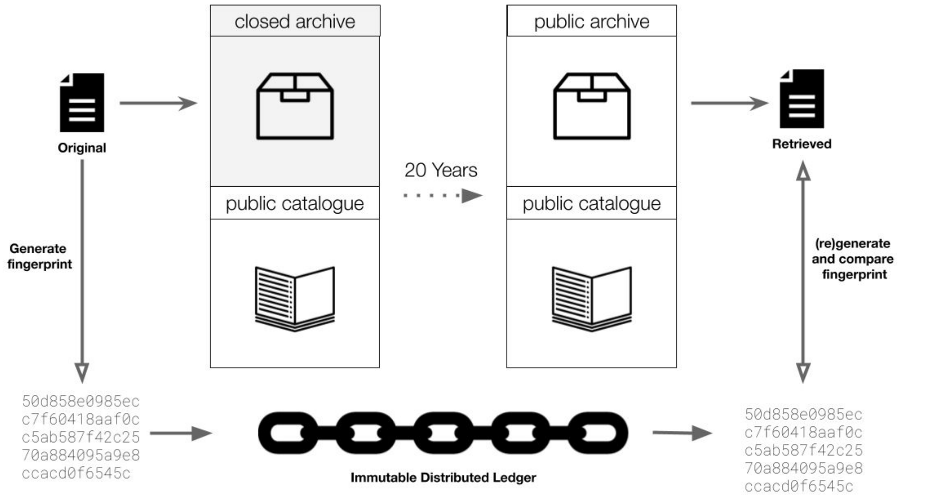
In this way, a researcher can lookup the archive reference in the blockchain to fetch its original checksum – recorded perhaps decades ago – and compare it with the checksum of the record they have just downloaded.

In summary, the ARCHANGEL blockchain enables:

- an archive to upload metadata that uniquely identifies specific records
- that data to be sealed into a 'block' that cannot be altered or deleted without detection
- a copy of the data to be shared with each of the other trusted members of the network for as long as the AMIs maintain it.

[7] Open Data Institute (2016), 'Applying blockchain technology in global data infrastructure' https://theodi.org/article/applying-blockchain-technology-in-global-data-infrastructure/

[8] Unlike proof-of-work blockchains, which rely on members of the chain computing intensive mathematical challenges ('mining') as a way to prevent 51% attacks. This makes such blockchains extremely energy-hungry. The proof-of-authority type of distributed ledger used in ARCHANGEL does not rely on such a mechanism, and is therefore unlikely to ever consume as much energy as a small country.


closed archive
public catalogue
public archive
public catalogue
Original
Generate fingerprint
20 Years
Retrieved
(re)generate and compare fingerprint
50d858e0985ec c7f60418aaf0c c5ab587f42c25 70a884095a9e8 ccacd0f6545c
Immutable Distributed Ledger
50d858e0985ec c7f60418aaf0c c5ab587f42c25 70a884095a9e8 ccacd0f6545c

Illustrating the storage of metadata about closed records in an immutable ledger
From ODI Lunchtime Lecture: Can technology help reinvent national archives for the 21st Century?

## Does this really need a blockchain?

The ARCHANGEL blockchain seems to be a good solution to the digital preservation needs we have identified earlier. As an immutable storage solution, it creates a barrier against tampering of the record after the fact. Its transparency and the ability for anyone to keep a copy makes the record verifiable. And for the sake of performance, the ledger is separate from storage. This also enables use for closed records.

But if the point of the ARCHANGEL project is to try to defend AMIs from allegations of improper conduct, is an infrastructure based on a blockchain the best vehicle?

There is a public perception that blockchains go hand in hand with cryptocurrencies, a domain of wild speculation and rampant crime.[9]

Its reputation also suffers from a tendency by technology vendors to oversell technical solutions to complex problems. For example, IBM's widely publicised blockchain work purports to be a groundbreaking example of secure new technology. Its electronic bill-of-lading proof-of-concept claimed, for example, that it reduced transaction time from five-to-seven days to under one second, failing to mention that the five-to-seven days was for a paper-based system that was rendered obsolete decades ago by EDI systems that IBM itself played a large part in designing.[10]

Can ARCHANGEL help strengthen trust in AMIs using a technology which is itself sometimes seen as untrustworthy, or at least over-hyped?

