NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

INNOVATION INCREASE: HOW TECHNOLOGY CAN CREATE OPEN, DECENTRALIZED, AND TRACKABLE DATA SHARING

by

Erica Hupka

March 2018

Thesis Co-Advisors: Rodrigo Nieto-Gomez
Ted Lewis

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE March 2018	3. REPORT TYPE AND DATES COVERED Master's thesis

4. TITLE AND SUBTITLE INNOVATION INCREASE: HOW TECHNOLOGY CAN CREATE OPEN, DECENTRALIZED, AND TRACKABLE DATA SHARING	5. FUNDING NUMBERS
6. AUTHOR(S) Erica Hupka	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB number ____N/A____.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

University research must be widely shared to increase innovation; however, regulated and sensitive information must be secured to prevent theft and malicious misuse. The ideal sharing environment will allow universities to openly and, with trust, share verified unique data that is both immutable and ultimately traceable. Many technologies today facilitate pieces of the ideal sharing environment, but are unable to provide all required capabilities. My proposed technology solution capitalizes on the benefits of existing technologies and also proposes new technologies to achieve the ideal sharing environment. If this technology proves successful for university research sharing, it can also be expanded to other fields, including private industry research and development.

14. SUBJECT TERMS blockchain, university, academic research, hyperledger fabric, Turnitin, iThenticate, Google search engine, information sharing, information security	15. NUMBER OF PAGES 95
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89) Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

INNOVATION INCREASE: HOW TECHNOLOGY CAN CREATE OPEN, DECENTRALIZED, AND TRACKABLE DATA SHARING

Erica Hupka
Manager of Emergency Management, University of Kansas Medical Center
B.A., University of Kansas, 2005

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF ARTS IN SECURITY STUDIES
(HOMELAND SECURITY AND DEFENSE)

from the

NAVAL POSTGRADUATE SCHOOL
March 2018

Approved by: Rodrigo Nieto-Gomez
Co-Advisor

Ted Lewis
Co-Advisor

Erik Dahl
Associate Chair for Instruction
Department of National Security Affairs

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

University research must be widely shared to increase innovation; however, regulated and sensitive information must be secured to prevent theft and malicious misuse. The ideal sharing environment will allow universities to openly and, with trust, share verified unique data that is both immutable and ultimately traceable. Many technologies today facilitate pieces of the ideal sharing environment, but are unable to provide all required capabilities. My proposed technology solution capitalizes on the benefits of existing technologies and also proposes new technologies to achieve the ideal sharing environment. If this technology proves successful for university research sharing, it can also be expanded to other fields, including private industry research and development.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

I. PROTECTING SENSITIVE RESEARCH INFORMATION 1
- A. WHY IT MATTERS TO HOMELAND SECURITY ENTERPRISES 2
- B. RESEARCH QUESTION 8
- C. RESEARCH DESIGN 10
  - 1. Data Sharing 10
  - 2. Data Trust and Immutability 10
  - 3. Data Uniqueness Verification 11
  - 4. Data Identification and Traceability 11
- D. LITERATURE REVIEW 11
  - 1. Sharing Information 12
  - 2. Information Security 14

II. DATA SHARING 19
- A. RESEARCHGATE 19
- B. ACADEMIA.EDU 20
- C. arXiv.ORG 22

III. DATA TRUST AND IMMUTABILITY 25
- A. ADVANCED ENCRYPTION STANDARD 25
- B. BLOCKCHAIN 26
- C. ESTONIA EXAMPLE 31
- D. HYPERLEDGER FABRIC 37

IV. DATA IDENTIFICATION AND TRACEABILITY 43
- A. DIGITAL OBJECT IDENTIFIER 43
- B. PERSISTENT UNIFORM RESOURCE LOCATOR 44
- C. INTERNATIONAL STANDARD SERIAL NUMBER 44

V. DATA UNIQUENESS 47
- A. TURNITIN 47
- B. iTHENTICATE 48
- C. GOOGLE SEARCH ENGINE 49

VI. ANSWERING THE RESEARCH QUESTION 51
- A. OPEN DATA SHARING 51
- B. TRUST AND IMMUTABILITY 51

C. DATA IDENTIFICATION AND TRACKING 52
D. DATA UNIQUENESS 55

VII. CONCLUSION 63

LIST OF REFERENCES 67

INITIAL DISTRIBUTION LIST 71

LIST OF FIGURES

Figure 1. How Bitcoin Works 28

Figure 2. Blockchain Use Cases.....30

Figure 3. Estonian Information System.....33

Figure 4. Hyperledger Fabric Model.....38

Figure 5. Hyperledger Membership 39

Figure 6. Hyperledger Contract Confidentiality.....40

Figure 7. Hyperledger Separating Transaction Endorsement from Consensus.....41

Figure 8. Charter Functions.....55

Figure 9. Artificial Intelligence Architecture.....56

Figure 10. How to Send a File to the Ledger 58

Figure 11. How to Add a File to the Ledger 59

Figure 12. How to Get a File from the Ledger.....60

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

Table 1. Data Sharing Technology Evaluation 24

Table 2. Comparison among Public Blockchain, Consortium Blockchain, and Private Bockchain 36

Table 3. Typical Consensus Algorithms Comparison 37

Table 4. Data Immutability and Trust Technology Evaluation 42

Table 5. Data Identification and Traceability Technology Evaluation.....46

Table 6. Data Uniqueness Technology Evaluation.....50

Table 7. Compiled Technology Evaluations.....65

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

DARPA	Defense Advanced Research Projects Agency
DOI	Digital Object Identifier
FBI	Federal Bureau of Investigation
ISSN	International Standard Serial Number
ORCiD	open researcher and contributor identification number
PIN	personal identification number
PKI	public-private key infrastructure
PURL	persistent uniform resource locator

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

Research universities need to share information, whether through teaching or disseminating key innovations to society. However, universities should not share all research with everyone.[1] If universities fail to protect sensitive data, it could harm citizens personally, financially, or potentially fatally if acquired by malicious actors.

The greater the number of researchers collaborating on a complex problem, like cancer, the less time it may take to find cure.[2] Information sharing is part of a university's mission and potentially a regulatory or legal obligation.[3] Because research is valuable, it needs to be propagated to reach its potential. As Steven Johnson writes, "The trick to having good ideas is not to sit around in glorious isolation and try to think big thoughts. The trick is to get more parts at the table."[4] Researchers must share the data "parts" on the global collaboration "table" to realize big ideas.

Scholars must balance open access with restricted access to sensitive information, when sharing university research data. Therefore, universities must track and secure sensitive data to prevent nefarious actors from stealing or weaponizing the information. For instance, if a malicious actor stole data generated from biodefense projects, United States service members and citizens, along with other United States allies, could be at great

---

[1] "Responsible Conduct of Research: Data Acquisition and Management," Columbia University, accessed May 11, 2017, http://ccnmtl.columbia.edu/projects/rcr/rcr_data/foundation/.

[2] Robert W. Rycroft, "Does Cooperation Absorb Complexity? Innovation Networks and the Speed and Spread of Complex Technological Innovation," *Technological Forecasting and Social Change* 74, no. 5 (June 1, 2007): 565–78, https://doi.org/10.1016/j.techfore.2006.10.005.

[3] To provide evidence for the assertion that information sharing is a part of the university's mission, I queried university mission statements cited here. The National Institutes of Health reference provides evidence for the regulatory and legal obligation of research universities to share information.

See "Mission," University of Kansas, accessed June 22, 2017, https://ku.edu; "History & Mission," Johns Hopkins University, accessed June 22, 2017, https://www.jhu.edu/about/history/; "Stanford's Mission," Stanford University, accessed June 22, 2017, http://exploredegrees.stanford.edu/stanfordsmission/; "NIH Data Sharing Policy," National Institutes of Health, accessed March 2, 2018, https://grants.nih.gov/grants/policy/data_sharing/data_sharing_brochure.pdf; "Responsible Conduct of Research: Data Acquisition and Management," Columbia University, accessed May 11, 2017, http://ccnmtl.columbia.edu/projects/rcr/rcr_data/foundation/.

[4] Steven Johnson, *Where Good Ideas Come from: The Natural History of Innovation* (New York: Riverhead Books, 2010).

risk. Furthermore, if a malevolent group exploited a lethal disease or toxic threat research, novel bioagents could be produced against which our country has no protection. These are just a couple of the many possibilities that could result from the theft and use of research data for pernicious purposes. Access management and tracking must take priority among research universities and homeland security experts.

Addressing these concerns will require creating a novel collaborative scientific environment, whereby researchers and other academically minded individuals openly share and debate ideas and findings, where research is verified as unique or properly attributed prior to publication, and where every participant is vetted as trustworthy. Perhaps most importantly, this ideal environment will prevent censoring and corruption of ideas, data, and progress by any nation, state, or malicious individual. This ideal collaborative space would immediately benefit scientists and universities; moreover, if successful for academic purposes, this environment could expand to include private industry research and government laboratories. By examining existing technologies and identifying gaps within these technologies, this thesis offers a hypothetical solution to the ideal research sharing environment. Using lessons from what exist today combined with ideas for the technology of tomorrow, this thesis outlines new technologies for an open and trusted sharing environment where unique data and ideas can be traceably shared without fear of deletion. This thesis answers this question: How can research universities openly and with trust share verified unique data that is both immutable and ultimately traceable? This involves several processes:

1. Securely storing sensitive information online so that it is accessible only to authorized individuals.

2. Authorizing access to trusted parties with minimal risk of exposure.

3. Verifying authorship and tracking access to guarantee that sensitive information is not tampered with or plagiarized.

4. Preventing research information from being deleted prior to submission.

The author proposes a combination of key management, encryption, and validation to allow sharing of information and simultaneously prevent its distribution to unauthorized parties. The general outline of this solution is as follows.

A standard public-private key infrastructure (PKI) and Document Object Architecture is proposed for sharing documents among authorized parties while maintaining immutability and secure access. Proposed artificial intelligence techniques guarantee uniqueness and immutability of sensitive data and documents. A highly modified blockchain ledger similar to the X-Road is used for tracking and keeping records of who has had access in the past and present.

A PKI process similar to https currently employed by Internet browsers may be used to establish a trusted path between document owners and document users. Public keys are used to encrypt requests and private keys are used to decrypt documents stored as Digital Object Identifier (DOI) objects. In place of a centralized certificate authority, a distributed blockchain ledger and associated algorithms are used to track and manage access. The blockchain mechanism and novel beacon technology guarantees traceability and symmetric key encryption of documents guarantees security.

A system of smart contracts provides a PUT operation for authorized parties to add sensitive information to the system, and a GET operation for document retrieval and authorization. The smart contract blockchain is similar to the X-Road system employed by Estonia, but with significant differences:

- Anti-plagiarism verification is integrated into the PUT operation.
- The distributed ledger smart contract manages PKI, rather than a central authority.
- Documents are encrypted and assigned a DOI that resides in the ledger(s).
- A beacon is inserted in each GET operation, allowing for files to be tracked, and muted as necessary, after being downloaded.

Innovation does not occur in a vacuum. As Steven Johnson writes, “Good ideas may not want to be free, but they do want to connect, fuse, recombine. They want to reinvent themselves by crossing conceptual borders. They want to complete each other as

much as they want to compete."[5] Innovators must collaborate. The greatest minds in the world must be able to work together to solve the world's most daunting problems. Facilitating on-demand global intellectual summits or collaboration colliders will make the world a better place, if done correctly. Achieving this on a daily basis will require a new digital collaboration and sharing environment. This environment will allow research universities, openly and with trust, to share verified unique data that is both immutable and ultimately trackable. What are the next steps to make this environment a reality? First, examining currently technology's ability to meet the define needs. Second, evaluating the identified technologies against the ideal environment as defined by the thesis question. Third, proposing a solution that will meet the ideal environment. And finally, propose future projects to bring the environment from theory to reality.

Though this technology can help many different sectors, including the government and private industry, the ideal test market for this new technology is the academic research setting. Universities have a need to share information. For financial, legal, and prestige reasons, research universities are an ideal market for this new technology to succeed. In addition to being generators of invention and innovation, universities also have highly intelligent workforces and understand the value of open information sharing. As discussed previously in the problem statement, university research, when used as intended, has the potential to improve life via gene therapies and replacement organs, and increasing nutrition and food security globally. Maintaining the safety and security of sensitive and potentially dangerous information while sharing it productively requires better technology than exists today. As examined in this thesis, existing technologies cannot meet the needs of researchers collaborating globally today.

Though existing technologies cannot create an open, trusted sharing environment of verified unique data that is immutable and trackable, they can, however, provide a foundation from which to build new technology. The solution applications proposed above can hypothetically meet all the prescribed needs of the thesis question. Unfortunately, drawbacks also exist to the proposed technologies. Universities, researchers, and homeland

[5] Johnson, *Good Ideas*, 22.

security experts must pursue a solution, perhaps the one described in this thesis, to protect our universities' sensitive research data, our country's health from bioengineered diseases, and our nation's security from threats posed by maliciously misused research data.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

This thesis would not have been possible without the support of the University of Kansas Medical Center. I am especially indebted to Rick Johnson, associate vice chancellor and chief of police, and Steffani Webb, vice chancellor of administration, who have been supportive of my career goals and who worked actively to provide me with academic guidance and protected time to pursue my goals.

In addition, I would not have been able to pursue this master's program without the sponsorship, mentorship, and support of my dearest friends, Kelly Dunn, Matt May, and Erinn Blaz. Thanks to their advice and championship, I was accepted into the Center for Homeland Defense and Security. Furthermore, their continued care and dedication to my academic and mental wellbeing ensured my successful completion of this rigorous program.

I am grateful to all of those with whom I have had the pleasure to work during this and other related projects, most particularly, Kate Egerton and Chloe Woida of the Graduate Writing Center. I could not have completed this thesis without Kate and Chloe's encouragement, guidance, and patiently delivered lessons in academic writing.

This thesis would also not be possible without the extensive guidance of both Dr. Rodrigo Nieto-Gomez and Dr. Ted Lewis. As my teachers and mentors, they have taught me more than I could ever give them credit for here.

However, nobody has been more important to me in the pursuit of this program than the members of my family. I would like to thank my parents, Nancy and Darwin, whose love and strength are with me in whatever I pursue. I would also like to thank my husband's parents, James and Loretta, for providing support and encouragement, without which I would not have been able to finish this program. Most importantly, I wish to thank my amazingly supportive and devoted husband, Joshua, as well as my wonderful child, Alice, who provide unending motivation, inspiration, happiness, and love.

THIS PAGE INTENTIONALLY LEFT BLANK

I. PROTECTING SENSITIVE RESEARCH INFORMATION

Research universities need to share information, whether through teaching or disseminating key innovations to society. However, universities should not share all research with everyone.[1] If universities fail to protect sensitive data, it could harm citizens personally, financially, or potentially fatally if acquired by malicious actors.

The greater the number of researchers collaborating on a complex problem, like cancer, the less time it may take to find cure.[2] Information sharing is part of a university's mission and potentially a regulatory or legal obligation.[3] Because research is valuable, it needs to be propagated to reach its potential. As Steven Johnson writes, "The trick to having good ideas is not to sit around in glorious isolation and try to think big thoughts. The trick is to get more parts at the table."[4] Researchers must share the data "parts" on the global collaboration "table" to realize big ideas.

Scholars must balance open access with restricted access to sensitive information, when sharing university research data. Therefore, universities must track and secure sensitive data to prevent nefarious actors from stealing or weaponizing the information. For instance, if a malicious actor stole data generated from biodefense projects, United States service members and citizens along with other United States allies could be at great

---

[1] "Responsible Conduct of Research: Data Acquisition and Management," Columbia University, accessed May 11, 2017, http://ccnmtl.columbia.edu/projects/rcr/rcr_data/foundation/.

[2] Robert W. Rycroft, "Does Cooperation Absorb Complexity? Innovation Networks and the Speed and Spread of Complex Technological Innovation," *Technological Forecasting and Social Change* 74, no. 5 (June 1, 2007): 565–78, https://doi.org/10.1016/j.techfore.2006.10.005.

[3] To provide evidence for the assertion that information sharing is a part of the university's mission, I queried university mission statements cited here. The National Institutes of Health reference provides evidence for the regulatory and legal obligation of research universities to share information.

See "Mission," University of Kansas, accessed June 22, 2017, https://ku.edu; "History & Mission," Johns Hopkins University, accessed June 22, 2017, https://www.jhu.edu/about/history/; "Stanford's Mission," Stanford University, accessed June 22, 2017, http://exploredegrees.stanford.edu/stanfordsmission/; "NIH Data Sharing Policy," National Institutes of Health, accessed March 2, 2018, https://grants.nih.gov/grants/policy/data_sharing/data_sharing_brochure.pdf; Columbia University, "Responsible Conduct of Research."

[4] Steven Johnson, *Where Good Ideas Come from: The Natural History of Innovation* (New York: Riverhead Books, 2010).

risk. Furthermore, if a malevolent group exploited a lethal disease or toxic threat research, novel bioagents could be produced against which our country has no protection. These are just a couple of the many possibilities that could result from the theft and use of research data for pernicious purposes. Access management and tracking must take priority among research universities and homeland security experts.

Addressing these concerns will require creating a novel collaborative scientific environment, whereby researchers and other academically minded individuals openly share and debate ideas and findings, where research is verified as unique or properly attributed prior to publication, and where every participant is vetted as trustworthy. Perhaps most importantly, this ideal environment will prevent censoring and corruption of ideas, data, and progress by any nation, state, or malicious individual. This ideal collaborative space would immediately benefit scientists and universities; moreover, if successful for academic purposes, this environment could expand to include private industry research and government laboratories. By examining existing technologies and identifying gaps within these technologies, this thesis offers a hypothetical solution to the ideal research sharing environment. Using lessons from what exit today combined with ideas for the technology of tomorrow, this thesis outlines new technologies for an open and trusted sharing environment where unique data and ideas can be traceably shared without fear of deletion.

A. WHY IT MATTERS TO HOMELAND SECURITY ENTERPRISES

In the *National Strategy to Secure Cyberspace*, universities are listed as critical infrastructure for national cybersecurity.[5] The document lists universities as critical infrastructure because "[Institutes of higher education] are subject to exploitation for two reasons: (1) they possess vast amounts of computing power; and (2) they allow relatively open access to those resources…many [university networks contain] research and significant central computing facilities."[6] Additionally, the types of information that

[5] Department of Homeland Security, *The National Strategy to Secure Cyberspace* (Washington, DC: Department of Homeland Security, 2003), xiii, https://www.us-cert.gov/sites/default/files/publications/cyberspace_strategy.pdf.

