CINE-GT.3401: The Culture of Archives, Museums, and Libraries

May 3, 2019

Claire Fox

Final Paper

Trusting the Trustless Ledger: Considering Blockchain for Long-Term Digital Preservation

> The prospect of a world in which all text, audio, picture, and video documents are in digital form on an easily modifiable media raises the issue of how to certify when a document was created or last changed. The problem is to time-stamp the data, not the medium.
>
> Stuart Haber and W. Scott Stornetta, *How To Time-Stamp a Digital Document*, 1991

> Claims of trustworthiness are easy to make but are thus far difficult to justify or objectively prove. Establishing more clear criteria detailing what a trustworthy repository is and is not has become vital.
>
> ISO 16363:2012, *Audit and Certification of Trustworthy Digital Repositories*
> Last reviewed and confirmed 2017

There is no set definition for blockchain technology. There are a range of assumptions and questions, cultural associations and contexts, use cases that have changed over time, and aspects of the technology that practitioners reject as "buzz words" that are "ambiguous and meaningless" (Reddit, 2018). Blockchain is still in a stage where it can mean different things to different people, and those differences in meaning have a significant impact on the way the technology is implemented. However, an excellent summary definition comes from Finn Brunton, a historian and assistant professor of Media, Culture, and Communication at New York University. It is as follows:

> Blockchain is a persistent, transparent, public, append-only ledger. It is a system that you can add data to, and not change previous data within it. It does this through a mechanism for creating consensus between distributed parties that do not need to trust each other: they just need to trust the mechanism by which their consensus is arrived at. (WIRED, 2017).

The primary points of the blockchain infrastructure are all present in this definition. Blockchain is distributed: there is no central authority, no "trusted third party" that governs the network. Blockchain is append-only: you can't edit, overwrite, or remove data, as the records (blocks) in the network (chain) are immutable. Perhaps most significantly, blockchain is trustless: even when it is distributed across a collective of 20,000 personal computers, no single member needs to worry about trusting any central governing body, or trusting any single member of the collective, because the system is governed by consensus algorithms (for example: proof of work, or proof of stake) which stand in for human decision-making. As Brunton writes, the distributed parties don't need to trust one another, they need only trust the mechanism: the consensus algorithm.

In contrast to the wild west of accepted blockchain definitions, archival digital preservation has precisely-defined recommended levels (the National Digital Stewardship Alliance (NDSA) Levels of Digital Preservation), prescriptive workflows (Open Archival Information System (OAIS) Reference Model), and peer-reviewed ISOs to use in tandem with those workflows (including the *Audit and Certification of Trustworthy Digital Repositories* (TDR), or ISO 16363:2012). These models and guidelines are generated by the archival community, in response to the practitioner work performed by archivists everyday. When assessing objects and records for preservation, archivists typically ask questions including: Who created this work, and how did it arrive here? Who owns it now? Where will it be stored? How will people access it, both now and in the future? These questions fall under the broad categories of provenance, copyright, storage, and access, and in digital preservation, each of these categories is constantly being re-defined and re-assessed.

What blockchain and digital preservation share in common is that they are both relatively new fields requiring iterative work that addresses the same urgent need: to find a way to to ensure that rapidly-proliferating digital files will be retrievable and uncorrupted in future generations. This is fragile work, as the digital landscape and the software that comes with it is always evolving to fit the needs and demands of complex webs of socio-technical networks. The hardware on which the software runs is always pushing toward the edge of obsolescence, and the labor that maintains the hardware is contingent on lagging resources. In this constant state of flux, archives are increasingly reliant on multinational technology companies like Amazon, via storage on Amazon Web Service's (AWS) S3 Glacier, to be custodians of their digital assets and the metadata for their physical ones. In some cases, when consensus can't be reached, digital workflows simply don't exist.