The core feature ARCHANGEL provides is the ability to verify the integrity of an electronic artefact produced by an archive. It does this by providing a tamper-proof log of artefact checksums. Are there alternative technologies we might be able to use to achieve the same ends?

[9] Coin Telegraph (2019), ‘Report: Indictment Reveals Connection to Bitfinex, QuadrigaCX’s Shadow Banking Services’, https://cointelegraph.com/news/report-indictment-reveals-connection-to-bitfinex-quadrigacxs-shadow-banking-services

[10] European Paten Register (1991), ‘European patent EP0507717A2: Method and apparatus for interchange of customization characteristics of formatted business data’, https://register.epo.org/application?number=EP92480032

## Could we have used a good old-fashioned database?

Do we need distributed ledger technology at all? Could ARCHANGEL simply take the form of additional metadata, published in the archive's catalogue database?

Consider how the process of verifying a document would work when using a database which is neither distributed nor immutable. Having obtained a document provided by an archive and generating the document fingerprints, we are then going to compare the fingerprints against a database provided by that selfsame archive. If the fingerprints match, it actually does nothing to establish the authenticity of the document.

What if a third party produces another copy of the database which differs from the archive's current database? We are then in the same position as trying to establish the true copy of the artifact in question. There is no way to prove which, if either, of the two versions is a true copy of the database.

In a situation like this, those disinclined to trust an archive will not be reassured and will remain unconvinced that integrity can be guaranteed.

## Could we have used a merkle-tree solution such as Git?

Rather than a conventional database, perhaps a more unconventional datastore might be more suitable. Perhaps something built on a Merkle tree (a tree in which every leaf node is labelled with the hash of a data block, and every non-leaf node is labelled with the cryptographic hash of the labels of its child nodes).[11] Something like Git[12] perhaps, the source-code control system used by millions of programmers around the world, notably in the popular GitHub platform.

Within a Git repository each entry – typically a set of software changes, but document metadata in the ARCHANGEL case – forms part of a journal. New entries are linked to one or more previous entries, and will in turn, be linked to by subsequently entries. Entries can be cryptographically signed, allowing their origin to be verified against published public keys.

The features sound very similar to those offered by a blockchain-based solution, which should not be a surprise: merkle trees are also used as the tamper-proof data structure in blockchain.

However, an ARCHANGEL built around Git, or a system that is similar, would still suffer the same fundamental flaw as the previous database example. It would again simply require institutional trust in a master copy for copies to stay consistent with each other. This is because, while the core data structures are tamper-proof, the design of Git is such that additions to the journal are not automatically duplicated to the copies. Instead the owner of one copy must ask for updates from a chosen peer in the network to synchronise with each other.

Even making the Git repository publically available and signing entries is not proof against deliberate tampering. Any publically available Git repository is a snapshot in time and, in any case, is not necessarily the 'master' copy.

## Could we have used durable storage instead of a distributed ledger technology?

There are a number of long-term durable storage providers, perhaps the most well known is Amazon's S3 Glacier product. Amazon describes S3 Glacier as "a secure, durable, and extremely low-cost cloud storage service for data archiving and long-term backup". ‘Secure’ here means secure against theft or tampering, while ‘durable’ means the data will be preserved as-is without loss or corruption.[13]

[11] Wikipedia (2019), ‘Merkle tree’, https://en.wikipedia.org/wiki/Merkle_tree
[12] https://git-scm.com/
[13] Amazon claims 99.999999999% durability.

The service offers a variety of access policies, including a write-once read-many policy that provides protection against tampering even by the data owner – data can be written once, but then can not be altered or deleted. S3 Glacier is sufficiently well developed, both in terms of technology and process, to achieve compliance with the UK Government's G-Cloud and the US Department of Defence Data Processing requirements, among others.

In a system based on Glacier, we could generate fingerprints of our archived records, and write them into our Amazon S3 Glacier storage. At a future time when the record is produced, archives can provide a pointer to the corresponding Glacier entry. Interested third-parties could find the fingerprint within Glacier and verify the record against it.

This would certainly be a workable system from a technological point of view. However, it fails a number of our non-functional requirements:

- This system is neither transparent nor open. Interested third parties cannot see what has been written to Glacier storage, which reduces public scrutiny
- Once our third parties do know what they want to retrieve, then that data is not open and available to them.
- They need to pay a data-retrieval fee to access the data.

S3 Glacier, and services like it, are built to serve a particular market for long-term data archival with infrequent retrieval which doesn't need to be fast – such as auditable accounts – and thus solve a different problem.