[6] Department of Homeland Security, 40.

universities store, specifically those relating to personal identifiable information and gene sequencing and modification, are of national concern.

Information is valuable and at universities it is like air; even though a person may not be able to see it, it is everywhere. Information saturation makes universities a very attractive target for malicious actors looking to leverage stolen information for monetary profit. For instance, medical records are extremely valuable. On the black market, patient health records are worth as much as $363 U.S. each, and the average price for a partial health record is $50 U.S. each, according to the Federal Bureau of Investigation (FBI).[7] According to that same FBI report, the reason these records are so valuable is that they can be used to "file fraudulent insurance claims, obtain prescription medication, and advanced identity theft."[8] Though most universities do not store many health records, unless they are a medical center, all universities do have student enrollment records and human resource files that contain many of the same valuable pieces of information, including social security numbers, dates of birth, addresses, and financial data.[9] This information, if stolen and sold on dark websites, has the power to affect the financial stability of affected individuals, credit issuers, and health insurance companies.

In addition to personally identifiable information, university networks also store research data that is valuable to malicious actors. One example is, research that involves genetic modification of plants and animals. According to the Defense Advanced Research Projects Agency (DARPA) website,

> From a national security perspective, [there are] inherent risks that arise from the rapid democratization of gene editing tools. The steep drop in the costs of genomic sequencing and gene editing toolkits, along with the increasing accessibility of this technology, translates into greater opportunity to experiment with genetic modifications. This convergence of

---

[7] Ashiq JA, "Hackers Selling Healthcare Data in the Black Market," Infosec Institute, July 27, 2015, http://resources.infosecinstitute.com/hackers-selling-healthcare-data-in-the-black-market/; "Health Care Systems and Medical Devices at Risk for Increased Cyber Intrusions for Financial Gains," Federal Bureau of Investigation (FBI), April 8, 2014, http://www.aha.org/content/14/140408--fbipin-healthsyscyberintr ud.pdf.

[8] FBI, "Health Care Systems."

[9] "Data Breaches," Privacy Rights Clearinghouse, accessed October 30, 2016, https://www.privacy rights.org/data-breaches.

> low cost and high availability means that applications for gene editing—both positive and negative- could arise from people or states operating outside of the traditional scientific community.[10]

For example, the University of Missouri has eleven unique agricultural research laboratories. According to its website, the University of Missouri's College of Agriculture, Food, and Natural Resources' "innovative research spans the globe, often taking a high-tech look at traditional agriculture, food and natural resource issues, and directly impacts our future, from health breakthroughs, to sustainable agriculture techniques to food safety."[11] Specifically, these laboratories study animal health, pest and weed control, alternative crops, wastewater management, and natural resource measurement.[12] These topics may seem benign; however, if a malicious actor gained unrestricted access to this research, he or she could, with the right knowledge and equipment, use this same information to create a biological weapon against our homeland.

According to the University of Missouri News Bureau, researchers at the University of Missouri have genetically engineered pigs that will accept any stem cell transplant or graft given to them.[13] This is a ground breaking discovery that can lead to porcine-generated cures for human diseases in the near future because pigs are very anatomically similar to humans.[14] However, should one of these immunocompromised pigs, or the information for how to create them, be stolen and used by a malicious actor with means, that actor would have the ability to use the pigs to replicate and disperse diseases that could be more virulent than the plague. Take, for instance, the initial H1N1 influenza pandemic of 1918. Though malicious actors did not initiate this pandemic, it does give an example of how pigs can easily spread novel diseases to humans with disastrous outcomes.

---

[10] "Setting a Safe Course for Gene Editing Research," Defense Advanced Research Projects Agency (DARPA), September 7, 2016, https://www.darpa.mil/news-events/2016-09-07.

[11] "Research," University of Missouri College of Agriculture, Food & Natural Resources, accessed November 7, 2016, https://cafnr.missouri.edu/research/.

[12] "CAFNR Research Centers," University of Missouri College of Agriculture, Food & Natural Resources, accessed November 7, 2016, https://cafnr.missouri.edu/research/centers/.

[13] Nathan Hurst, "MU Scientists Successfully Transplant, Grow Stem Cells in Pigs," *MU News Bureau*, June 4, 2014.

[14] Hurst.

According to the Centers for Disease Control and Prevention and Anhlan et al., the pandemic of 1918 killed between 50 and 100 million people, and was caused by birds transmitting avian flu to pigs, that mutated the virus and transmitted it to humans.[15] Similarly, if a malicious actor inserted a highly virulent mutated virus strain into the University of Missouri research pigs, the pigs could be used to replicate and unknowingly transmit the virus to anyone with whom they had contact.

Though this example may seem far-fetched, genetically engineered material is much more accessible today than it was even a few years ago. The University of California Berkeley's Innovation website reports that there are a total of 1,530 active inventions right now.[16] Berkeley has invented medications that provide the university with royalties monetized for $87.5 million that currently help fund biological research facilities among other investments.[17] One of the university's biological research success stories, as reported by Wallace Ravven with Berkeley Research, is the "'molecular scissors' approach, known as CRISPR/Cas9, [which] can be used with great precision to selectively disable or add several genes at once to human cells."[18] This discovery by biochemist Jennifer Doudna has made creating and editing genetic material easier than ever before. Created to help cure bloodborne diseases *ex vivo*, this technology could help to treat sickle cell anemia and human immunodeficiency virus in the future.[19]

Similarly, according to the University of Nebraska Medical Center's *Newsroom*, the University of Nebraska Medical Center is performing ground- breaking research

---

[15] "2009 HiNi Pandemic (Hinipdm09 Virus)," Centers for Disease Control and Prevention, last updated November 2, 2017, http://www.cdc.gov/flu/pandemic-resources/basics/past-pandemics.html; Darisuren Anhlan et al., "Origin of the 1918 Pandemic H1N1 Influenza A Virus as Studied by Codon Usage Patterns and Phylogenetic Analysis," *RNA* 17, no. 1 (2011): 64, http://doi.org/10.126/ma.2395211.

[16] "Innovation & Entrepreneurship," University of California at Berkeley, accessed November 7, 2016, http://vcresearch.berkeley.edu/excellence/innovation-and-entrepreneurship.

[17] University of California at Berkeley.

[18] Wallace Ravven, "Crispr Goes Global," University of California Berkeley Research, March 18, 2014, https://vcresearch.berkeley.edu/news/profile/doudna_jennifer.

[19] Ravven.

creating a vaccine for the Ebola virus that has been ravaging Africa.[20] This same article goes on to describe how the Department of Defense is sponsoring this research because the Ebola virus has been classified as a class A bioterrorism agent, killing nearly 90 percent of those infected with the virus.[21] The intended purpose of this research is to create a vaccine for United States soldiers; however, once sufficient amounts of the vaccine have been made and distributed, the vaccine would likely become available to civilians who live, work, and travel to Ebola-stricken regions.[22] In this research, according to the *Newsroom* article, scientists have combined aluminum salt with a virus-like particle that mimics a vulnerable piece of genetic code within the Ebola virus.[23] To identify the critical genetic piece of the virus, significant research must have been performed and documented, identifying large portions of the genetic material of the Ebola virus. If a malicious actor breached the University of Nebraska Medical Center's network and obtained that information, they could potentially use it for great harm.

Continuing this scenario, a malicious actor could use available techniques to kill millions worldwide. If a malicious actor with significant scientific training used the CRISPR technique to splice a mutation into the Ebola virus to change the key piece of the anti-body defense genetic code then use an immunocompromised pig to replicate or deliver the virus, anyone infected with the virus would likely die. The probability of this happening is very slim, given the multiple points of failure within the provided scenario; however, these are just a few examples of research done in university laboratories across the country that could cause significant harm to our nation if a malicious actor gained access to university data.

As another hypothetical example of research data being used for malicious intent, if data on crop research were misappropriated, it could be used to decimate the United States' food supply. Kansas State University's Agronomy Department performs genetic

---

[20] Tim O'Connor, "UNMC Research Team Working on Vaccine for Ebola Virus," University of Nebraska Medical Center Newsroom, February 13, 2015, https://www.unmc.edu/news.cfm?match=16495.

[21] O'Connor.

[22] O'Connor.

[23] O'Connor.

modification to staple agricultural products in order to enhance them for drought tolerance, thriving in nutrient poor soil, and pest resistance.[24] These programs aim to help the United States and other countries internationally achieve food security.[25] However, the information collected from these research projects also detail gene modifications and outcomes that were not successful.[26] A malicious actor exploiting a failed experiment's genetic modification instructions could mass produce and sell seeds that would not fruit or mature, decimating America's agrarian economy. In 2015, agriculture and related products accounted for $992 billion of the United States' gross domestic product (GDP), accounting for 5.5% of the overall GDP.[27] Though this type of a malicious attack is hypothetical, the damage from such an attack would be devastating to the nation; decimating the nation's agrarian economy, disrupting the nation's food supply, and threatening the nation's security.

A similar situation of counterfeit product entering the market via online sales happened in 2017. Amazon sold counterfeit solar eclipse glasses to thousands of people across the United States prior to the 2017 solar eclipse.[28] Though Amazon is a trusted online retailer, the Amazon sub-distributors misrepresented the counterfeit solar eclipse viewing glasses and sold them to the American masses.[29] It is plausible that a malignant counterfeit seed manufacturer could do something similar that would impact the American economy on an even broader scale than the Amazon solar eclipse glasses.

---

[24] "Plant Breeding & Genetics Research Areas," Kansas State University Department of Agronomy, December 8, 2015, http://www.agronomy.k-state.edu/research/plant-breeding-and-genetics/index.html.

[25] Kansas State University Department of Agronomy.

[26] "Research Reports," Kansas State University Department of Agronomy, February 15, 2017, https://www.agronomy.k-state.edu/research/research-reports/.

[27] "Ag and Food Sectors and the Economy," United States Department of Agriculture Economic Research Service, October 18, 2017, https://www.ers.usda.gov/data-products/ag-and-food-statistics-charting-the-essentials/ag-and-food-sectors-and-the-economy/.

[28] Nicole Pelletiere, "Amazon Issues Refunds to Customers Who Bought Counterfeit Solar Eclipse Glasses," ABC News, August 14, 2017, http://abcnews.go.com/US/amazon-issues-refunds-customers-bought-counterfeit-solar-eclipse/story?id=49206091.

[29] Pelletiere.

DARPA is also concerned about this type of possibility. In September of 2016, DARPA announced the Safe Genes program. According to the DARPA website, "Safe Genes program aims to build a biosafety and biosecurity toolkit to reduce potential risks and encourage innovation in the field of genome editing."[30] The three objectives of the program are to create reversible control of gene editors, develop countermeasures for prophylactic and treatment of wild-type organisms to protect from mutations, and design a method to remove unwanted genes from an environment.[31] Though these objectives will help national security, they do not address keeping gene editing data safe.

For the purposes of this thesis, sensitive information is defined as any document or data that may be used for harm. Such data must be simultaneously made available to researchers to advance science, while being denied to malicious actors. Striking a balance between access and denial of access is the objective of this research.

B. RESEARCH QUESTION

This thesis answers the question: How can research universities openly and with trust share verified unique data that is both immutable and ultimately traceable? This involves four processes:

1. Securely storing sensitive information online so that it is accessible only to authorized individuals,
2. Authorizing access to trusted parties with minimal risk of exposure,
3. Verifying authorship and tracking access to guarantee that sensitive information is not tampered with or plagiarized,
4. Preventing research information from being deleted prior to submission.

[30] DARPA, "Setting a Safe Course for Gene Editing Research."

[31] DARPA.

The author proposes a combination of key management, encryption, and validation to allow sharing of information and simultaneously preventing its distribution to un-authorized parties. The general outline of this solution is as follows.

A standard public-private key infrastructure (PKI) and Document Object Architecture is proposed for sharing documents among authorized parties while maintaining immutability and secure access. Proposed artificial intelligence techniques guarantee uniqueness and immutability of sensitive data and documents. A highly modified blockchain ledger similar to the X-Road is used for tracking and keeping records of who has had access in the past and present.

A PKI process similar to https currently employed by Internet browsers may be used to establish a trusted path between document owners and document users. Public keys are used to encrypt requests and private keys used to decrypt documents stored as Digital Object Identifier (DOI) objects. In place of a centralized certificate authority, a distributed blockchain ledger and associated algorithms are used to track and manage access. The blockchain mechanism and novel beacon technology guarantees traceability and symmetric key encryption of documents guarantees security.

A system of smart contracts provides a PUT operation for authorized parties to add sensitive information to the system, and a GET operation for document retrieval and authorization. The smart contract blockchain is similar to the X-Road system employed by Estonia, but with significant differences:

- Anti-plagiarism verification is integrated into the PUT operation
- The distributed ledger smart contract manages PKI, rather than a central authority
- Documents are encrypted and assigned a DOI that resides in the ledger(s)
- A beacon is inserted in each GET operation, allowing for files to be tracked, and muted as necessary, after being downloaded

C. RESEARCH DESIGN

To answer the research question, I compared representative samples of current technologies used by university faculty for information sharing. This exploration included data sharing, uniqueness trust, immutability, verification, identification, and traceability. I explored university websites, trade journals, and blogs from the technology community to determine the most applicable technologies for my course of inquiry. A minimum of three technologies for each sub-topic has been explored to reduce sample bias. Once I identified key strengths and weaknesses of each available technology, I used those data to determine whether the technology meets the needs identified within the research question. Data for the purposes of this thesis will be defined as data and information generated from research, key research findings, and publications. Unfortunately, the limitations of current technologies prevent exploration of raw and granular research data because these technologies are limited to publish works only. With that limitation defined, moving forward in this thesis the reference to data in the technology evaluations will refer only to university research publications.

1. Data Sharing

A key piece to grasping research universities' current sharing and securing challenges is to explore technologies used for these purposes. To explore this topic, I gathered information on current research sharing and storage solutions including popular research collaboration and cloud storage websites. I also analyzed the advantages and disadvantages of each type of website, as well as the website's ability to fulfill prescribed parameters.

2. Data Trust and Immutability

Data trust and immutability are hard problems to solve. Emerging technologies make trust and immutability a possibility through algorithmic encryption and decentralized data storage. Through reading scholarly articles, trade journals, and technology community blogs, I determined what technologies currently exist and are emerging to address this problem. Once identified, I compared each technology to the parameters defined within the research question.

3. Data Uniqueness Verification

Determining the uniqueness of documented research becomes increasingly challenging as more and more work is published. However, verifying the uniqueness of research prior to publication has never been more important. Plagiarism has far-reaching consequences. For example, according to Martin Enserink, physicist Etienne Klein is accused of plagiarizing the works of many scientists and other experts in his many published papers.[32] As a result Klein is reportedly losing his appointment to the Institute for Advanced Studies for Science and Technology in Paris, France.[33] Fortunately, many tools exist to help detect plagiarism, including Turnitin, iThenticate, and Google search engine.[34] To better understand these tools and these technology's ability to detect uniqueness of data, I examined how each of these products work, including the benefits and drawbacks and compare them to the parameters defined within the research question.

4. Data Identification and Traceability

To explore data identification and tracking, I studied current document tracking systems and methods. Exploring historical and current document identification and tracking methods allowed me to assess the advantages and disadvantages of each type of technology. I gathered information from technology manuals and technology source websites for my analysis.

D. LITERATURE REVIEW

To better understand the dichotomies universities face regarding research sharing, collaboration, and innovation, it is important to recognize the fundamental struggle of sharing versus securing university research information. These opposing needs directly

---

[32] Martin Enserink, "French Physicist Accused of Plagiarism Seems Set to Lose Prestigious Job," *Science*, April 6, 2017, http://www.sciencemag.org/news/2017/04/french-physicist-accused-plagiarism-seems-set-lose-prestigious-job.

[33] Enserink.

[34] "Technology to Improve Student Writing," Turnitin, accessed June 18, 2017, http://turnitin.com/; "Prevent Plagiarism in Published Works," iThenticate, accessed June 18, 2017, http://www.ithenticate.com; "Google Search Help," accessed February 21, 2018, https://support.google.com/websearch/?hl=en#topic=3036132.

impact the ability of university research scientists to globally collaborate and innovate. Furthermore, universities must share enough information to facilitate research innovation without endangering the reputation or security of the institution. This literature review explores the research that exists for sharing versus securing information in higher education institutions.

1. Sharing Information

In today's knowledge-driven society, universities must share research to remain relevant. According to a study by Ming-Yu Cheng, Jessica Sze-Yin Ho, and Pei Mey Lau, "once created, knowledge needs to be distributed quickly and widely because active knowledge is the 'gem' while idle knowledge is the 'stone.'"[35] In other words, knowledge is only valuable if it is shared and appreciated immediately following its creation. Even potentially sensitive data needs to be shared. As J. Robert Oppenheimer said,

> the trouble with secrecy is that it denies to the government itself the wisdom and the resources of the whole community, and the whole country, and the only way you can do this is to let almost anyone say what he thinks, and to let men deny what they think is false, argue what they think is false. You have to have a free and uncorrupted communication.[36]

To sequester knowledge is a burdensome weight preventing forward progress. This is particularly true for universities. As organizations, universities increase its own status and prestige by sharing and publishing relevant data. An article by David Wiley further confirms this saying that universities must be willing to share information to remain relevant.[37] This elevation of status can result in economic gains for the university by way of private donations and additional grant funds.[38] Additionally, broad information sharing

---

[35] Ming-Yu Cheng, Jessica Sze-Yin Ho, and Pei Mey Lau, "Knowledge Sharing in Academic Institutions: A Study of Multimedia University Malaysia," *Electronic Journal of Knowledge Management* 7 (2009): 313–24.

[36] J. Robert Oppenheimer, "J. Robert Oppenheimer on Government Secrecy," History.com video, accessed August 3, 2017, http://www.history.com/topics/world-war-ii/world-war-ii-history/videos/j-robert-oppenheimer-on-government-secrecy.

[37] David Wiley, "Open Source, Openness, and Higher Education," *Innovate: Journal of Online Education* 3, no. 1 (October 2006), https://www.learntechlib.org/p/104321/.

[38] Cheng, Ho, and Lau, "Knowledge Sharing in Academic Institutions."

allows new researchers to use this information to expand on existing topics without having to repeat work as they enter the field.[39]

For researchers, sharing information allows for collaboration between different groups and organizations. According to the National Institutes of Health, information sharing results in more efficient use of resources. It also decreases the amount of time required for foundational research leading up to discovery by reducing the amount of repetitive work.[40] In addition to improving the time to discovery, collaborating, publishing, and sharing information can also lead to more sound science. The University of Alaska Fairbanks Publication and Peer Review website describes the necessity of peer review for publications and findings to verify that the research is sound.[41] University and faculty benefit from information sharing. In many organizations of higher education, contributing a major discovery to the field is how one secures a promotion and potentially tenured status.[42] These faculty works are only valuable if shared and validated by a peer review group.