A Note on AWS Pervasiveness in Archives

It would be difficult to identify a better known central authority than Amazon. What began as a bookseller expanded into an e-commerce marketplace, television and film production and distribution, and massive cloud computing infrastructure, which is the venture supporting all other Amazon endeavors. At the top of it all is Jeff Bezos, who in addition to his role as CEO of Amazon is the owner of *The Washington Post*. This role presents some complications, as delineated by Maria Bustillos during an interview on Mozilla podcast *IRL*. "There's one guy on top who really owns that property," she says, "and if *The Washington Post* wants to criticize Amazon, they've got to think really carefully about whom they're going to offend. And the same principle applies all throughout media. Who gets to decide what we talk about, and how we talk about it?" (Zomorodi, 2019).

Similarly, it would be difficult to overstate the influence of AWS on digital preservation. Though statistics regarding the percentage of archival repositories using AWS aren't available, an indication of the influence of AWS on digital preservation is through its partnership with Preservica, a software company that specializes in digital preservation technology. Preservica's clients include MoMA, Yale University Libraries, The UK National Archives, and the Frick Collection, among a range of other libraries, archives, and museums, including archives for state agencies. On its AWS Partner Story page, Preservica marketing director Michael Hope notes that many of the company's customers don't have the resources to support their own local IT department. With the aim of helping resource-lacking state agencies and other organizations with complex metadata and storage needs in mind, Preservica joined the AWS Partner Network (APN) and based its Preservica Cloud Edition on the AWS Cloud. Said Hope: "We selected AWS to host our digital preservation and access software because we knew we could provide a very cost-effective solution that helped to ensure security and storage durability for our customers" (AWS, 2019).

AWS, with its range of cloud-based services geared toward scalability from the smallest community archive to the largest hedge fund, could certainly bear the brunt of a state agency's digital preservation infrastructure needs. But one can't help but wonder: who decides what we get to archive, and how we make it accessible?

Absence of File-Based Workflows

An example of a lack of digital workflows can be found in the premier research library in the United States: the Library of Congress. A walk through the Recorded Sound Processing Unit at the National Audiovisual Conservation Center (also known as the Packard Campus for Audiovisual Conservation) in Culpeper, Virginia is as close an approximation to an embodied preservation workflow as one might find. Physical donations to the NAVCC are accessioned, grouped into collections on shelves and pallets by donation number, and then begin their progression through a series of rooms where they are labeled, refined, removed from original boxes and interleaved for consolidation with similar donations, prioritized, cleaned, and digitized for access via requests from the Public Services office.

When visitors and archivists walk through the various processing rooms provides a visual understanding of how metadata is used and created throughout the processing workflow: donors are accounted for, as is the content of the donations. The condition of the materials and their housings is assessed, and preservation actions are taken based on the assessments. Materials are prepared for long-term storage, both in terms of their physical condition and the descriptive and administrative metadata associated with them. Formats are monitored for obsolescence, and digitization workflows are available for the sake of providing access and ensuring playback in the future. Patrick Midtlyng, the Head of the Unit, emphasizes that the workflows in the Recorded Sound Processing Unit are constantly being examined to create more space and process more material with greater efficiency (Fox, 3).

Strangely absent from this workflow is a methodology for processing digital file-based donations. How might a collection be processed if it was born-digital? Or if it were digitized prior to donation, and donated via hard drive? Can hard drives join in with this multi-room processing workflow? And if they can, what about donations made via cloud-based file-transfer, that don't have any physical manifestation an archivist could hold in their hand?

While these questions are being researched within the Recorded Sound Processing Unit, they have not yet influenced the workflow for physical donation processing. At present, digital files do not have a set processing workflow. Midtlyng notes that the lack of workflow is problematic within the context of the Library of Congress, which aims to fulfill the mantra of "preservation for access" (Mashon, 2018). Without a workflow, processing actions cannot be taken. As such, no digital file-based donations are currently being processed for preservation in the Recorded Sound Unit. All these files are in danger of losing contextual metadata, experiencing data degradation, and hardware obsolescence before they even begin processing. By the time these files may begin processing, archivists may not know what metadata is trustworthy enough to write into the historical record.