For example, a financial institution has a number of reporting and data retention obligations set by the industry regulator. It needs to report the specified data promptly, perhaps at the end of the tax year, and then store it securely for a period of time. In practice, so long as the financial institution operates in a normal way, within the bounds of accepted business practice, it does not attract the attention of the regulator and so may never need to retrieve data from the store.

S3 Glacier  – and the clue is in the name – provides *cold storage* – a data storage system that is accessed less frequently and doesn't require fast access. The expectation is that the data is rarely, if ever, needed again. Consequently, the fee structure actively discourages data retrieval: writing and storage is cheap (indeed, very cheap), but data recovery is comparatively expensive. Data retrieval can take hours to complete – unless you pay an even bigger fee – and the retrieved data is available for only a limited period.

Hash verification requires immediate access. If someone wants to verify the authenticity of a document, they should be able to do it immediately, not many hours later after paying a fee. Observers should, if they wish, be able to access the entirety of records. Records need to be transparent and this form of durable storage hides the record.

Setting this problem aside – perhaps an ARCHANGEL-like system could come to some suitable commercial arrangement with a storage provider. But the presence of a commercial third party presents another barrier to trust. Public trust in large technology companies is eroding[14] and this is unlikely to change in the near future. It is not particularly farfetched to argue that if an archive is paying a service provider to store the artefact fingerprints, that archive might also pay that provider to change or delete fingerprints if necessary.

Lastly, there is the additional question of organisational longevity. ARCHANGEL aims to be a system for the very-long term. The average lifetime of an S&P listed company (widely-used gauge of performance for US companies, often used as a

[14] EY (2019), 'Why trust in tech giants is eroding, and how it can be rebuilt', https://www.ey.com/en_gl/trust/why-trust-in-tech-giants-is-eroding--and-how-it-can-be-rebuilt

proxy for ‘the market’) is 15 years[15]. ARCHANGEL would need to rely only on organisations guaranteed to exist for decades and beyond.

### A technological solution, a social contract

None of the technologies we considered above fit the bill for ARCHANGEL, but not only for purely technological reasons. One could build the most secure system yet, but unless that system can be comprehensively examined and critiqued by others, we are reliant solely on the system builder's assurance whom we then also need to trust.

Blockchain technology provides the transparency and openness demanded by ARCHANGEL by enabling the following:

- The ARCHANGEL blockchain is, by design, publicly readable to provide scrutiny; while the ability to write to the blockchain is limited to participating AMIs
- Various tools – many of them open source – can be used to access and keep a copy of the chain[16].
- The consensus mechanism replaces a single institution's assurance with a collective assurance.
- No one participant can change the record, no matter how much they may wish to.
- In the unlikely event that a cohort attempted to subvert the record, that attempt would immediately be detected by others in the consortium.

## Temporal content hashes for content-aware document fingerprinting

Building and delivering this system is complicated by another challenge in digital archiving – changing digital formats. Digital formats shift over time – new ones are created, old ones are retired – and new software might no longer support the same formats supported by previous versions. For example, some modern video players will not open older video files. This presents a fundamental challenge for digital archivists striving to preserve documents for the future.

Video formats rapidly become obsolete, motivating format shifting (transcoding) as part of the curatorial duty to keep content viewable over time. This leaves video preservation at risk of modification – either due to direct attack (tampering), or due to accidental corruption, such as truncation or frame corruption due to bulk transcoding errors.

Cryptographic hashes operate at the bit level. They are effective at detecting video tampering, but not for videos undergoing transcoding: a bit-level fingerprint of a video using format A would be entirely different to that of the exact same video in a different format B. This means that, to guarantee the integrity of video records (such as video proceedings of the UK Supreme Court, which are deposited and kept at the National Archives), one has to find a solution involving content-aware and format-agnostic hashing of the audio-visual stream itself.

Our approach for ARCHANGEL was therefore to explore and prototype the creation of hashes using machine-learning methods, particularly for image and video content, rather than ‘traditional’ bit-level hashes.

[15] The situation varies around the world. In Japan, a number of businesses are over 1000 years old. However, none of those are large technology outfits that could provide this kind of service.
[16] The ARCHANGEL prototype was built on the Ethereum stack, which provides a number of tools and libraries at https://geth.ethereum.org/downloads/

### Content-aware hashing: not new, still a challenge

A number of technologies and solutions already exist that attempt to tackle the challenge of content-aware hashing for video. One of the most well-known ones is ContentID, used to detect and prevent copyright infringement on the Youtube video platform. While Google says[17] they have invested over $100m into ContentID, the system is not infaillible[18].