Each university, as an educational institution, has a mission to educate those willing to learn. This includes sharing information generated within the organization with the larger community so that others can learn from what the university has accomplished.[43] Universities, researchers, and faculty all benefit from sharing information. Yet, when is it prudent to share information? The National Institutes of Health state that data sharing is without question the most important outcome of any research.[44] The National Institutes of

---

[39] NIH, "Data Sharing Policy."

[40] NIH.

[41] "Publication & Peer Review," University of Alaska Fairbanks, August 25, 2015, http://www.uaf.edu/ori/responsible-conduct/peer-review/.

[42] Thomas R. McDaniel, "Rethinking Scholarly Publication for Tenure," in *Faculty Promotion and Tenure: Eight Ways to Improve the Tenure Review Process at Your Institution*, 13–14 (Madison, WI: Magna, 2012), http://www.jsums.edu/academicaffairs/files/2012/08/Tenure-and-Promotion.pdf?x19771; Timothy J. Fogarty and Donald V. Saftner, "Academic Department Prestige: A New Measure Based on the Doctoral Student Labor Market," *Research in Higher Education* 34, no. 4 (August 1, 1993): 427–49, https://doi.org/10.1007/BF00991853.

[43] NIH, "Data Sharing Policy."

[44] NIH.

Health, along with many other grant-providing entities, encourages sharing of information to verify the funded research has indeed borne fruit.[45] Sharing information from research also reduces redundant investment for very expensive, rare, or specifically timed resources or events.[46] Articles advocating for open sharing also note that there must be a balance between knowledge sharing, and information security. To maintain viability, universities must clearly designate which types of information are advantageous to share versus which types of information are better to retain within the organization.

2. Information Security

The field of information security focuses on network architecture and protection tools rather than guidance for information dissemination with an audience outside the university. Few articles or white papers broadly outline how to classify information, and none provide any specific details on how to protect and share information generated at universities. The majority of documents examined here were from academic articles, government documents, industry white papers, and university policies.

Publication and collaboration require researchers to share information online. However, this presents challenges for university information security. Quey-Jen Yeh and Arthur Jung-Ting Chang, assert that information security is protection against unauthorized access, adulteration, or other misuse of information.[47] Often information security conflicts with information sharing. A Universities UK report describes how universities must manage the risk associated with each type of information to maintain adequate information security.[48] Many large universities have embraced this concept and have developed information security and privacy policies. Harvard's Research Data Security Policy states

---

[45] NIH.

[46] NIH.

[47] Quey-Jen Yeh and Arthur Jung-Ting Chang, "Threats and Countermeasures for Information System Security: A Cross-Industry Study," *Information & Management* 44, no. 5 (July 2007): 480–91, https://doi.org/10.1016/j.im.2007.05.003.

[48] Universities UK, *Cyber Security and Universities: Managing the Risk* (London: Universities UK, 2013), http://www.universitiesuk.ac.uk/policy-and-analysis/reports/Documents/2013/cyber-security-and-universities.pdf.

that research confidentiality is based on laws, regulations, and university policies.[49] Another supplemental policy also binds researchers, faculty, students, and staff to contractual obligations for the protection of information deemed confidential.[50] The terms outlined in these policies are very broad, yet similar blanket policies exist for other institutions. Stanford, for example, offers similar advice to protect data, and mentions that financial and ethical reasons also may qualify information as sensitive or protected.[51] However, as a foil to Harvard's policy, Stanford's Information Security Policy mentions that the balance between information security and information sharing is necessary to further academic objectives.[52] Though a small sample, these two universities exemplify the developing culture of information security at many universities.

The scope of this thesis assumes classification has already been performed and focuses on actions taken to protect, distribute, track, and retrieve information that has previously been classified as sensitive. Such information is treated differently through processes described below.

Other challenges information security professionals face beyond data classification include direct and indirect systems attacks and legal consequences of a data breach. According to Randy Marchany and Trend Micro posit that threats to information security include "exploits against internal database systems and servers, malware delivered to staff endpoints via a variety of vectors, exploits against websites or servers, and phishing attacks."[53] In addition to the many attacks that university intelligence infrastructure is expected to withstand, universities also face regulations and laws with very severe consequence should a breach occur. Some of the more prominent laws and regulations

---

[49] "Harvard Research Data Security Policy," Harvard University, accessed March 8, 2017, http://vpr.harvard.edu/pages/harvard-research-data-security-policy.

[50] "Information Security Policy," Harvard University, accessed March 8, 2017, http://policy.security.harvard.edu/home.

[51] "Information Sharing," Stanford University, accessed March 8, 2017, http://financialaid.stanford.edu/undergrad/policy/sharing.html.

[52] Stanford University.

[53] Randy Marchany, *Higher Education: Open and Secure?* (North Bethesda, MD: SANS Institute, 2014), 7, https://jp.trendmicro.com/cloud-content/us/pdfs/business/articles/sans_higher_education_open_and_secure_research_study_trend_micro_edition_final.pdf.

include Family Educational Rights and Privacy Act (FERPA), Health Insurance Portability and Accountability Act of 1996 (HIPAA), and the Payment Card Industry Data Security Standard.[54] These are in addition to the broad information management and handling statutes and regulations that cities, states, and the federal government impose. The combined pressure from the threats and the regulations is enough to motivate the over-classification and protection of data generated by university researchers, faculty, and staff.

Once again, the proposed solution conforms to technical and legal protections without loss of generality. It is assumed document classification is done according to the laws and any system of collection and dissemination must conform to the law and be technically secure.

The majority of university information security policies do not include the process for data redaction for research findings that include sensitive information. The National Institutes for Health define data redaction as the process of removing sensitive information from a dataset to prepare the larger conceptual idea for release without compromising any protected information.[55] For example, when a doctor is studying patients for disease trends, the doctor can disclose the trends without releasing any protected patient information. However, not all studies are as straightforward as this example, including studies involving patents or ethical dilemmas.

The proposed solution also assumes there is no loss of confidentiality due to data aggregation even though aggregation may occur outside of the proposed solution.[56] The use of PKI and blockchain technology does not address the important problem of aggregation that may result in escalation of classification.

---

[54] "Family Educational Rights and Privacy Act (FERPA)," Department of Education, June 26, 2015, https://ed.gov/policy/gen/guid/fpco/ferpa/index.html; "Health Insurance Portability and Accountability Act of 1996," Department of Health and Human Services, November 23, 2015, https://aspe.hhs.gov/report/health-insurance-portability-and-accountability-act-1996; PCI Security Standards Council, *PCI DSS Quick Reference Guide: Understanding the Payment Card Industry Data Security Standard Version 2.0* (Wakefield, MA: PCI Security Standards Council, 2010), https://www.pcisecuritystandards.org/documents/PCI%20SSC%20Quick%20Reference%20Guide.pdf.

[55] NIH, "Data Sharing Policy."

[56] Julia Lane et al., *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (New York: Cambridge University Press, 2014), https://doi.org/10.1017/CBO9781107590205.

Unfortunately, information security in the digital age is a very young field. Many recent articles address the need for additional staffing so that a more balanced approach for data management and sharing can be achieved.[57] Even with this basic awareness of additional staffing needs, it is unknown how the developing information security field will balance security with the need for greater sharing in the future.

[57] Though many articles exist detailing the shortages of information technology and other computer professionals in the United States, these two articles were chosen as representative samples: Robert Half Technology, *2018 Salary Guide for Technology Professionals* (Menlo Park, CA: Robert Half Technology, 2017), https://www.roberthalf.com/sites/default/files/documents/2018_salary_guide_NA_technology_1.pdf; Stella Fayer, Alan Lacey, and Audrey Watson, "STEM Occupations: Past, Present, and Future," Bureau of Labor Statistics, January 2017, https://www.bls.gov/spotlight/2017/science-technology-engineering-and-mathematics-stem-occupations-past-present-and-future/home.htm.

THIS PAGE INTENTIONALLY LEFT BLANK

II. DATA SHARING

Many websites exist to facilitate idea exchange and collaboration worldwide via publications, white papers, or other articles. ResearchGate, Academia.edu, and arXiv are a few prominent sites in scientific communities.[58] This chapter evaluates these information sharing sites' ability to answer the research question. Though each of these sites provides an avenue to share ideas, there are also drawbacks to each.

A. RESEARCHGATE

ResearchGate, founded in 2008, is a German-based, privately owned website that allows anyone to register to use the service.[59] Though designed toward academics, the site is geared to be a multi-disciplined forum. There are four separate options on the site's registration page: academic or student; corporate, government, or non-governmental organization; medical; or not a researcher, to include journalists, citizen scientists, or anyone interested in reading and discovering research.[60] The sign up page encourages individual registrants to use his or her institutional or business email to sign up for an account, to share papers and publish scientific findings. To verify identity, the site sends a confirmation email to the user's account. The site also allows each user to comment on publications, sign up to follow other users' work, follow work of other peers of a common organization, and even offers suggestions for articles users may want to read. This open and searchable forum allows for collaborations between users who otherwise may not have discovered one another.[61]

---

[58] ResearchGate, accessed May 30, 2017, https://www.researchgate.net/; "Academia.edu," accessed June 5, 2017, https://www.academia.edu/user_unflag_requests/new; "e-Print Archive," arXiv.org accessed October 14, 2017, https://arxiv.org/.

[59] Ingrid Lunden, "ResearchGate Raises $52.6M for Its Social Research Network for Scientists," *TechCrunch* (blog), accessed October 14, 2017, http://social.techcrunch.com/2017/02/28/researchgate-raises-52-6m-for-its-social-research-network-for-scientists/; "Sign up," ResearchGate, accessed October 14, 2017, https://www.researchgate.net/signup.SignUp.html?ev=su_chnl_index&_sg=TE_HpXW7FB7IPBOi2ybu0a9XjiwDB4PPW0WPzX4qr-nbP9JwcoFoqsCOY9BNZk9Ry6XrOzCjKhiL.

[60] ResearchGate, "Sign up."

[61] ResearchGate "Share and Discover Research."

Though ResearchGate is free to use, the website generates funding by targeting advertisements for lab materials and jobs to users. Currently, ResearchGate has more than 13 million users and 100 million publications from 193 countries worldwide.[62] The company has also been successful in generating over $100 million in venture capital.[63] As of February 2017, ResearchGate is publishing approximately 2.5 million articles each month.[64] A great success of ResearchGate is the ability to measure the impact and reputation of each publication and researcher. Each publication is tracked for the quantity and identity of readers, citations, and recommendations.[65] These metrics are then combined to create the "scientific reputation."[66] The ResearchGate "scientific reputation" is a tool that can help publication consumers better assess the credibility and expertise of an author. This feature further establishes trust within the ResearchGate community.

B. ACADEMIA.EDU

Another academic publication and collaboration website is Academia.edu. Academia.edu is a San Francisco-based privately owned social networking website designed for information sharing. Created prior to 2001, Academia.edu was created with an .edu domain name despite having no affiliation to a United States institute of higher education.[67] Since its creation, Academia.edu has garnered 56 million users, 19 million publications, and $17.7 million in venture capital.[68] Like ResearchGate, any Academia.edu user can create an account by entering a name, email address, and organization.[69] To better categorize each user, the registration page offers the following options for new registrants:

---

62 "About Us," ResearchGate, accessed February 14, 2018, https://www.researchgate.net/about.

63 ResearchGate.

64 Lunden, "ResearchGate Raises $52.6M."

65 "Stats and Scores," ResearchGate, accessed October 14, 2017, https://explore.researchgate.net/display/support/Stats+and+Scores.

66 ResearchGate

67 Sarah Bond, "Dear Scholars, Delete Your Account at Academia.Edu," Forbes, accessed February 14, 2018, https://www.forbes.com/sites/drsarahbond/2017/01/23/dear-scholars-delete-your-account-at-academia-edu/.

68 "About," Academia.edu, accessed October 14, 2017, http://www.academia.edu/about.

69 Academia.Edu, "Share Research."

professor, alum, graduate student, undergraduate, or other. To verify identity, the site sends a confirmation email to the user's email account. However, Academia.edu's registration site also offers opportunities to connect one's Academia.edu account to other social media platforms, further expanding opportunities for collaboration while also confirming the identity of the user.

When publishing on Academia.edu, each user certifies that they are the owner of all submissions posted to the site.[70] However, the site does not offer any duplication or attribution scans. If there is a conflict with information published on the site, users must report the error to site administrators for further assessment.[71] However, when searching the "Help Center" on the site there are no results for plagiarism or duplicate publications.

Publications on Academia.edu, like ResearchGate, are tracked for statistical purposes.[72] However, unlike ResearchGate, Academia.edu users must pay to have access analytics gathered on individuals viewing and citing their papers. In addition to analytics, the premium subscription provides users a personalized website, and provides a search engine that will search the body of all published articles not just the title and publisher-supplied metadata. Each Academia.edu account has the ability to publish, edit, and delete any work submitted.[73] Going through a simple process, each account retains the ability to change or remove any work previously submitted.[74] Though the ability to remove mistakes or incorrect information is important, the ability to easily edit or remove a publication could be exploited by individuals or nation-states who would want to censor information regardless of its usefulness or potential.

---

[70] "Terms," Academia.edu, accessed May 29, 2017, https://www.academia.edu/terms.

[71] "Reporting a Fake, Offensive, or Spam Profile," Academia.edu, accessed October 14, 2017, http://support.academia.edu/customer/en/portal/articles/1734342-reporting-a-fake-offensive-or-spam-profile.

[72] "What Is Academia Premium?," Academia.edu, accessed February 14, 2018, http://support.academia.edu/customer/en/portal/articles/2405880-what-is-academia-premium-.

[73] "Deleting Your Paper," Academia.edu, accessed October 14, 2017, http://support.academia.edu/customer/portal/articles/2250705.

[74] Academia.edu.

C. arXiv.ORG

Lastly, Cornell University Library's arXiv.org is another academic information sharing website. Unlike ResearchGate and Academia.edu, arXiv.org is associated with Cornell University and funded by the Simons Foundation and other member institutions.[75] Also unlike ResearchGate and Academia.edu, resources on arXiv.org are openly and freely available to any visitor to the site without creating a user account or providing any personally identifiable information.[76] However, to publish on the site a user must register providing a name, username and password, email address, affiliation, career status, and research category of interest.[77] Once the request is submitted, users must activate the account with a verification code or link provided in the verification email. Once complete any user can begin to upload publications. In addition to the user-provided name, arXiv.org request that all publishers include their open researcher and contributor identification number (ORCiD) member number so that all associated works, regardless of platform, can be attributed to the publisher. An open researcher and contributor identification, or ORCiD, is a non-proprietary alphanumeric string used to uniquely identify scientific and other academic authors and contributors. This furthers the works available for collaboration while also verifying the identity of the publisher to establish trust.[78]

In addition to the ORCiD number, each publication must also have an arXiv identifier.[79] The site continues to explain that the identifier is made from the month and year of publication followed by a four or five digit number representing the numerical order of publication, and if it is a document that will be versioned, that number is to be included as well: arXiv:yymm.numbervV, or 1710.0013v1 as an example of the first version of the 13th document published in October of 2017. The arXiv identifier allows the document to

---

[75] "arXiv Member Institutions (2017)," Cornell University, accessed October 14, 2017, https://confluence.cornell.edu/pages/viewpage.action?pageId=340900096.

[76] arXiv.org, "e-Print Archive."

[77] "Your arXiv.org Account," accessed October 14, 2017, https://arxiv.org/user/.

[78] "ORCID Identifiers," arXiv, accessed October 14, 2017, https://arxiv.org/help/orcid.

[79] "Understanding the arXiv Identifier," arXiv, accessed October 14, 2017, https://arxiv.org/help/arxiv_identifier.

be more easily traced regardless of potential relocation or title change.[80] Another deviation from the ResearchGate and Academia.edu models is that arXiv.org documents cannot be completely deleted. A formal request to withdraw an article, once published, will result in the title and abstract still being available although the paper itself will be withdrawn from public access.[81] Document removal requires approval from the site administrator before it becomes official, and all previous versions of the paper will still be publicly accessible.[82] However, articles can be replaced or updated by simply creating and uploading a new version.[83]

Another deviation from ResearchGate and Academia.edu is arXiv.org's review process. arXiv.org provides moderation for each submission.[84] This is not to be confused with a peer-review process, it is simply to verify that all publication submissions are unique, properly categorized, and appropriate for the specific scientific community that arXiv.org serves.[85]

Each of these research information sharing websites has distinct advantages and disadvantages. Though all provide a global stage to share information, none allow for complete immutability, or uniqueness verification. ResearchGate provides a community that self-polices and establishes end-user trust, and an algorithm to track each publication's downloads.[86] ResearchGate makes download statistics available to the publisher and end users at no cost which further provides academic relevance and credibility scores for each submission. Academia.org also tracks publication downloads and has a self-policing community; however, access to credibility and academic impact scores require further

---

80 arXiv.

81 "To Withdraw an Article," arXiv, accessed October 14, 2017, https://arxiv.org/help/withdraw.

82 arXiv.

83 "To Replace an Article," arXiv, accessed October 14, 2017, https://arxiv.org/help/replace.

84 "The arXiv Moderation System," arXiv, accessed October 14, 2017, https://arxiv.org/help/moderation.

85 arXiv.

86 "RG Score FAQ," ResearchGate, accessed February 14, 2018, https://www.researchgate.net/RGScore/FAQ.

financial investment to Academia.edu.[87] Unlike ResearchGate or Academia.edu, arXiv.org allows for completely open sharing of data that has been moderated for uniqueness, but it does not individually track each publication's downloads. Though the least user-friendly, arXiv.org had the most robust identifiers for each publication, which improves the ability to locate publications over time. For each of these websites hosting is the risk of domain name and/or address change. If ownership of the site address is not maintained, the site and the information contained within are at risk of corruption or deletion.

Table 1 summarizes the evaluated technologies' ability to meet the requirements of the thesis question.

Table 1. Data Sharing Technology Evaluation

Data Sharing					
	Open Data Sharing	Data Uniqueness Verification	Data User Trust	Data Immutability	Data Identification and Traceability
ResearchGate	No	Some	Some	No	Some
Academia.edu	No	No	No	No	Some
arXiv	Yes	Some	Some	No	Some

Table Key:

Yes	The site does offer solutions for the defined parameter
Some	The site offers a partial solution for the defined parameter
No	The site does not offer any solutions for the defined parameter
N/A	Not applicable based on the services provided by the site

[87] Academia.edu, "What Is Academia Premium?"