Blockchain's Foundations

Though the Recorded Sound Unit is conducting research regarding possible file-based workflows, the current absence of infrastructure leaves space to consider what that future workflow might need. Going further, the absence leaves space to consider which technology – maybe one with a perceived lack of precedent in the field – could be tested, with an eye toward finding new preservation workflows along the way. It is in this space that archivists could consider blockchain technology as a possible infrastructure for digital preservation.

Blockchain technology, popularized through cryptocurrencies in the wake of the 2008 financial collapse, has early implementations that align with the values of archival digital preservation. Blockchain was first developed in the late 1980s by Dr. Scott Stornetta, a physicist, and Dr. Stuart Haber, a cryptographer, who were colleagues at a research center called Bellcore (Whitaker, 2018). Haber and Stornetta had been thinking through the accelerating rate of digital file creation, along with the ability to alter those files, which came hand-in-hand with the rise of personal computing. As Amy Whitaker writes in her *Wall Street Journal* article "The Eureka Moment," "[t]hey wondered how we might know for certain what was true about the past. What would prevent tampering with the historical record – and would it be possible to protect such information for future generations?" (Whitaker, 2018).

Dr. Stornetta realized (somewhat infamously while having dinner with his family at a Friendly's restaurant in Morristown, New Jersey) that this could be achieved by removing the need for a central trusted record-keeping authority. Instead, records could be kept across many computers, making it difficult for one bad actor to manipulate all copies of the dispersed, shared ledger. This realization eventually led to Drs. Haber and Stornetta to create a cryptographically secure archive that would store items of data in time-stamped digital groups called blocks. Drs. Haber and Stornetta wrote about this blockchain framework in a 1991 paper titled "How to Time-Stamp a Digital Document," which was published that year in the *Journal of Cryptology*. As Whitaker explains:

> Each block includes an alphanumeric code called a "hash" summing up its data. The hash of each completed block also appears in the next one in the chain, which means that to alter one block you would have to alter all the ones connected to it. These cryptographic dominos function together to protect against tampering or fraud. (Whitaker, 2018).

Drs. Haber and Stornetta's blockchain framework places blockchain technology squarely within the mandate of archives: archives exist for the sake of providing future generations with a window into the past, and their blockchain's ability to create a time-stamped chain of records builds a chain of custody (or provenance) directly into the technology. But this only achieves part of an archivist's mandate. An archivist needs to make sure that any digitized, transferred, or born-digital files will be renderable in a far-distant future: those files need to be findable, retrievable, and played back. The Digital Preservation Coalition calls this challenge the "always emerging digital preservation challenge" (Digital Preservation Coalition, n.d.). "Information can only be accessed and functions can only be executed through a computer," their website reads. "As technology becomes more sophisticated this dependence becomes an ever more elaborate chain of inter-dependencies that are hard to track and tricky to maintain."

Digital Preservation and Contemporary Blockchain Workflows

The "ever more elaborate chain of inter-dependencies" is written into the literature on digital preservation, specifically in the ISO Standard Reference Model for an Open Archival Information System (OAIS). As written by Kara Van Malssen in her report, *Planning Beyond Digitization: Digital Preservation of Audiovisual Collections*, OAIS gives a general overview of the fundamentals that make up a repository's operation (Van Malssen, 79). She defines three main areas that work together collectively to achieve long-term preservation and access of digital information. These areas include the external environment, the functional environment, and the information model.