Using a content-aware hashing to guarantee the integrity of records in a national archive is extremely challenging:

- False negatives are unacceptable: failing to detect tampering would make trust in the system collapse, and make the whole thing useless
- False positives are also highly undesirable: if the system over diagnoses tampering, trust collapses too – or a trusted party has to take on manual checking, limiting the usefulness of the automated system.

The solution proposed by the research team on the ARCHANGEL project uses Deep Neural Networks to create a ‘temporal content hash’[19] that is trained to ignore transcoding artifacts, but is capable of detecting tampers of a few seconds duration within typical video clip lengths within a short amount of time (ie minutes or hours).

By using both original video as well as derived, transcoded videos as training material, the machine-learning system is able to accurately differentiate whether glitches and noise – in either the audio or video signal – are caused by transcoding and format-shifting, or are caused by any undesirable process, including corruption of the files in storage, or tampering of the record.

It is worth noting that this system is very different, in its approach and its purpose, from other kinds of content-aware digital fingerprinting systems. Systems like ContentID attempt to answer the questions like: “Does this piece of content closely resemble anything else in our catalogue”; the temporal content hashes in ARCHANGEL aim to answer the question: “Can we say with high certainty whether differences between content pieces A and B are artefacts of format-shifting or the result of tampering?”.

[17] *Protecting what we love about the internet: our efforts to stop online piracy*, Google Policy blog, Nov 2018 https://www.blog.google/outreach-initiatives/public-policy/protecting-what-we-love-about-internet-our-efforts-stop-online-piracy/

[18] *Sorry professor, old Beethoven recordings on YouTube are copyrighted*, Ars Technica, March 2018 https://arstechnica.com/tech-policy/2018/09/how-contentid-knocked-down-decades-old-recordings-of-beethoven/

[19] More on the algorithm at https://arxiv.org/abs/1904.12059

# Reinventing national archives

## The ARCHANGEL prototype

To test the technology solution proposed in the ARCHANGEL project, a prototype was created to explore the feasibility of the technology, as well as the desirability and suitability of the solution for AMIs.

The prototype consisted mainly of two intertwined systems:

1. A permissioned network built on the Ethereum blockchain stack – the ARCHANGEL distributed ledger.

2. A digital preservation tool, implemented both as a simple web application and a desktop app. Designed as closely as possible to the typical user experience of the digital archivist, the tool uses the language and metaphors of the Open Archival Information System (OAIS).

   The tool integrates with another system typically used by archives such as the National Archives', the DROID (Digital Record Object IDentification) file format characterisation tool. We used it to help the digital archivist describe the digital records being deposited, create fingerprints, and store those in the ARCHANGEL blockchain network.

   The tool also included the feature described above of creating temporal content hashes for video. Given the resource-intensive process of training deep neural networks, the feature was achieved not directly in the application, but instead integrated through an API, by software deployed on the University of Surrey digital infrastructure.

Most of the software developed for the prototype is available on GitHub at: https://github.com/archangel-dlt/.

Early tests of the prototype demonstrated the technical feasibility of the system. But – given that such a distributed system requires several archives to use it and participate in hosting nodes of the blockchain network – we still needed to discover whether external digital preservation practitioners would: understand what the prototype did; see value in the methodology; and have an interest beyond a simple pilot.

## An international pilot

To address this question, in April 2019, the ARCHANGEL project team conducted a pilot of the prototype with national AMIs in five different countries: the UK, Australia, Norway, Estonia, and the US.

The aim of the pilot was to gather insights on the ARCHANGEL concept and its prototype from participating institutions as they used the prototype system in parallel with their routine archival process for a limited period of time.

To better understand the viability, desirability, feasibility and usability of the prototype, the project team ran user research with pilot participants. Some of the questions covered by the user research included:

- Did they engage with the prototype? Was it appealing? Did it fit their usual mental model? Did the prototype help solve problems and pain points known to the archivists?
- Did the archivists understand the underlying technology of ARCHANGEL (distributed ledger technologies, hashing and computer vision)? Would that understanding matter in order for them to have trust in the system and, importantly, to see it as an improvement on existing processes and structures?
- Did the participating archivists understand that they are mutually underwriting the integrity of other archives in return for others doing the same? Did they assign value in being part of a blockchain?
- Was it more important to them that all tampered videos should be reported (with the occasional false positives); or should videos be reported as 'tampered with' only when the system is fully confident?