III. DATA TRUST AND IMMUTABILITY

Open sharing of academic data and papers must occur in an environment of trust and immutability. For data to be shared openly it is necessary to establish a sharing environment where both consumers and publishers have established trust within the community. In addition to trusting the members within the sharing environment it is also critical that information not be deleted, redacted, or withheld. To the greatest possible extent, data needs to be immutable. There are circumstances where information will need to be classified or kept secret to protect patents or other legal obligations; however, as discussed in Chapter I, it is absolutely necessary to share as much as possible to further innovation. This chapter will evaluate current technologies designed to create environments of trust with immutable records, to include advanced encryption standard, blockchain, and hyperledger fabric.

A. ADVANCED ENCRYPTION STANDARD

The advanced encryption standard technology is capable of establishing trust. Developed in the late 1990s, the advanced encryption standard was created via a partnership between government agencies, academia, and private industry to securely send files from one person or location to another.[88] On October 2, 2000, the United States government chose the Rijndael algorithm to be the new standard for encryption for point-to-point digital file. It was also widely praised in private industry. The Rijndael algorithm as adopted as the official encryption backbone in November 2001.[89]

This new block cipher method is capable of encrypting with keys of 128, 192, and 256 bits for blocks of 128 bits.[90] According to Dworkin et al., it is unfeasible to use brute

---

[88] Mitchell C. Richards, *AES: The Making of a New Encryption Standard* (North Bethesda, MD: SANS Institute, 2001), https://www.sans.org/reading-room/whitepapers/vpns/aes-making-encryption-standard-740.

[89] Morris J. Dworkin et al., *Advanced Encryption Standard (AES)*, Federal Information Processing Standards Publication 197 (Gaithersburg, MD: National Institute of Standards and Technology, 2001), https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.197.pdf.

[90] Dworkin et al.

force to break the advanced encryption standard, which is also effective against mathematical hacks attempting to reverse engineer the algorithm. The algorithm is so successful it was first adopted by the National Security Administration, then many other industries also adopted the advanced encryption standard.[91]

Yet, the advanced encryption standard is not only used by large industries. It is also used to secure many web transactions.[92] The AS2, FTPS, HTTPS, SFTP, and WebDAVS protocols all use the advanced encryption standard to transfer files securely.[93] However, the advanced encryption standard is not without vulnerabilities, including vulnerability to timing attacks.[94]

Advanced encryption standard can encrypt data to send between trusted sources from point-to-point. By design, it does not allow for open sharing. However, the advanced encryption standard does not contain a mechanism to verify uniqueness, track where a document goes beyond the initiator and receiver(s), or prevent deletion from the sender or receiver. There is, however, an encryption technology that allows for open sharing of immutable information on a broad scale called blockchain.

B. BLOCKCHAIN

Blockchain is a technology created in 2008 by "Satoshi Nakamoto"[95] to revolutionize financial transactions and digital currency.[96] This technology was originally used as a means to move bitcoins, the first blockchain digital currency, from one location

---

91 Dworkin et al.

92 Dworkin et al.

93 John Carl Villanueva, "What AES Encryption Is and How It's Used to Secure File Transfers," *JSCAPE* (blog), May 19, 2015, http://www.jscape.com/blog/aes-encryption.

94 Abdullah Al Hasib and Abdul Ahsan Md. Mahmudul Haque, "A Comparative Study of the Performance and Security Issues of AES and RSA Cryptography," *Convergence and Hybrid Information Technology, 2008. ICCIT '08. Third International Conference On* 2 (2008): 505–10, https://doi.org/10.1109/ICCIT.2008.179.

95 "Satoshi Nakamoto" is the fictitious name used by the creator or creators of bitcoin.

96 Wikipedia, s.v. "Blockchain (Database)," February 28, 2017, https://en.wikipedia.org/w/index.php?title=Blockchain_(database)&oldid=767916348.

or person to another without having to use a bank or other third party vendor.[97] A particularly enterprising piece of technology, this process is the first created able to track and confirm assets are only distributed once, effectively preventing the possibility of double expenditure of any singular asset.[98] Blockchain technology is based on the building of complete data sets, called blocks, that are built one atop another and chained together.[99] Figure 1, an infographic from Quora.com, shows how blockchain technology works for financial transactions.[100]

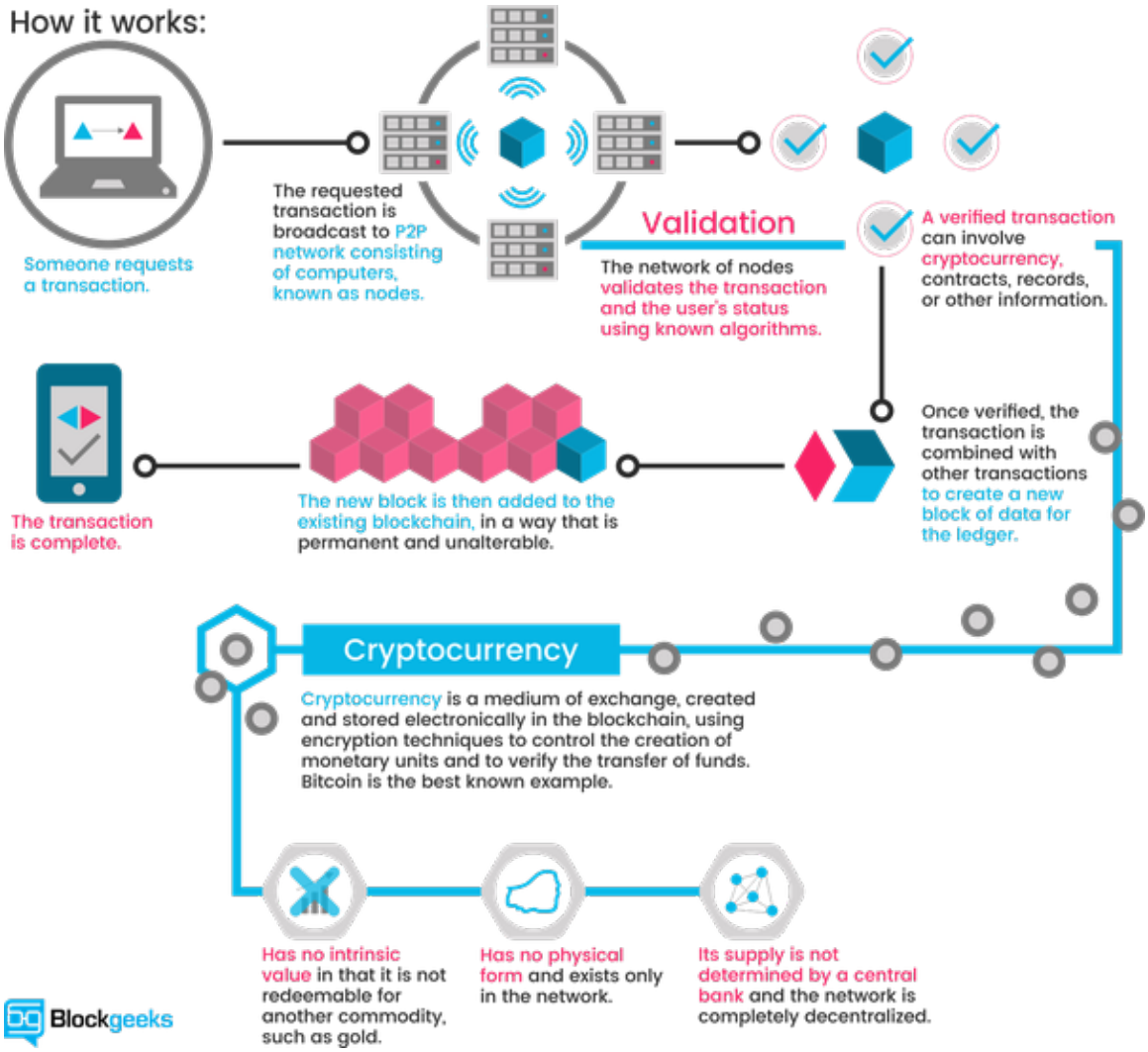
[97] Dan Bradbury, "In Blocks We Trust (Bitcoin Security)," *Engineering Technology* 10, no. 2 (March 2015): 68–71, https://doi.org/10.1049/et.2015.0208.

[98] Francois Zaninotto, "The Blockchain Explained to Web Developers, Part 1: The Theory," *Marmelab* (blog), April 28, 2016, http://marmelab.com/blog/2016/04/28/blockchain-for-web-developers-the-theory.html.

[99] Zaninotto.

[100] "How Does Bitcoin Blockchain Work and What Are the Rules behind it?," Quora.com, October 1, 2016, https://www.quora.com/How-does-Bitcoin-Blockchain-work-and-what-are-the-rules-behind-it.

Figure 1. How Bitcoin Works[101]



This diagram describes how the blockchain technology underlying bitcoin works.

Each block is built of transaction information that has an encryption algorithm that must be solved to officially validate it.[102] It is through the process of validation that bitcoins are created. Each computer in the peer network that does the work to solve the algorithm to add another block to the chain, also referred to as mining, is paid for services rendered in bitcoin.

---

[101] Source: Quora.com.

[102] Quora.com.

However, it is also necessary to verify the authority of those validating the transaction, and that is where the brilliance of the hash system for the blockchain comes into play. According to an article by Xingping Min, Qingzhong Li, Lei Liu, and Lizhen Cui, in order to add another block to the chain, a person would have to have an appropriate hash, or code created with data from the previous block.[103] This process, called ratcheting, prevents people from just creating a block and attempting to insert it into an existing chain. The block has to be consistent with the rest of the hash process throughout the entirety of the chain.[104] Any generated hash created to mimic the actual one in an effort to defraud the system would be easy to detect because of inconsistencies with the rest of the chain.[105]

This added level of protection allows the information contained within the transaction chain to be open for audit by anyone, but only those with the specific encrypted permissions are allowed to modify the transaction.[106] According to Zaninotto, all bitcoin transaction hashes start with a series of zeros to separate them from other information that may exist within the block.[107] However, bitcoin transactions are not the only type of data stored on blockchain blocks; blockchain technology is simply a peer provided publicly accessible database.[108] When considered in this most basic form, the potential applications of blockchain are significant, including those in Figure 2, by Elena Mesropyan.[109]

---

[103] Xinping Min et al., "A Permissioned Blockchain Framework for Supporting Instant Transaction and Dynamic Block Size," in *2016 IEEE Trustcom/BigDataSE/ISPA* (2016): 90–96, https://doi.org/10.1109/TrustCom.2016.0050.

[104] Zaninotto, "Blockchain Explained to Web Developers."

[105] Bradbury, "In Blocks We Trust."

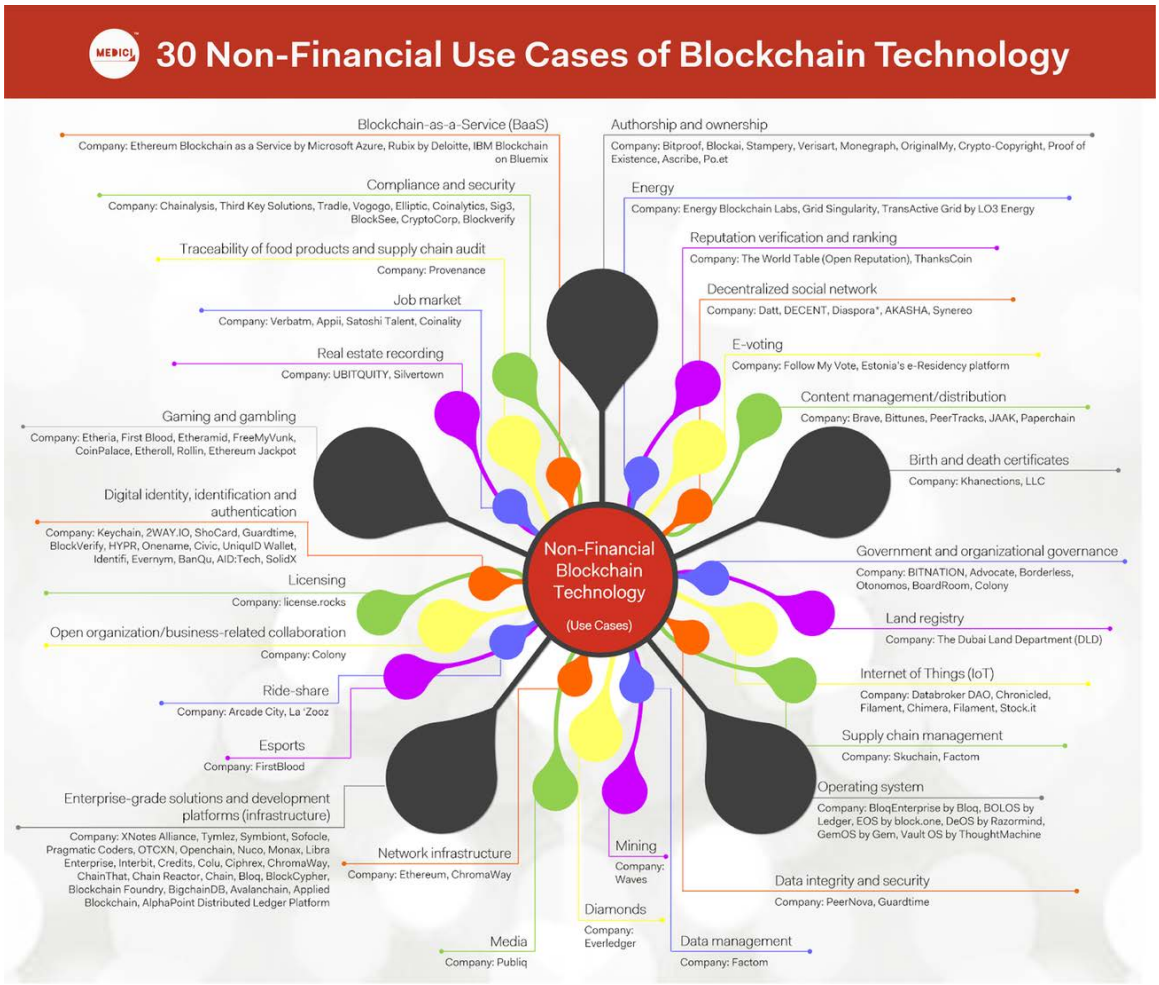
[106] Zaninotto, "Blockchain Explained to Web Developers."

[107] Zaninotto.

[108] Kevin Peterson et al., "A Blockchain-Based Approach to Health Information Exchange Networks," (research paper, Health Information Technology, 2016), 2, https://www.healthit.gov/sites/default/files/12-55-blockchain-based-approach-final.pdf.

[109] Bradbury, "In Blocks We Trust"; "Blockchain Use Cases: Comprehensive Analysis & Startups Involved," *Blockchain* (blog), July 29, 2015, https://gomedici.com/blockchain-use-cases-comprehensive-analysis-startups-invoved/.

Figure 2. Blockchain Use Cases[110]



Though a nascent technology, many innovators are finding alternate uses for blockchain in other economic and business sectors.

Perhaps the most promising non-cryptocurrency application for blockchain is handling smart contacts. Smart contracts are executions on the blockchain built around transactional parameters. Essentially, the blockchain has an "if"/"then" program to execute transactions.[111] For instance, if a car-sharing company wanted to use blockchain smart

[110] Source: Elena Mesropyan, "30 Non-financial Use Cases of blockchain Technology," *Blockchain*, December 18, 2017, https://gomedici.com/30-non-financial-use-cases-of-blockchain-technology-infographic/.

[111] "What Are Smart Contracts? A Beginner's Guide to Smart Contracts," Blockgeeks, accessed February 8, 2018, https://blockgeeks.com/guides/smart-contracts/.

contracts to rent its cars, the company would create a smart contract that consumers would activate by paying the assigned rental fee. Sandra, wanting to rent Car Y, would initiate the contract by paying the rental fee in cryptocurrency. The contract would complete by sending Sandra the code to unlock the car that will last for the duration of the rental contract. No other intervention or intermediary is required for the transaction. Once complete, the transaction is published on the blockchain and distributed to all peers.

C. ESTONIA EXAMPLE

Estonia has capitalized on the potential of blockchain. In 1991, when Estonia separated from the Soviet Union, it did not have many financial, natural, or human capital resources to draw on for establishing the new government.[112] Realizing the importance of efficiency, the newly formed Estonian government created a three-part strategy to build a new electronic government, or e-government.[113] The first part detailed how to create a mechanism to identify unique citizens and their associated government information. To meet this need, each individual citizen is assigned a unique identity code generated by the population registry.[114] The next part of the strategy established a way for citizens to interact securely with the government. To solve this need, Estonian national identification cards with embedded electronic chips were issued to citizens.[115] In addition to the card, each citizen has two unique codes or personal identification numbers (PINs) that have to be used in combination with the identification card for online requests and transactions.[116] The first PIN is the personal identifier of the citizen, the second PIN number is the digital signature code.[117] The combination of these elements, the identification card, and the PINs

---

[112] Jaan Priisalu and Rain Ottis, "Personal Control of Privacy and Data: Estonian Experience," *Health and Technology* 7, no. 4 (December 1, 2017): 441–51, https://doi.org/10.1007/s12553-017-0195-1.

[113] Priisalu and Ottis.

[114] Priisalu and Ottis, 443.

[115] Kristjan Vassil, "Estonian E-Government Ecosystem: Foundation, Applications, Outcomes" (report, World Bank, 2016), http://pubdocs.worldbank.org/en/165711456838073531/WDR16-BP-Estonian-eGov-ecosystem-Vassil.pdf.

[116] Kristjan Vassil, 4.

[117] Kristjan Vassil.

constitute a digital signature that prevents false identity claims.[118] The third and final part securely connected the citizens and government services with all necessary information. The backbone for data connection is called the X-Road.[119] Dr. Vassil, from Estonia's University of Tartu, describes the X-Road:

> This open design is accompanied by rigid security measures—authentication, multilevel authorization, high-level log processing and monitoring, encrypted and time stamped data traffic—the basic functionalities that are covered within the very structure of X-Road.[120]

The key to this system is storing information in only one place.[121] Data and files are not to be duplicated or distributed anywhere else on the network.[122] The blockchain on the X-Road simply encrypts and records the transaction data from across the X-Road. The functionality of the X-Road is depicted in Figure 3, created by Gary Anthes.[123]

---

[118] Priisalu and Ottis, "Personal Control of Privacy and Data, "443–4.