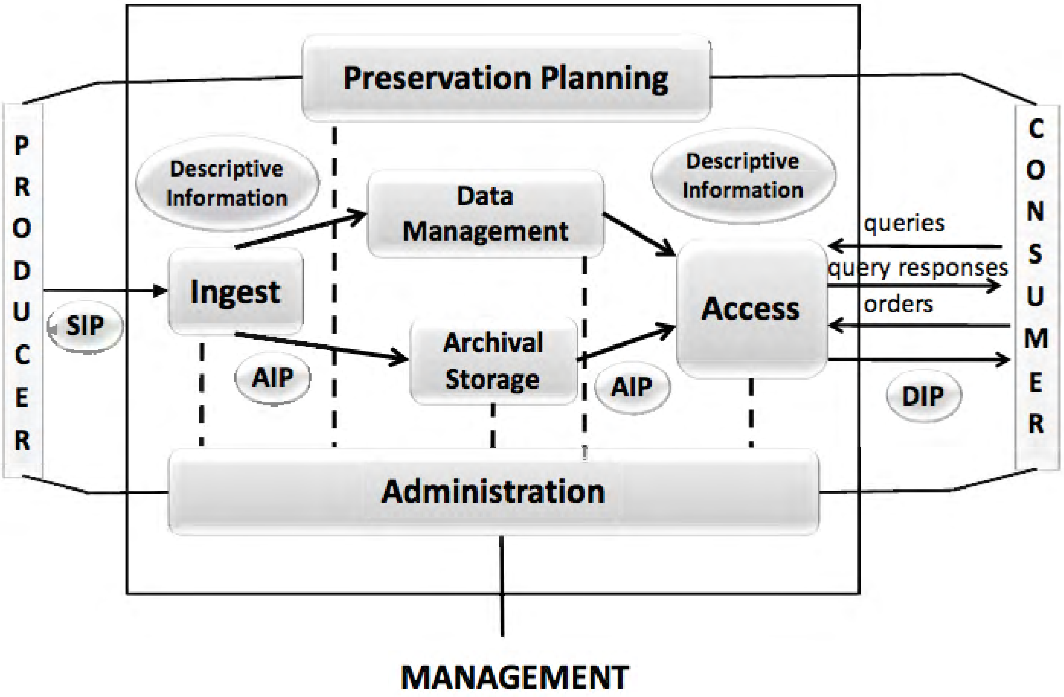

Figure 1: The functional model. Screenshot from the *Reference Model for an Open Archival Information System.*

Within the information model, OAIS describes three different iterations of information packages to illustrate how information is received, transformed, and disseminated within the framework of the repository. An information package describes a digital object and its associated metadata, and it exists across the repository as a Submission Information Package (SIP), an Archival Information Package (AIP), and a Dissemination Information Package (DIP). The SIP is the package acquired from the submitter, and contains the minimum required amount of metadata along with the files (Van Malssen, 81). The AIP includes further metadata added to the SIP by the repository, creating a complete archival object, and additionally includes actions for format migration. The DIP is a limited version of an AIP created depending on the needs of the user requesting the information.

ARCHANGEL, a digital preservation project based at the UK National Archives, is currently performing research regarding the ways in which a digital object might change over time, and how an archive could know whether or not those changes were legitimate and that the record is still trusted as the authentic record (Green, 2018). The National Archives are specifically investigating how blockchain technology might be used to secure these records. In a 2018 presentation at the Blockchain@UBC "The Future of Blockchain Technology" Mini-Conference in Vancouver, B.C., National Archives digital researcher Mark Bell used the OAIS model to illustrate how records might be created on the blockchain and verified against each other. This process served to answer two main questions: how would an archivist know that the initial document hasn't changed, and how might a consumer know that the document they're viewing is the same as the archival document? (Bell, 2018). Bell modeled ways in which records would be created on the blockchain, and then how iterations of objects would be created based on the initial records. These models were based on OAIS information packages, and were further examined to note that the immutability and distribution of the blockchain provides a measure of trust that may not be found in the OAIS without it.