## Results and findings

> *The key thing for us is blockchain environment really. [...] Where I work there's a 20-year closed period. That's a long time for files and metadata to be sitting around. If we can prove authenticity and integrity through a tried and tested technology, [...] then that's going to help us a lot just through multiple generations of technology.*

The user research-generated insights and gave us access to a wide range of ideas, and a broad perspective on the experiences of various kinds of archivists.

Overall, the participants of the pilot showed positive sentiment towards the ARCHANGEL concept and prototype. They found the interface clean and simple. However, some were not clear about the inner workings of the prototype.

The main differentiator in the response to the pilot was whether the participants had prior understanding and knowledge of blockchain technology.

- Those who did, understood the underlying value of mutually underwriting the integrity of other archives and trusted the system.
- Those who did not, while appreciating the value of software that creates fingerprints for digital records, understood ARCHANGEL's trust framework as inherently residing with the participating organisations.

The pilot also generated insights that would be useful for future iterations, such as making the software more modular to enable configurable workflows, and an additional interface for independent validation of checksums. Those insights are discussed in detail in the research report[20].

[20] ARCHANGEL pilot - User Research Report, ODI 2019 https://docs.google.com/document/d/1GQL23E9aQNpc5vB1UtETEFz0tt8hYAb_qNlmldDQY8Y

# We shape our tools, and in turn they shape us

Collaboration is essential to the value offered by blockchain technology. In the proof-of-authority system chosen for ARCHANGEL, every participating organisation has an equal role in providing assurance of integrity.

In a large enough network, this means that collusion and tampering would be difficult to achieve and relatively easy to notice. And since every participating AMI keeps a copy of the ledger, this creates a form of insurance against crises such as political turmoil in a given country or region – outsourcing some of the preservation effort to the network.

This means the success of the ARCHANGEL blockchain relies on a minimum number of participants – we think at least seven – from as many different institutions as possible. Without a minimum number of participants the trust that the technology engenders is in danger of being lost.

In a live, future environment we would hope to involve participants beyond the archive sector such as news organisations, and other transparency-minded groups who – as well as providing external oversight – also have a stake in the assurance of public records.

A distributed approach to assuring trust is another step in the long-running trend of memory institutions relying on each other. Archival capability is distributed and shared in terms of know-how and in the development and maintenance of tools and archival resources (for example PRONOM[21] and LOCKSS[22]). In the case of web archives, the collections themselves overlap and content is shared, as archives supply each other with content to fill gaps in their collections.

Because the challenge is great and as institutions' archives are quite small, this approach is the key to winning the technology arms race between archives and those parties who use the tools to falsify our digital inheritance.

If this approach to distributed services is successful, it is exciting to think about what else could be distributed in the future.

[21] National Archives (2006), 'PRONOM',http://www.nationalarchives.gov.uk/pronom
[22] Stanford University (2019), 'LOCKSS', https://www.lockss.org

# Further reading and references

The project team created papers, articles, blog posts and presentations in the course of the two-year ARCHANGEL project. Several of them were used in part and summarised in this document, but they typically go much further, especially in explaining the technical aspects of the system.

- *ARCHANGEL: Trusted Archives of Digital Public Documents* – ACM Document Engineering *2018.* https://arxiv.org/abs/1804.08342
- *ARCHANGEL: Tamper-proofing Video Archives using Temporal Content Hashes on the Blockchain – CVPR Blockchain Workshop 2019.* https://arxiv.org/abs/1904.12059
- *A Blockchain For Archives: Trust Through Technology,* *https://www.archivoz.es/en/a-blockchain-for-archives-trust-through-technology-2/*
- *Underscoring archival authenticity with blockchain technology*, Insights, https://insights.uksg.org/articles/10.1629/uksg.470/
- *Blockchain's potential role in the future of archiving,* ODI blog https://theodi.org/article/blockchains-potential-role-in-the-future-of-archiving/
- *Challenges in using blockchains to build trust in digital archiving*, ODI blog. https://theodi.org/article/challenges-in-using-blockchain-to-build-trust-in-digital-archiving/