[119] Vassil, "Estonian E-Government Ecosystem."

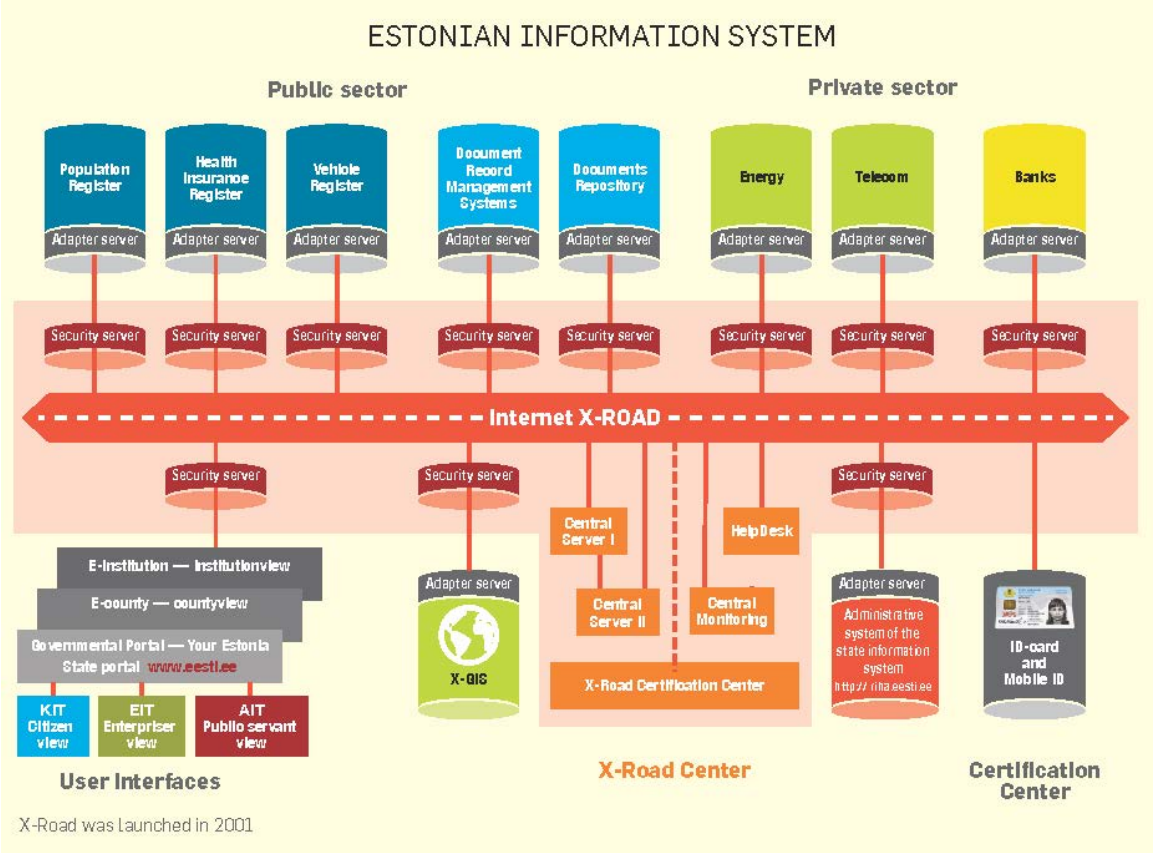
[120] Vassil, 30.

[121] Helen Margetts and Andre Naumann, "Government as a Platform: What Can Estonia Show the World?" (research paper, University of Oxford, 2017), https://www.politics.ox.ac.uk/publications/government-as-a-platform-what-can-estonia-show-the-world.html.

[122] Margetts and Naumann.

[123] Gary Anthes, "Estonia: A Model for e-Government," *Communications of the ACM* 58, no. 6 (May 2015): 18–20, https://doi.org/10.1145/2754951.

Figure 3. Estonian Information System[124]



The X-Road, as shown above, provides secure centralized access to government databases.

Through careful planning, this open and transparent government model has thrived for more than 15 years.[125] According to Helen Margetts and Andre Naumann, the first true challenge to the system came in 2007, when a denial of service attack was launched against Estonia. Parliament, a few government agencies, newspapers, banks, and broadcast agencies were affected by the attacks.[126] The government acted quickly, blocking all connections outside Estonia for a few hours so the attacks could be stopped.[127] None of

[124] Source: Anthes.

[125] Margetts and Naumann, "Government as a Platform."

[126] Margetts and Naumann.

[127] Margetts and Naumann.

the databases or e-government services were breached during the attack.[128] The security measures in place protected the privacy and security of the Estonian people.

Though blockchain has many benefits for financial transactions and data storage, it also has some limitations. Coding blocks requires time and significant computing power.[129] To increase the speed of mining, some bitcoin miners have begun to pool resources. Each of the contributing miners is then compensated relative to the amount of progress each contributed to create a given block.[130] Pooling of resources has allowed for block creation time to remain around ten minutes for each block.[131] However, pooling resources can also break the blockchain's security.

If 51% or more of a blockchain is controlled by any one pool, that pool has enough information to break the algorithm and ruin the security of the whole system.[132] The Homeland Security Institute defines a 51% attack as:

> A condition in which more than half the computing power on a cryptocurrency network is controlled by a single miner or group of miners. That amount of power theoretically makes them the authority on the network and gives them power to (1) interfere with issuing and confirming transactions, (2) double-spend bitcoin, or (3) prevent other miners from mining valid blocks.[133]

As of December 31, 2014, only one 51% attack had been detected on the bitcoin blockchain. In June of 2014, GHash.IO controlled 51% of the hashing power for the bitcoin

---

[128] Margetts and Naumann.

[129] Department of Homeland Security, *Risks and Threats of Cryptocurrencies* (Falls Church, VA: Homeland Security Studies & Analysis Institute, 2014), https://www.anser.org/docs/reports/RP14-01.03.03-02_Cryptocurrencies%20508_31Dec2014.pdf.

[130] Department of Homeland Security.

[131] Min et al., "A Permissioned Blockchain Framework," 90.

[132] Kamamnashis Biswas and Vallipuram Muthukkumarasamy, "Securing Smart Cities Using Blockchain Technology," in *IEEE 14th International Conference on Smart City* (2016): 1392–93, https://doi.org/10.1109/HPCC-SmartCity-DSS.2016.0198.1392.

[133] Department of Homeland Security, *Risks and Threats of Cryptocurrencies*, xiv.

blockchain.[134] After that, GHash.IO announced they would control no more than 39.99% share of the hashing power to prevent a collapse of the system.[135]

Blockchain could also be advantageous is in the storage and security of university research data. To help with the security of data stored on blockchains, two types of blockchain hash management systems are available. The first is blockchain with public keys.[136] With a public key blockchain, researchers would be able to collaborate globally while simultaneously allowing others to use the data generated. By making the data publicly available with a blockchain, the researchers are still able to control and time-stamp record all the data that they are generating while allowing others to review it as soon as the block is verified. This helps to keep the data from being adulterated or stolen after the fact, effectively securing the data while allowing it to be used by others.

However, for studies with information that should not be publicly available, private blockchain keys allow for secure blockchain storage that can only be accessed by a specific hash key held by the data originator.[137] Specifically, those projects that concern personal information, health information, patent information, or are related to homeland security need to have the data secured.

As an example, researchers completing laboratory trials on a public key blockchain would keep sensitive patient information on private key blockchain when moving onto the clinical trials. Moving patient information to the private key blockchain keeps patent information protected from those without a need to know as the collaborating teams gather necessary data for the Food and Drug Administration's therapy approval.

With multiple teams generating innovative research data, the Food and Drug Administration approval will be granted much more quickly than if any of the teams had attempted this process alone. Because the level of encryption is so rigorous, private key blockchain provides an even greater level of security for sensitive data than traditional

---

[134] Department of Homeland Security.

[135] Department of Homeland Security.

[136] Department of Homeland Security.

[137] Department of Homeland Security.

encryption for internet connected and local storage solutions today.[138] Table 2, created by Zheng et al., summarizes the benefits of public key and private key blockchains. The third type included on the table is the consortium blockchain, which will be discussed later in this chapter.

Table 2. Comparison among Public Blockchain, Consortium Blockchain, and Private Bockchain[139]

Property	Public blockchain	Consortium blockchain	Private blockchain
Consensus determination	All miners	Selected set of nodes	One organization
Read permission	Public	Could be public or restricted	Could be public or restricted
Immutability	Nearly impossible to tamper	Could be tampered	Could be tampered
Efficiency	Low	High	High
Centralized	No	Partial	Yes
Consensus process	Permissionless	Permissioned	Permissioned

This table defines each type of blockchain as it relates to consensus determination, read permission, immutability, efficiency, centralization, and consensus process.

Though blockchain offers an immutable openly accessible platform, it still does not have the capability of tracking where information goes once downloaded from the blockchain. It also does not have any mechanism for verifying uniqueness of the data contained within the blocks.

In addition to traditional blockchains, new technologies are emerging that improve on the speed and storage requirements of traditional blockchains, making them more

[138] Curtis Miles, "Blockchain Security: What Keeps Your Transaction Data Safe?" *Blockchain Unleashed* (IBM Blockchain Blog), December 12, 2017, https://www.ibm.com/blogs/blockchain/2017/12/blockchain-security-what-keeps-your-transaction-data-safe/.

[139] Source: Zibin Zheng et al., "Blockchain Challenges and Opportunities: A Survey," *International Journal of Web and Grid Services* (December 2017).

accessible, faster, and more functional. Table 3, created by Zheng et al., provides a summary of the different emerging blockchain-based technologies being created to meet business needs.[140] Across the top of table are the different types of encryption algorithms used; the properties listed on the left hand column are the key features for blockchain-based technologies to determine the best technology for a desired business use.

Table 3. Typical Consensus Algorithms Comparison[141]

Property	PoW	PoS	PBFT	DPOS	Ripple	Tendermint
Node identity management	open	open	permissioned	open	open	permissioned
Energy saving	no	partial	yes	partial	yes	yes
Tolerated power of adversary	$< 25\%$ computing power	$< 51\%$ stake	$< 33.3\%$ faulty replicas	$< 51\%$ validators	$< 20\%$ faulty nodes in UNL	$< 33.3\%$ byzantine voting power
Example	Bitcoin	Peercoin	Hyperledger Fabric	Bitshares	Ripple	Tendermint

This table defines each type of blockchain as it relates to node identity management, energy savings, tolerated power of adversary (or malignant nodes on the chain), and provides a brand name as an example of each type.

D. HYPERLEDGER FABRIC

Among those listed in the previous section is the hyperledger fabric, which is a permissioned, energy-saving, blockchain-based technology. Started by the Linux Foundation in 2015, the hyperledger fabric has a modular organization, allowing for more

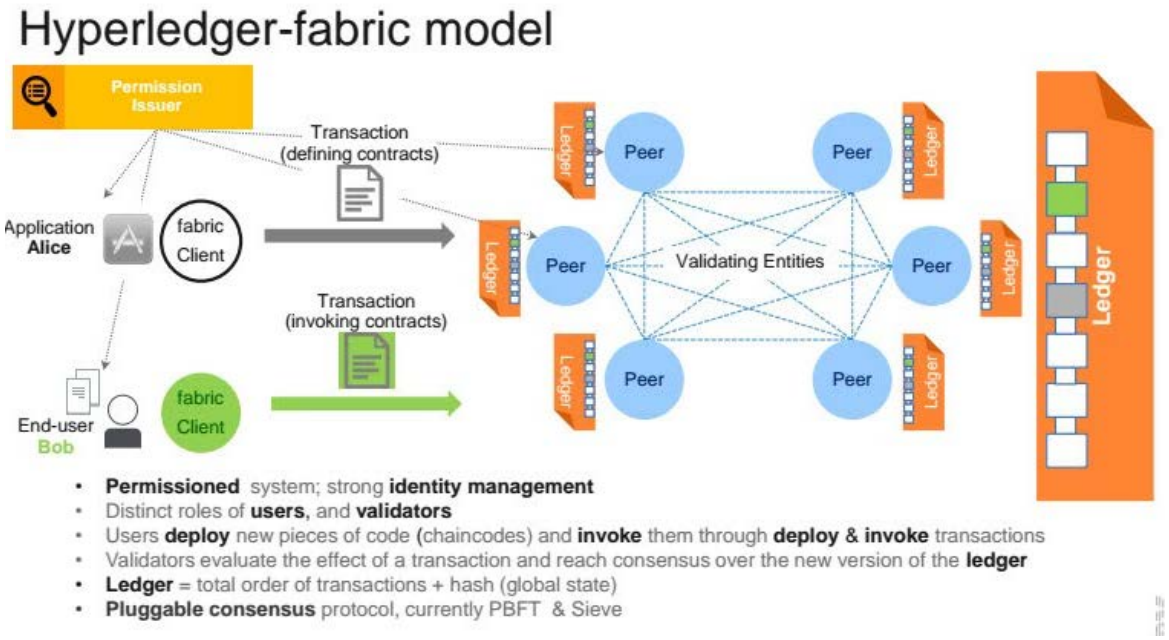
[140] Zheng et al.

[141] Source: Zheng et al.

efficiency and flexibility, enabling it to be used with other technologies and types of code than traditional blockchain.[142]

Figure 4, created by IBM, gives a graphical representation of how the hyperledger fabric works, for example, where Alice is initiating a transaction and Bob is completing it.[143] The entire transaction is recorded on the immutable ledger in the hyperledger fabric.

Figure 4. Hyperledger Fabric Model[144]



This diagram shows how hyperledger fabric processes transactions, and adds them to the distributed ledger.

---

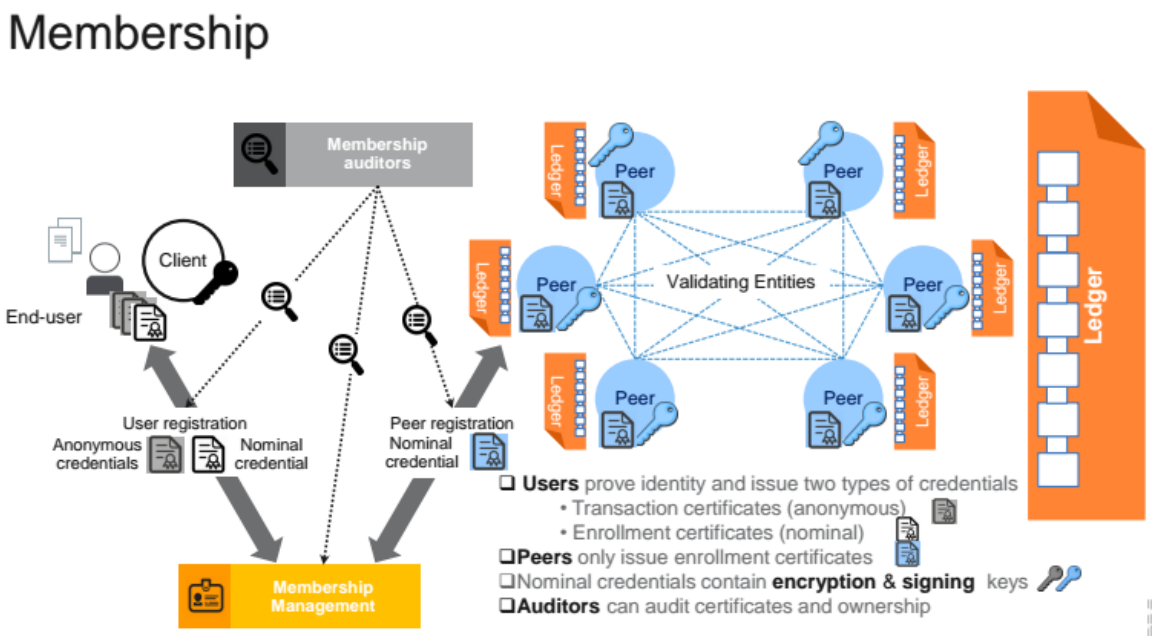
142 Lakshmi Siva Sankar, M. Sindhu, and M. Sethumadhavan, "Survey of Consensus Protocols on Blockchain Applications," in *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)* (2017): 1–5, https://doi.org/10.1109/ICACCS.2017.8014672; Marko Vukolic, "Hyperledger Fabric: An Open-Source Distributed Operating System for Permissioned Blockchains" (report, IBM Research, 2017), https://blockchain-summer.epfl.ch/talks/hyperledger-fabric-vukolic.pdf; "Linux Foundation Unites Industry Leaders to Advance Blockchain Technology," *The Linux Foundation* (blog), December 17, 2015, http://www.linuxfoundation.org/press-release/linux-foundation-unites-industry-leaders-to-advance-blockchain-technology/.

143 Roger Strukhoff, "How Hyperledger Fabric Delivers Security to Enterprise Blockchain," *Altoros* (blog), November 14, 2016, https://www.altoros.com/blog/how-hyperledger-fabric-delivers-security-to-enterprise-blockchain/.

144 Source: Strukhoff.

To reduce the risk of corruption to the blockchain, hyperledger requires each member to register with an enrollment Certificate Authority.[145] This membership allows the peer to submit transactions with appropriate transaction Certificate Authority.[146] IBM also created Figures 5 and 6, which depict how the Certificate Authorities manage and audit memberships.

Figure 5. Hyperledger Membership[147]



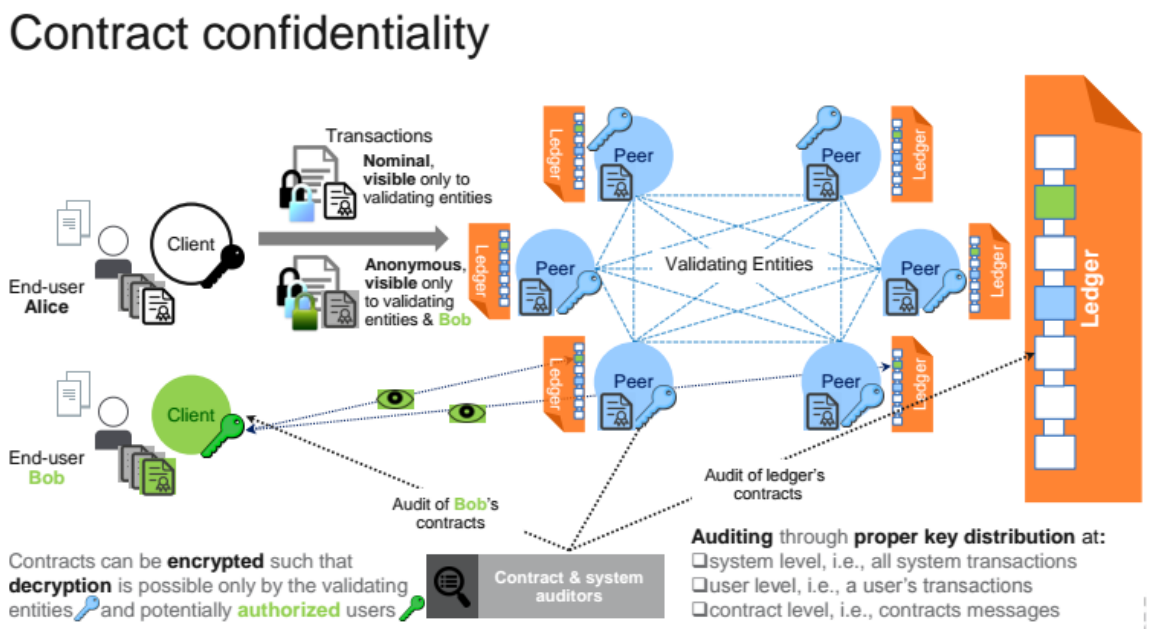
This diagram shows how hyperledger fabric audits and manages members of the hyperledger fabric.