Next Steps: Establishing Trust Through Iteration and Collaboration

In a blog post for *The National Digital Stewardship Residency New York* titled "On the Subject of Trust," Shira Peltzman writes about the use of trust frameworks as a digital preservation planning tool. She specifically writes about the *Audit and Certification of Trustworthy Digital Repositories* (ISO 16363:2012) as 109 distinct criteria for measuring trustworthiness that were designed to be used in tandem with the OAIS (Peltzman, 2014). As Peltzman outlines, trust frameworks perform functions like establishing metrics for evaluating a repository's progress over time, and extends to include external contingencies including fiscal responsibility and the establishment of escrow arrangements in case the repository is no longer able to operate. While establishing these measures of trust might seem out of scope when considering all that is already folded into blockchain, trust in software is just one part of a larger trust-based system.

As has been demonstrated by OAIS, long-term digital preservation is dependent on a range of interconnected dependencies. Blockchain, with its persistent, transparent, distributed, append-only ledgers and consensus algorithms, will initiate new resources within these dependencies for building trust, should archivists choose to implement them. New challenges will arise: chiefly, if digital preservation is going to rely on a consensus algorithm, new standards will need to be written and peer-reviewed to account for algorithmic bias. At the base of this all, there are still humans writing the code, whether or not they need to trust one another.

Reference List

AWS. AWS Partner Story: Preservica. Retrieved May 3, 2019 from https://aws.amazon.com/partners/success/preservica.

Bell, M. (2018, May 25). Does blockchain have a future? Retrieved May 3, 2019 from https://blockchainubc.ca/2018/05/04/blockchain-101-next-to-the-blockathon-for-social-good/

Consultative Committee for Space Data Systems. (2011 September). *Audit and Certification of Trustworthy Digital Repositories*. Retrieved May 3, 2019 from https://public.ccsds.org/pubs/652x0m1.pdf.

Digital Preservation Coalition. (n.d.). Why digital preservation matters. Retrieved May 3, 2019 from https://www.dpconline.org/handbook/digital-preservation/why-digital-preservation-matters.

Fox, C. (2019, February 24). What Can Data Do for Us?: Reconsidering MAVIS with Patrick Midtlyng. (Unpublished class assignment.) New York University, New York, NY.

Green, A. (2018, June 5). Trustworthy technology: the future of digital archives. [Blog post]. Retrieved May 3, 2019 from https://blog.nationalarchives.gov.uk/blog/trustworthy-technology-future-digital-archives/.

Haber, S., & Stornetta, W.S. (1991). How to Time-Stamp a Digital Document. *Journal of Cryptology, 3,* 99-111.

Mashon, M. (2018, September 01). Film Loans from the Library of Congress: September 2018. Retrieved May 3, 2019 from https://blogs.loc.gov/now-see-hear/2018/09/film-loans-from-the-library-of-congress-september-2018/.

Peltzman, S. (2014, October 30). On the Subject of Trust. Retrieved May 3, 2019 from https://ndsr.nycdigital.org/on-the-subject-of-trust/.

Reddit r/bitcoin. (2018). Are Consensus Algorithms and Smart Contracts same? Retrieved May 3, 2019 from https://www.reddit.com/r/Bitcoin/comments/a0xtml/are_consensus_algorithms_and_smart_contracts_same/.

WIRED. (2017, November 28). Blockchain Expert Explains One Concept in 5 Levels of Difficulty [Video File]. Retrieved May 3, 2019 from https://youtu.be/hYip_Vuv8J0.

Van Malssen, Kara. (2011). Planning Beyond Digitization: Digital Preservation for Audiovisual Collections. *Making Invisible Assets: The Preservation of Digital AV Collections*, 68-91. Retrieved May 3, 2019 from https://www.weareavp.com/planning-beyond-digitization-digital-preservation-for-audiovisual-collections/.

Zomorodi, Manoush. (Producer). (2019, February 18). *Decentralize It* [Audio podcast]. Retrieved May 3, 2019 from https://irlpodcast.org/season4/episode6/.