145 Strukhoff.

146 Sankar, Sindhu, and Sethumadhavan, "Protocols on Blockchain Applications."

147 Source: Strukhoff, "Hyperledger Fabric Delivers Security."

Figure 6. Hyperledger Contract Confidentiality[148]



This diagram shows how confidentiality is maintained on the hyperledger fabric.

One of the key features that separates the hyperledger fabric from traditional blockchain is the ability to scale easily to meet the business needs.[149] Scaling is achieved through separating endorsers and committers from concenters, freeing up computing resources.[150] In Figure 7, IBM graphically describes how this process occurs.

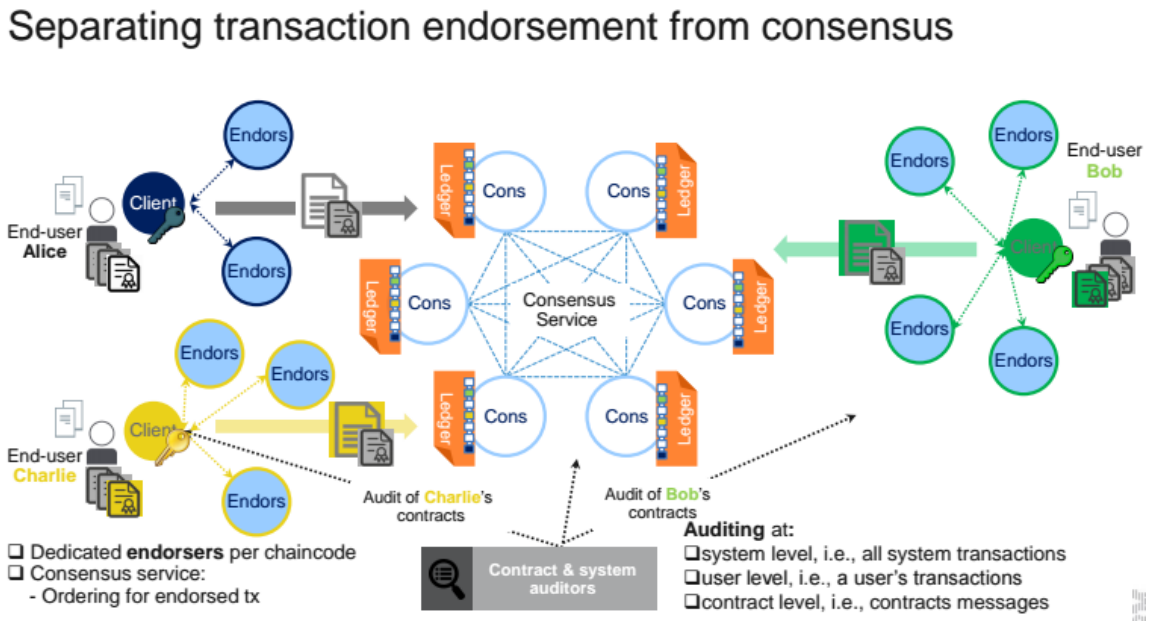
---

148 Source: Strukhoff.

149 "Hyperledger Fabric Explainer," YouTube video, uploaded by Hyperledger, April 28, 2017, https://www.youtube.com/watch?v=js3Zjxbo8TM.

150 Hyperledger.

Figure 7. Hyperledger Separating Transaction Endorsement from Consensus[151]



This diagram shows how hyperledger fabric separates endorsements and consensus to increase efficiency and scalability of the system.

Advanced encryption standard, blockchain, and hyperledger fabric offer options to share information with trust. Advanced encryption standard provides the ability to securely share information from one peer to another, but does not offer the ability for sharing outside the initial peers. Blockchain offers more open sharing, however it requires significant computing and energy resources. Blockchain is also difficult to search and limits scalability. Hyperledger fabric offers more sharing and scalability options, but has not been used in this application before. Unfortunately, none of the solutions examined in this chapter are capable of verifying data uniqueness or tracking data outside its native environment.

Table 4 summarizes the evaluated technologies' ability to meet the requirements of the thesis question.

151 Source: Strukhoff, "Hyperledger Fabric Delivers Security."

Table 4. Data Immutability and Trust Technology Evaluation

Immutability and Trust					
	Open Data Sharing	Data Uniqueness Verification	Data User Trust	Data Immutability	Data Identification and Traceability
Advanced encryption standard	No	No	Yes	No	Some
Blockchain	Some	No	Yes	Yes	Some
IBM Hyperledger	Some	No	Yes	Yes	Some

Table Key:

Yes	The site does offer solutions for the defined parameter
Some	The site offers a partial solution for the defined parameter
No	The site does not offer any solutions for the defined parameter
N/A	Not applicable based on the services provided by the site

IV. DATA IDENTIFICATION AND TRACEABILITY

As important as it is to share data, it is also important to track all the locations where the data exists. Both for authors and consumers, it is necessary to ensure that all copies of the data are accurate. As well, it is also necessary to ensure sensitive data has not fallen into the wrong hands. This chapter will explore three tracking technologies that currently exist to meet the need are digital object identifier (DOI), persistent uniform resource locator, and international standard serial number.

A. DIGITAL OBJECT IDENTIFIER

DOIs allow each creation in cyberspace to be assigned a unique string of characters that create the foundation for tracking the object through cyberspace. According to the *DOI Handbook*, a DOI can be assigned to any entity, whether digital or physical by a designated registration agency or by developing a community to create one.[152] The value assigned to a document is a combination of an existing identifier, location (URL), and metadata to create a handle.[153] Any agency can be assigned a prefix to begin all agency assigned DOIs.[154] The prefix is followed by a / then a string of numbers uniquely identifying the particular object.[155] If there is a conflict, DOI names can be deconflicted for free through a DOI registration agency.[156] It is possible to reassign a DOI to allow interoperability into the future.[157]

Though these numbers can be reassigned, each provides registration and location information for the object, and the registered owner. If the location of an object changes the DOI can be updated to include the new location, but there is no mechanism for

[152] International DOI Foundation, *DOI Handbook* (Wilmington, DE: The Corporation Trust Company, 2016), https://www.doi.org/hb.html.

[153] International DOI Foundation.

[154] International DOI Foundation.

[155] International DOI Foundation.

[156] International DOI Foundation.

[157] International DOI Foundation.

preventing an object from being deleted, or removed entirely from the system.[158] Unfortunately, DOIs only track the origin copy of the object, not the location or movement of any additional copies or prints made by the consumer of the object.[159] Additionally, the framework offers no method for identifying the originality of the work receiving the DOI; there is no mechanism to verify the object does not include plagiarism.

B. PERSISTENT UNIFORM RESOURCE LOCATOR

Whereas a DOI has a location associated, persistent uniform resource locators (PURLs) are the wayfinding to an object's location. According to the Online Computer Library Center, PURLs, created in 1995, are uniform resource identifiers that direct users to an intermediate resolution service.[160] This service points the user to the object, much like a traditional URL, however if an object moves the intermediate PURL service will redirect the user to the correct new address without having to change the identifier for the object.[161] To increase PURL implementation, the Online Computer Library Center openly shares the PURL source code so that others may generate PURLs at will.[162] Though this technology is openly available, and allows for the traceability of the original object as the internet evolves, it does not prevent the object from being deleted or have any mechanism for verifying the object is free of plagiarism.[163]

C. INTERNATIONAL STANDARD SERIAL NUMBER

Another tracking method was developed by the International Standards organization in the early 1970s called the International Standard Serial Number to identify newspapers, journals, magazines, and periodicals across all mediums, according to the

---

158 International DOI Foundation.

159 International DOI Foundation.

160 "PURL," Online Computer Library Center, accessed October 14, 2017, http://www.oclc.org/research/themes/data-science/purl.html.

161 Online Computer Library Center.

162 Online Computer Library Center.

163 Online Computer Library Center.

ISSN (International Standard Serial Number) Manual Section 0.[164] Established in 1976 the International Centre for the Registration of Serial Publications manages the International Standard Serial Number registration process.[165] Over 1.9 million publications have a registered ISSN.[166] Each nation participating in ISSN has a National Centre responsible for assigning ISSN numbers for that nation.[167] The International Centre for the Registration of Serial Publications assigns a block of ISSN numbers to each National Centre, who are then responsible for assigning the numbers sequentially to any works upon approval of a valid request.[168] Each request for an ISSN includes the key title, location, bibliographic record, and country codes.[169] This allows the publication to be tracked in perpetuity.

Despite the large number of registered publications, an ISSN may be reassigned or deleted if a publication has changed or been removed, according to the ISSN manual Section 2.5.[170] ISSNs are also granted based on request to the National Centre, which does not verify that materials within the publication are original or properly cited.[171] Additionally, there is no trust inherent in the system for users of ISSN registered publications. The ISSN is for cataloging and retrieval. Once the document has reached the end user, there is no accountability for its use or further transmission.

Though there are multiple methods for identifying and tracking the current publication origin location of unique publications, none of the methods allow for tracking

---

164 ISSN, *ISSN Manual* (Paris: ISSN InterNational Centre, 2015), http://www.issn.org/understanding-the-issn/assignment-rules/issn-manual/.

165 "The International Centre for the Registration of Serial Publications," ISSN, accessed October 14, 2017, http://www.issn.org/the-centre-and-the-network/our-mission/the-international-centre-for-the-registration-of-serial-publications-cieps/; "What Is an ISSN?" ISSN, accessed October 14, 2017, http://www.issn.org/understanding-the-issn/what-is-an-issn/.

166 "The ISSN International Register," ISSN, accessed October 15, 2017, http://www.issn.org/understanding-the-issn/the-issn-international-register/.

167 "The ISSN Network Today," ISSN, accessed October 15, 2017, http://www.issn.org/the-centre-and-the-network/members-countries/the-issn-network-today/#.

168 ISSN, *Manual*.

169 ISSN, *Manual*.

170 ISSN, *Manual*.

171 ISSN, *Manual*.

beyond the initial end user. Even with registries, none of the options prevent record deletion, either. Though PURL did offer its source code to anyone, neither of the other two organizations offered open solutions, and none of the options established trust with the end user community or verified that publications were unique prior to publication or assigning a tracking number.

Table 5 summarizes the evaluated technologies' ability to meet the requirements of the thesis question.

Table 5. Data Identification and Traceability Technology Evaluation

Identification and Traceability					
	Open Data Sharing	Data Uniqueness Verification	Data User Trust	Data Immutability	Data Identification and Traceability
Digital Object Identifier (DOI)	N/A	No	No	No	Some
Persistent Universal Resource Locator (PURL)	N/A	No	No	No	Some
Internet Standard Serial Number (ISSN)	N/A	No	No	No	Some

Table Key:

Yes	The site does offer solutions for the defined parameter
Some	The site offers a partial solution for the defined parameter
No	The site does not offer any solutions for the defined parameter
N/A	Not applicable based on the services provided by the site

V. DATA UNIQUENESS

Verifying that data is unique is an increasingly difficult challenge. To date, no technology currently exists to verify data uniqueness. The technologies explored in this chapter, Turnitin, iThenticate, and Google search engine are capable of verifying strings of words in a specific language, rather than the data or true content of the works submitted for review. For the purpose of this chapter, the technologies will be evaluated on its ability to meet its own self-defined capabilities, acknowledging that these technologies are unable to meet the capability of true data uniqueness verification as defined for this thesis. The algorithms compare word order and sentence composition; however, language has human nuances that make detecting plagiarism a constantly evolving challenge. Add in that millions of documents are published every day globally, and it becomes a near-impossible task. However, there are technologies that exist today laying the groundwork for how to verify data uniqueness.

A. TURNITIN

One technology used for verifying language uniqueness is Turnitin. Turnitin is a website designed for use in educational settings.[172] Offering options for K-12 schools and Higher Education, Turnitin allows administrators, instructors, and students to compare their created documents to existing documents, for a fee.[173] In addition to English, the website has the capability of processing documents written in 29 different languages by translating any non-English writings to English then running them through the originality checking software.[174] According to the Turnitin website:

> Turinitin does not detect plagiarism per se; Turnitin just finds the text that matches other sources in the vast Turnitin databases and shows those

[172] Turnitin, "Technology to Improve Student Writing."

[173] "Turnitin," accessed October 14, 2017, http://turnitin.com/en_us/home.

[174] "Translated Matching," Turnitin, October 25, 2016, https://guides.turnitin.com/01_Manuals_and_Guides/Administrator_Guides/User_Guide/Translated_Matching.

> matches. It is up to a human being to determine whether those text matches are a problem or not.[175]

Turnitin technology is a basic tool to detect document uniqueness, however it is unable to operate without human intervention. The effectiveness of the document scans are also limited by the number of documents stored in the database. According to the Turnitin website, the software checks documents against "billions of internet documents, archived internet data that is no longer available live on the web, a local repository of previously submitted papers, and subscription repository of periodicals, journals, and publications." Additionally, unless a student "opts out" of the database, their paper will be added to the repository.[176]

B. iTHENTICATE

Another resource for checking language uniqueness is iThenticate. Owned by the same parent company as Turnitin, iThenticate is described as the "largest scholarly comparison database/[for] plagiarism detection," iThenticate is targeted toward determining the uniqueness of scholarly and professional publications.[177] iThenticate compares files to "590+ global, scientific, technical, and medical publishers…more than one million abstracts and citations from PubMed, and more than 20,000 research titles from EBSCOhost and the Gale Info Trac OneFile. iThenticate also maintains its own web crawler, indexing over 10 million web pages daily and totaling over 50 billion web pages."[178]

Unlike Turnitin, iThenticate's interface only supports English, Korean, and Japanese languages. However, the database does include documents from 30 different languages, and will match the text to the native language for a more accurate text

---

175 "Does Turnitin Detect Plagiarism?," Turnitin, accessed October 14, 2017, http://turnitin.com/en_us/resources/blog/421-general/1643-does-turnitin-detect-plagiarism.

176 "Top 15 Misconceptions about Turnitin," Turnitin, accessed October 14, 2017, http://turnitin.com/en_us/resources/blog/421-general/1644-top-15-misconceptions-about-turnitin.

177 "Plagiarism Detection Software," iThenticate," accessed October 14, 2017, www.ithenticate.com.

178 "FAQs | Plagiarism Software," iThenticate, accessed October 14, 2017, www.ithenticate.com/products/faqs.

comparison.[179] iThenticate also does not upload personally checked documents to a larger database. Any documents checked by an individual remain with that individual's account only and will not be available to anyone else for plagiarism comparisons. iThenticate also allows individual users to delete their work from the account at any time, allowing it to be a highly controlled and mutable record.[180] Much like Turnitin, iThenticate also provides users a similarity score to be interpreted by the report recipient for determining the level of plagiarism. Therefore, though iThenticate is a basic tool for checking document uniqueness, it is unable to meet the needs defined within the thesis question.

C. GOOGLE SEARCH ENGINE

Finally, an openly available resource for checking documents for originality is Google search engine. Though not as user friendly as Turnitin or iThenticate, Google's search engine does allow users to check the originality of a work 50–150 words at a time without creating a personal account or requiring any other barriers to use. Unfortunately, Google's search engine does not allow for sharing or maintaining records. The search is stored within a users' browser history and used for Google's metrics, but the original document remains unattainable until published on an alternative sharing source.

Table 6 summarizes the evaluated technologies' ability to meet the requirements of the thesis question.

[179] iThenticate, "FAQs."

[180] iThenticate, "FAQs."

Table 6. Data Uniqueness Technology Evaluation

Data Uniqueness					
	Open Data Sharing	Data Uniqueness Verification	Data User Trust	Data Immutability	Data Identification and Traceability
Turnitin	No	Some	N/A	No	N/A
iThenticate	No	Some	N/A	No	N/A
Google Search Engine	Yes	Some	N/A	No	N/A

Table Key:

Yes	The site does offer solutions for the defined parameter
Some	The site offers a partial solution for the defined parameter
No	The site does not offer any solutions for the defined parameter
N/A	Not applicable based on the services provided by the site

VI. ANSWERING THE RESEARCH QUESTION

To answer the research question—how can research universities openly and with trust share verified unique data that is both immutable and ultimately trackable?—I explored current technologies. Unfortunately, none of the examined technologies can meet all of these requirements. Therefore, my design proposal determines how to meet all the prescribed parameters by improving on existing technology to build an entirely new information sharing platform.

A. OPEN DATA SHARING

The first step in establishing a new information sharing platform is creating an open sharing environment. As evaluated in Chapter II, current technologies, including ResearchGate, Academia.edu, and arXiv allow for sharing of papers and publications, but are not designed to share raw research data. Each of these technologies disrupted the traditional journal and book publishing oligopoly. Though effective at providing a platform to widely share information, none has identified mechanisms to ensure each record is immutable. Therefore, an acceptable solution to the research question must be able to openly share with a multitude of sources while also preventing deletion and unauthorized editing to the information. The new technology must be accessible from the internet, but not solely reliant on a single source server.

B. TRUST AND IMMUTABILITY

Open sharing is directly tied to immutability. Simply being able to share information is not enough to solve the research question; the information must be protected against deletion or fraudulent editing. The solution for this problem lies in part with the technologies discussed in Chapter III. Of those discussed in Chapter III, advanced encryption standard, blockchain, and IBM's hyperledger fabric all offer secure transmission of shared information, but none was capable of open sharing. Estonia's X Road, on the other hand, is a user-friendly, easily searchable secure file transfer technology. The X Road also has the capability of automating searches and information retrieval to reduce the time, confusion, and stress that otherwise accompanies manually searching for

information. Therefore, solving the research question will require taking the best features of the X Road and hyperledger fabric to achieve a trusted sharing environment with an easily searchable repository of immutable data, files, and publications.

Estonia's X Road is an exceptional search and encryption tool. The X Road provides automated information retrieval, secure transfer, and tracks each step as it processes the requested information transaction. This is an exceptional service. However, it is not designed as a distributed ledger. By design, files are only stored where generated. This is extremely useful in a system that does not need to preserve data for posterity. To solve the thesis question, however, preserving data and publications in their original state is critical.

By contrast, the hyperledger fabric is designed to make files immutable by securing copies of all information submitted to the fabric across all members' servers. Though extremely useful at immutability, the hyperledger fabric is difficult to generally search, and is not designed to share information broadly across the system. Therefore, solving the research question will require a hyperledger fabric foundation with an X Road search functionality.

C. DATA IDENTIFICATION AND TRACKING

Earlier in Chapter IV, technology for publication identification and tracking were evaluated against the thesis question. Though each technology can identify publications, each has limited tracking capabilities. Unfortunately, current technology is limited to noting the original publication's location as it is updated or moves across the internet. The examined technologies do not track downloads or copies distributed by intermediates and end-users. To solve this part of the research question, different technology solutions are necessary.

As described previously in this chapter, Estonia's X Road provides automated search, retrieval, and secure delivery of requested information. Automated retrieval of information is possible because certain types of information are categorized and only stored in a few locations. As an example, health information about an individual will only be stored in a clinic or hospital where the files originated. The X Road prevents copies of files

from being sent across the network; each file stays where it originated and the information contained within the file is used to populate other forms or be used as a reference for the individual or pertinent doctors. Each health record belonging to a sample individual will be marked with the citizen identifier code belonging to that individual. Therefore, all that person's health records will have that single identifier code and be stored in the clinic or hospital where they were created. The simplicity of having one individual identifier to pull records from a few specific sources is powerful; however, it does not directly translate to a university research setting. Therefore, categorization must be incorporated into the hyperledger fabric submission identifier.

Explored in Chapter IV, the DOI system creates a unique alphanumeric string to identify each document by agency, URL, and metadata. Using a DOI-based system for file identification on the fabric will allow for coherent, easily searchable, and retrievable file identification in this proposed solution. Building on the DOI information, the identifier must include the date, version, and author identification code, as well as the university identification code, location on the fabric, and metadata. If the fabric submission is updated so, too, is the identification code. The previous version of the submission is not deleted or removed; it simply branches off the new version, with access nested within the new version. Generating detailed initial identifiers is key to programming the automation for accurate submission searching and retrieval.

Initial submission identification is only one piece of the tracking challenge. The other piece, which is arguably the more difficult part, is tracking where the submission has gone once it has been accessed or downloaded. Again, the Estonian X Road has a method of identifying access permissions and records. The X Road is effective at intra-system file transmission; however, it is not designed to track permissioned data beyond the initial destination. For instance, if some member of the system downloads a file with permission, then shares that file with someone outside the system (who does not have permission to access that file) using a flash drive or an SD card, there is no way to identify the file at its new destination. Solving this problem will require a new technology.

A beacon will be embedded within each submission file that is a non-mutable execution that connects the file back to the host system. Using enhanced text

steganography, the beacon will be woven into the document itself, unable to be detected by the operating system on any device.[181] Cloaking the beacon is vital to preventing its removal or modification. Additionally, this new enhanced text steganography allows more bits to be woven into the document which facilitates the executable instructions fitting within the document without adding considerably to the file size.

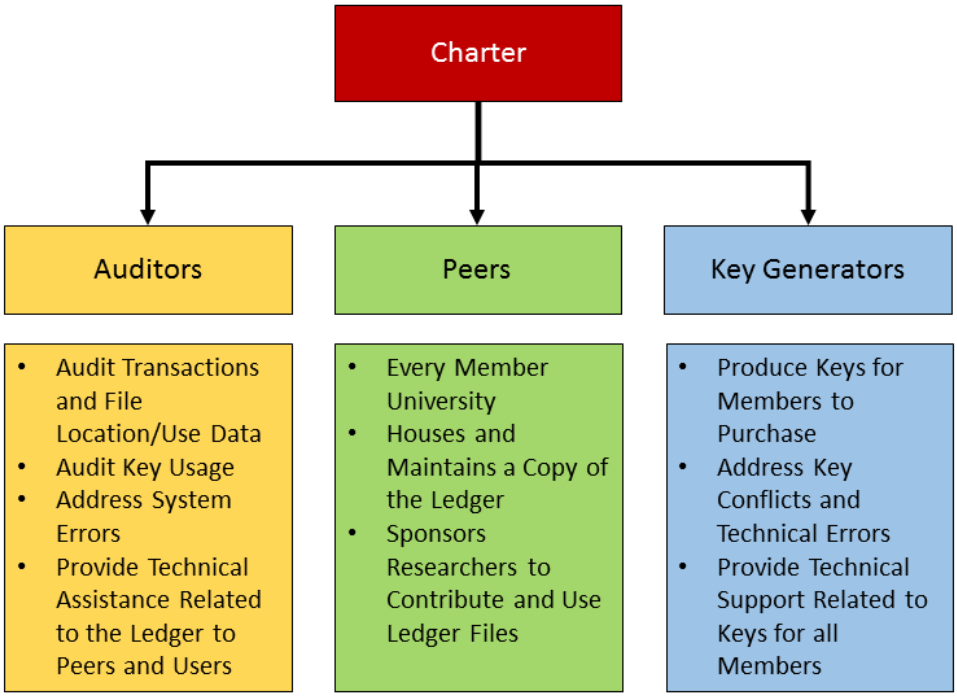
If it is a file with public permissions, the only action beyond recording the new file location will be pushing updates to it, when and if they become available. If the shared file is restricted, then the system will capture the user identification code, block the user from further accessing the system, re-encrypt all restricted access system files stored locally on the user's device(s), and notify the sponsoring university of the breach.

Integrity of the system is paramount. Each member university will be responsible for monitoring and auditing its users. Additionally, a member university Charter will be created. The Charter will be responsible for establishing a constitution for membership and use of the fabric. Both the constitution and its enforcement must be transparent to all users and members. Individuals, or in extreme circumstances, organizations that do not abide by the constitution will have their access and/or membership revoked; however, their transactions and data will remain on the fabric with obvious markings across the document denoting the revoked membership and a broad categorization as to why. Including a broad categorization of why a membership has been revoked will better allow consumers to weigh the validity and reliability of the data. See Figure 8.

[181] Text steganography is the ability to hide secret information within a text document. Usually this type of steganography is limited on the amount of information that can be hidden, since text files are generally very small. Additionally, text steganography can be limited by the lack of redundancy within the file as compared to digital images or video.

Khan Muhammad et al., "A Novel Image Steganographic Approach for Hiding Text in Color Images Using HSI Color Model," arXiv Preprint arXiv:1503.00388 (2015), 2.

Figure 8. Charter Functions



This diagram lists how different Charter functions would be divided among key Charter groups.

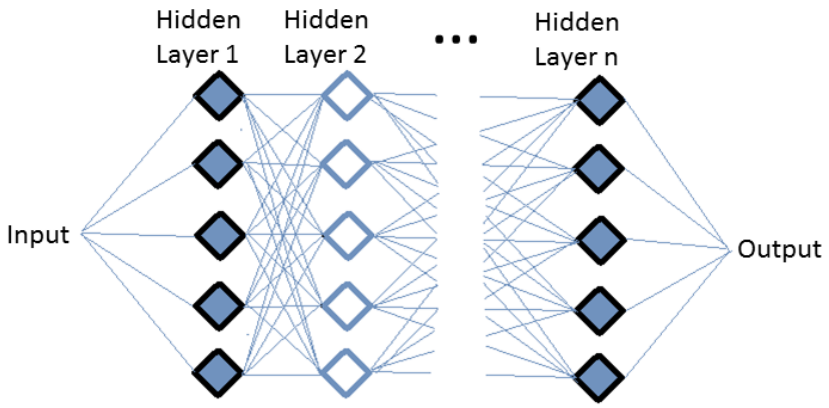
D. DATA UNIQUENESS

The final parameter of the thesis question is data uniqueness. Current technologies evaluated in Chapter V included Turnitin, iThenticate, and the Google Search Engine. Today's technology determines word strings within a document as compared to those documents available online or within the software's database, but the technology cannot determine whether the voice, intent, or any raw data within a document is original work. Artificial intelligence and machine learning may provide the solution to this problem. Currently used in business decision making and financial auditing, artificial intelligence is able to determine human intent and validate numerical data.[182] Figure 9 depicts how

[182] Mark van Rijmenam, "How Artificial Intelligence Will Change Corporate Governance," LinkedIn, December 12, 2017, https://www.linkedin.com/pulse/how-artificial-intelligence-change-corporate-mark-van-rijmenam; Julia Kokina and Thomas H. Davenport, "The Emergence of Artificial Intelligence: How Automation Is Changing Auditing," *Journal of Emerging Technologies in Accounting* 14, no. 1 (March 2017): 116–17.

artificial intelligence works. By gathering vast amounts of information to compare the input value, or in this case, the research being submitted to the fabric, artificial intelligence can be used to determine authenticity.

Figure 9. Artificial Intelligence Architecture



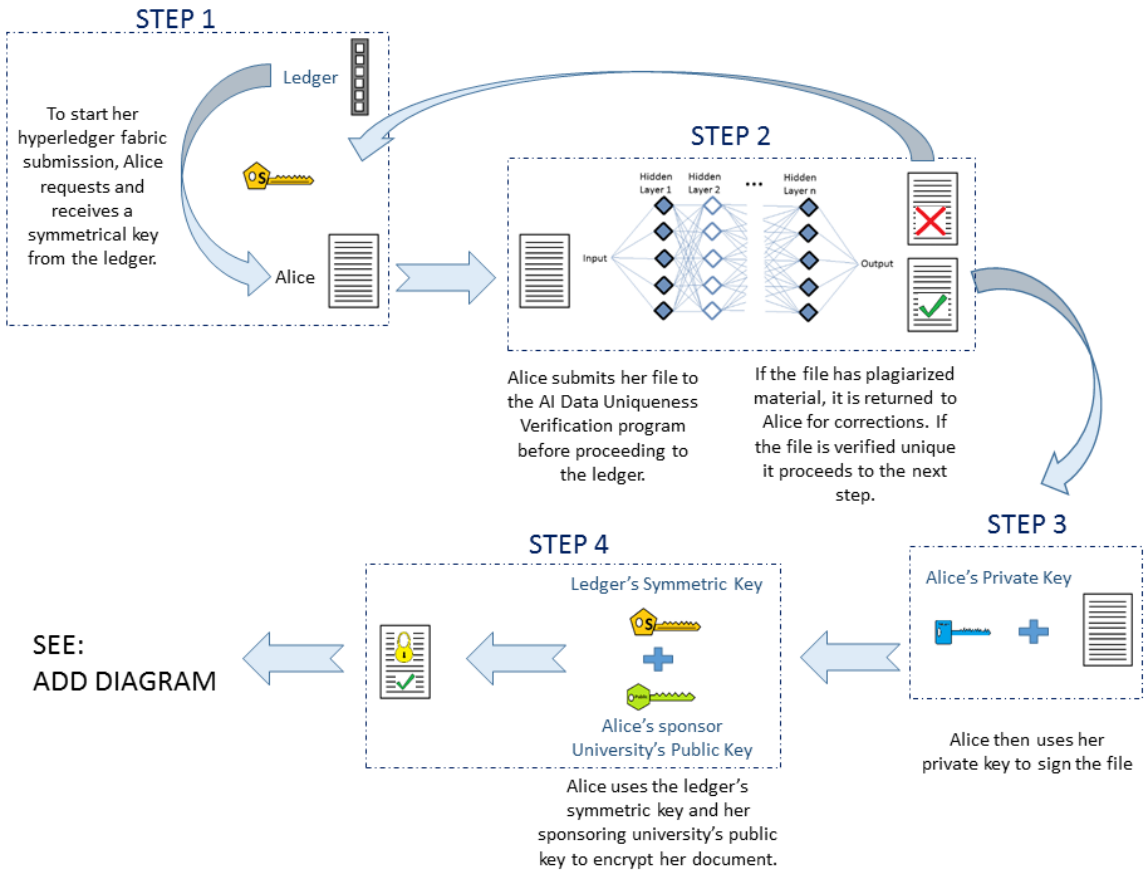
For the purpose of verifying data uniqueness, the multitude of hidden layers would compare all known research publications and files for comparison to the input value. If none of the values modeled a defined numerical value, then the work would be verified as unique. However, if the input value was too similar to another work, the artificial intelligence system would also need to be trained to measure amounts of ambiguity or plagiarism and make a determination whether it is repeated work, similar work, or, indeed, plagiarized work.

Combining these two capabilities, artificial intelligence could be created to verify the uniqueness of data and documents. Current word-string comparison technologies, like Turnitin and iThenticate, are limited to English language word-string comparisons.[183] Although the services are advertised to compare published works against a multitude of languages, according to each website both of these programs first translate the published works into English for comparison. This inability to work with the author's native language, combined with the inability to compare numerical data, figures, and charts makes these technologies unable to compare research files globally. Artificial intelligence, however, would be capable of comparing research files in their original language. Yet using artificial intelligence for data uniqueness will require significant data accumulation with programming and weighting for machine training. Once the system is operational, it will also require automating machine learning as more and more material is made available. Though this is an unusual use for artificial intelligence, the rapid advancement in the field and decreasing cost of storage for hidden layer data accumulation and training make this an excellent solution for determining data and publication uniqueness.

Figures 10 through 12 provide a visual representation of how these technologies would come together to provide the solution.

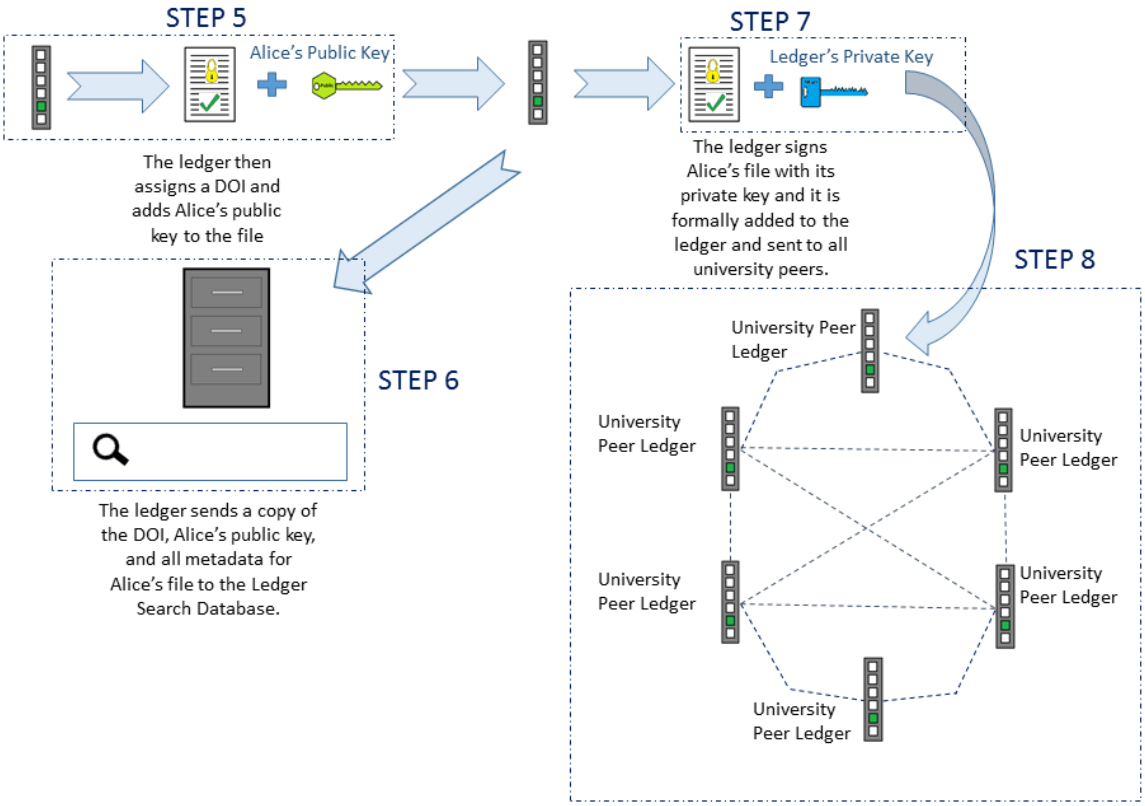
[183] Turnitin, "Turnitin: Technology to Improve Student Writing"; "Plagiarism Detection Software | IThenticate."

Figure 10. How to Send a File to the Ledger



In the diagram, Alice is a researcher who is submitting research files to the ledger. To start the transaction, Alice requests a copy of the ledger's symmetrical key from the ledger which is required to submit files to the ledger (Step 1). Before committing her submission to the ledger, Alice runs her submission through the artificial intelligence data uniqueness verification program. If it is verified as unique, it continues to Step 3. If it is not verified as unique it is sent back with a report to Alice for corrections (Step 2). Alice signs the file with her private key (Step 3), then uses the ledger's symmetric key and her sponsoring university's public key to encrypt the file so it can be transmitted to the ledger (Step 4).

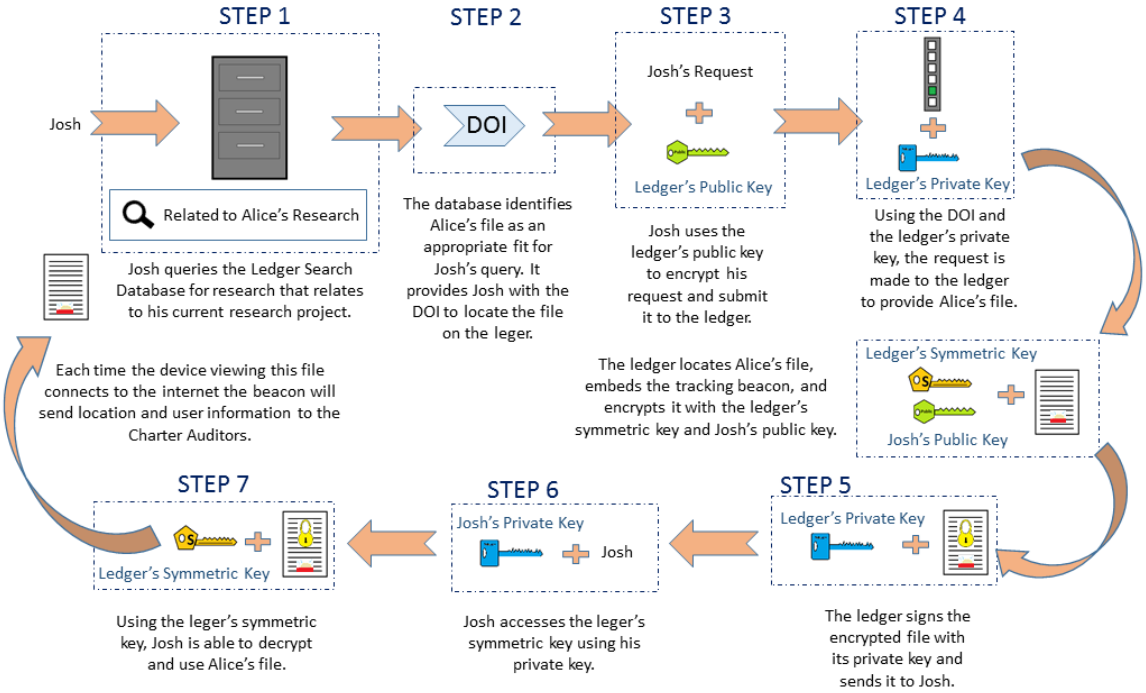
Figure 11. How to Add a File to the Ledger



Having been verified as unique, Alice's file is added to the ledger. When Alice's file is added to the ledger, the ledger generates a DOI to ensure each file has a unique identifier. The DOI is then combined with Alice's public key allowing both the file and the author to be located easily on the fabric (Step 5). The ledger then sends the DOI, file metadata, and Alice's public key information to the Ledger Search Database (Step 6). Alice's full file, DOI, public key are then encrypted for ledger addition with the ledger's private key (Step 7). Once added to the ledger, a copy of Alice's file is added to each peer copy of the ledger (Step 8).

Retrieving files from the fabric is a multi-step process as well. Figure 12 depicts each step of file gathering using the proposed solution's technologies.

Figure 12. How to Get a File from the Ledger



Josh is an end user of the solution. He is beginning a research project, but wants to ensure he has the latest information, and potentially find a collaborator interested in his research topic. To begin his search, Josh queries the Ledger Search Database for key phrases associated with his research. Alice's file is returned as a match for Josh's query. To access Alice's file, Josh uses the ledger's public key to formally request a copy of Alice's file, and the request is registered on the ledger as a transaction. To process Josh's transaction, the ledger uses its private key to decrypt Josh's request. Then after retrieving Alice's file, the ledger uses its symmetric key and Josh's public key to encrypt Alice's file. At this same time the ledger embeds the beacon into the file so that the file's transmission outside the ledger can be tracked later. To send the file to Josh and complete the transaction, the ledger must sign the file with its private key. The file is sent to Josh who uses his private key to access the ledger's symmetric key. Then using the ledger's symmetric key, Josh is able to decrypt the file and determine the usefulness of Alice's file. If Josh transmits Alice's file after downloading it, the beacon embedded in the file will alert the Charter's Auditors. The Auditors will open an investigation, and if the file was stolen or mishandled, it is muted or deleted from all fraudulent devices. Furthermore, the user, in this case Josh, may have his access to the ledger and database revoked.

Though this proposed solution theoretically answers all facets of the thesis question, certain limitations must be considered.

- Creating the Charter of universities will take time and resources from already budget-restricted universities.
- Establishing membership and user accounts, generating keys, and perform audit capabilities will require many information technology and administrative professionals to establish and maintain the systems.
- Participating universities will need to reallocate current server space or purchase new equipment to adequately store the ledger. Universities may also need to reallocate computing power and energy resources to support the new technology as well.
- Developing and deploying beacon technology will take time and money for research and development. This delay would prevent the initial solution roll out from being capable of meeting all prescribed requirements.
- Developing and deploying the artificial intelligence capabilities for determining data uniqueness will also require take time and money for research and development while delaying the full solution.
- Risking the security of the system if any user of the system loses his or her keys.

Though the proposed solution technologies have limitations, the benefits outweigh the challenges. The proposed solution is the only solution that addresses all prescribed needs for university research as defined in the thesis question. Developing a phased roll-out of the solution would reduce initial costs and allow the foundational technology of the solution to be well established before combining it with the nascent tracking and uniqueness verification technologies required for the full solution. Also, once created, these technologies have application in a multitude of other economic and business sectors. If successfully deployed and marketed, these alternative applications could generate revenue, partnerships, and prestige for member universities.

THIS PAGE INTENTIONALLY LEFT BLANK

VII. CONCLUSION

Innovation does not occur in a vacuum. As Steven Johnson writes, "Good ideas may not want to be free, but they do want to connect, fuse, recombine. They want to reinvent themselves by crossing conceptual borders. They want to complete each other as much as they want to compete."[184] Innovators must collaborate. The greatest minds in the world must be able to work together to solve the world's most daunting problems. Facilitating on-demand global intellectual summits or collaboration colliders will make the world a better place, if done correctly. Achieving this on a daily basis will require a new digital collaboration and sharing environment. This environment will allow research universities openly and with trust share verified unique data that is both immutable and ultimately trackable. What are the next steps to make this environment a reality? First, by examining currently technology's ability to meet the define needs. Second, by evaluating the identified technologies against the ideal environment as defined by the thesis question. Third, proposing a solution that will meet the ideal environment, and finally, proposing future projects to bring the environment from theory to reality.

Though this technology can help many different sectors, including the government and private industry, the ideal test-market for this new technology is the academic research setting. Universities have a need to share information. For financial, legal, and prestige reasons, research universities are an ideal market for this new technology to succeed. In addition to being generators of invention and innovation, universities also have highly intelligent workforces and understand the value of open information sharing. As discussed previously in the problem statement, university research, when used as intended, has the potential to improve life via gene therapies and replacement organs, and increasing nutrition and food security globally.

Maintaining the safety and security of sensitive and potentially dangerous information while sharing it productively requires better technology than exists today. As

[184] Johnson, *Good Ideas*, 22.

examined in this thesis, existing technologies cannot meet the needs of researchers collaborating globally today. See Table 7.

Table 7. Compiled Technology Evaluations

	Open Data Sharing	Data Uniqueness Verification	Data User Trust	Data Immutability	Data Identification and Traceability
Data Sharing					
ResearchGate	No	Some	Some	No	Some
Academia.edu	No	No	No	No	Some
arXiv	Yes	Some	Some	No	Some
Immutability and Trust					
Advanced encryption standard	No	No	Yes	No	Some
Blockchain	Some	No	Yes	Yes	Some
IBM Hyperledger	Some	No	Yes	Yes	Some
Identification and Traceability					
Digital Object Identifier (DOI)	N/A	No	No	No	Some
Persistent Universal Resource Locator (PURL)	N/A	No	No	No	Some
Internet Standard Serial Number (ISSN)	N/A	No	No	No	Some
Data Uniqueness					
Turnitin	No	Some	N/A	No	N/A
iThenticate	No	Some	N/A	No	N/A
Google Search Engine	Yes	Some	N/A	No	N/A

Table Key:

Yes	The site does offer solutions for the defined parameter
Some	The site offers a partial solution for the defined parameter
No	The site does not offer any solutions for the defined parameter
N/A	Not applicable based on the services provided by the site

Though existing technologies cannot create an open, trusted sharing environment of verified unique data that is immutable and trackable, they can provide a foundation from which to build new technology. The solution applications proposed in this thesis can hypothetically meet all the prescribed needs of the research question. Unfortunately, the proposed technologies also have drawbacks. Future researchers should further explore how to mitigate the challenges to realizing the solution proposed in this thesis. Future research must also be done to create the artificial intelligence uniqueness verification tool and the beacon tracking technology. These tools, once created, have a multitude of applications beyond this initial solution, and will be academically and financially fruitful.

Universities, researchers, and homeland security experts must pursue a solution, similar to the one described in this thesis, to protect our universities' sensitive research data, our country's health from bioengineered diseases, and our nation's security from threats posed by maliciously misused research data. This thesis is merely the start of the conversation; it is now up to university administrators, academic research professionals, and homeland security experts to find and realize the solution: an open and trusted sharing environment where unique data and ideas can be traceably shared without fear of deletion by nation-states or other malicious actors.

LIST OF REFERENCES

Al Hasib, Abdullah, and Abdul Ahsan Md. Mahmudul Haque. "A Comparative Study of the Performance and Security Issues of AES and RSA Cryptography." *Convergence and Hybrid Information Technology, 2008. ICCIT '08. Third International Conference On* 2 (2008): 505–10. https://doi.org/10.1109/ICCIT.2008.179.

Anhlan, Darisuren, Norbert Grundmann, Wojciech Makalowski, Stephan Ludwig, and Christoph Scholtissek. "Origin of the 1918 Pandemic H1N1 Influenza A Virus as Studied by Codon Usage Patterns and Phylogenetic Analysis." *RNA* 17, no. 1 (2011): 64–73. http://doi.org/10.126/ma.2395211.

Anthes, Gary. "Estonia: A Model for e-Government." *Communications of the ACM* 58, no. 6 (May 2015): 18–20. https://doi.org/10.1145/2754951.

Biswas, Kamamnashis, and Vallipuram Muthukkumarasamy. "Securing Smart Cities Using Blockchain Technology." In *IEEE 14th International Conference on Smart City* (2016): 1392–93. https://doi.org/10.1109/HPCC-SmartCity-DSS.2016.0198.1392.

Bradbury, Dan. "In Blocks We Trust (Bitcoin Security)." *Engineering Technology* 10, no. 2 (March 2015): 68–71. https://doi.org/10.1049/et.2015.0208.

Centers for Disease Control and Prevention. "2009 HiNi Pandemic (Hinipdm09 Virus)." Last updated November 2, 2017. http://www.cdc.gov/flu/pandemic-resources/basics/past-pandemics.html

Cheng, Ming-Yu, Jessica Sze-Yin Ho, and Pei Mey Lau. "Knowledge Sharing in Academic Institutions: A Study of Multimedia University Malaysia." *Electronic Journal of Knowledge Management* 7 (2009): 313–24.

Department of Homeland Security. *The National Strategy to Secure Cyberspace*. Washington, DC: Department of Homeland Security, 2003. https://www.us-cert.gov/sites/default/files/publications/cyberspace_strategy.pdf.

———. *Risks and Threats of Cryptocurrencies*. Falls Church, VA: Homeland Security Studies & Analysis Institute, 2014. https://www.anser.org/docs/reports/RP14-01.03.03-02_Cryptocurrencies%20508_31Dec2014.pdf.

Dworkin, Morris J., Elaine B. Baker, James R. Nechvatal, James Foti, Lawrence, E. Bassham, E. Roback, and James F. Dray, Jr. *Advanced Encryption Standard (AES)*. Federal Information Processing Standards Publication 197. Gaithersburg, MD: National Institute of Standards and Technology, 2001. https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.197.pdf.

Fayer, Stella, Alan Lacey, and Audrey Watson. “STEM Occupations: Past, Present, and Future.” Bureau of Labor Statistics, January 2017. https://www.bls.gov/spotlight/2017/science-technology-engineering-and-mathematics-stem-occupations-past-present-and-future/home.htm.

Federal Bureau of Investigation (FBI). “Health Care Systems and Medical Devices at Risk for Increased Cyber Intrusions for Financial Gains.” April 8, 2014. http://www.aha.org/content/14/140408--fbipin-healthsyscyberintrud.pdf.

Fogarty, Timothy J., and Donald V. Saftner. “Academic Department Prestige: A New Measure Based on the Doctoral Student Labor Market.” *Research in Higher Education* 34, no. 4 (August 1, 1993): 427–49. https://doi.org/10.1007/BF00991853.

Hyperledger. “Hyperledger Fabric Explainer.” YouTube video. April 28, 2017. https://www.youtube.com/watch?v=js3Zjxbo8TM.

International DOI Foundation. *DOI Handbook*. Wilmington, DE: The Corporation Trust Company, 2016. https://www.doi.org/hb.html.

ISSN. *ISSN Manual*. Paris: ISSN InterNational Centre, 2015. http://www.issn.org/understanding-the-issn/assignment-rules/issn-manual/.

Johnson, Steven. *Where Good Ideas Come from: The Natural History of Innovation.* New York: Riverhead Books, 2010.

Kokina, Julia, and Thomas H. Davenport. “The Emergence of Artificial Intelligence: How Automation Is Changing Auditing.” *Journal of Emerging Technologies in Accounting* 14, no. 1 (March 2017): 116–17.

Lane, Julia, Victoria Stodden, Stefan Bender, and Helen Nissenbaum. *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. New York: Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781107590205.

Marchany, Randy. *Higher Education: Open and Secure?* North Bethesda, MD: SANS Institute, 2014. https://jp.trendmicro.com/cloud-content/us/pdfs/business/articles/sans_higher_education_open_and_secure_research_study_trend_micro_edition_final.pdf.

Margetts, Helen, and Andre Naumann. “Government as a Platform: What Can Estonia Show the World?” Research paper, University of Oxford, 2017. https://www.politics.ox.ac.uk/publications/government-as-a-platform-what-can-estonia-show-the-world.html.

McDaniel, Thomas R. “Rethinking Scholarly Publication for Tenure.” In *Faculty Promotion and Tenure: Eight Ways to Improve the Tenure Review Process at Your Institution*, 13–14. Madison, WI: Magna, 2012. http://www.jsums.edu/academicaffairs/files/2012/08/Tenure-and-Promotion.pdf?x19771.

Min, Xinping, Qingzhong Li, Lei Liu, and Lizhen Cui. “A Permissioned Blockchain Framework for Supporting Instant Transaction and Dynamic Block Size.” In *2016 IEEE Trustcom/BigDataSE/ISPA* (2016): 90–96. https://doi.org/10.1109/TrustCom.2016.0050.

Muhammad, Khan, Jamil Ahmad, Haleem Farman, and Muhammad Zubair. “A Novel Image Steganographic Approach for Hiding Text in Color Images Using HSI Color Model.” arXiv Preprint arXiv:1503.00388 (2015).

Oppenheimer, J. Robert. “J. Robert Oppenheimer on Government Secrecy.” History.com video. Accessed August 3, 2017. http://www.history.com/topics/world-war-ii/world-war-ii-history/videos/j-robert-oppenheimer-on-government-secrecy.

PCI Security Standards Council. *PCI DSS Quick Reference Guide: Understanding the Payment Card Industry Data Security Standard Version 2.0*. Wakefield, MA: PCI Security Standards Council, 2010. https://www.pcisecuritystandards.org/documents/PCI%20SSC%20Quick%20Reference%20Guide.pdf.

Peterson, Kevin, Rammohan Deeduvanu, Pradip Kanjamala, and Kelly Boles. “A Blockchain-Based Approach to Health Information Exchange Networks.” Research paper, Health Information Technology, 2016. https://www.healthit.gov/sites/default/files/12-55-blockchain-based-approach-final.pdf.

Priisalu, Jaan, and Rain Ottis. “Personal Control of Privacy and Data: Estonian Experience.” *Health and Technology* 7, no. 4 (December 1, 2017): 441–51. https://doi.org/10.1007/s12553-017-0195-1.

Quora.com. “How Does Bitcoin Blockchain Work and What Are the Rules behind it?” October 1, 2016. https://www.quora.com/How-does-Bitcoin-Blockchain-work-and-what-are-the-rules-behind-it.

Richards, Mitchell C. *AES: The Making of a New Encryption Standard*. North Bethesda, MD: SANS Institute, 2001. https://www.sans.org/reading-room/whitepapers/vpns/aes-making-encryption-standard-740.

Robert Half Technology. *2018 Salary Guide for Technology Professionals*. Menlo Park, CA: Robert Half Technology, 2017. https://www.roberthalf.com/sites/default/files/documents/2018_salary_guide_NA_technology_1.pdf.

Rycroft, Robert W. “Does Cooperation Absorb Complexity? Innovation Networks and the Speed and Spread of Complex Technological Innovation.” *Technological Forecasting and Social Change* 74, no. 5 (June 1, 2007): 565–78. https://doi.org/10.1016/j.techfore.2006.10.005.

Sankar, Lakshmi Siva, M. Sindhu, and M. Sethumadhavan. “Survey of Consensus Protocols on Blockchain Applications.” In *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)* (2017): 1–5.https://doi.org/10.1109/ICACCS.2017.8014672

Strukhoff, Roger. “How Hyperledger Fabric Delivers Security to Enterprise Blockchain.” *Altoros* (blog), November 14, 2016. https://www.altoros.com/blog/how-hyperledger-fabric-delivers-security-to-enterprise-blockchain/.

United States Department of Agriculture Economic Research Service. “Ag and Food Sectors and the Economy.” October 18, 2017. https://www.ers.usda.gov/data-products/ag-and-food-statistics-charting-the-essentials/ag-and-food-sectors-and-the-economy.

Universities UK. *Cyber Security and Universities: Managing the Risk*. London: Universities UK, 2013. http://www.universitiesuk.ac.uk/policy-and-analysis/reports/Documents/2013/cyber-security-and-universities.pdf.

University of Alaska Fairbanks. “Publication & Peer Review.” August 25, 2015. http://www.uaf.edu/ori/responsible-conduct/peer-review/.

Vassil, Kristjan. “Estonian E-Government Ecosystem: Foundation, Applications, Outcomes.” Report, World Bank, 2016. http://pubdocs.worldbank.org/en/165711456838073531/WDR16-BP-Estonian-eGov-ecosystem-Vassil.pdf.

Vukolic, Marko. “Hyperledger Fabric: An Open-Source Distributed Operating System for Permissioned Blockchains.” Report, IBM Research, 2017. https://blockchain-summer.epfl.ch/talks/hyperledger-fabric-vukolic.pdf

Wiley, David. “Open Source, Openness, and Higher Education.” *Innovate: Journal of Online Education* 3, no. 1 (October 2006). https://www.learntechlib.org/p/104321/.

Yeh, Quey-Jen, and Arthur Jung-Ting Chang. “Threats and Countermeasures for Information System Security: A Cross-Industry Study.” *Information & Management* 44, no. 5 (July 2007): 480–91. https://doi.org/10.1016/j.im.2007.05.003.

Zheng, Zibin, Shaoan Xie, Hong-Ning Dai, Xiangping Chen, and Huaimin Wang. “Blockchain Challenges and Opportunities: A Survey.” *International Journal of Web and Grid Services* (December 2017).

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California