guardtime

SEPTEMBER 2025

Disruptive Technologies for EO Data Provenance Study Report

ID: ESA-TRACE4EO-SR-0001
Version: 1.0
Status: Released
Date of Issue: 09/09/2025
Classification: ESA UNCLASSIFIED – For Official Use Only

GUARDTIME.COM

Contents

Document versioning 3
1. Introduction 4
1.1. Purpose and scope 4
1.2. Relevant requirements 4
1.3. Structure of the document 4
1.4. Reference documents 5
1.5 Acronyms 5
2. Disruptive Technologies for Data Provenance 8
2.1. Blockchain-based solutions 8
2.2. Quantum-Resistant Cryptographic Timestamping 18
2.3. Self-Verifiable Data Objects 22
2.4. Zero-Knowledge Proofs 27
2.5. Federated Learning and AI-powered anomaly detection 31
2.6. Confidential computing & homomorphic encryption 35
2.7. Decentralised Knowledge Graphs 39
2.8. Two-Dimensional Hash Trees for Image Certification 41
3. System Requirements 48
3.1. Business/User requirements 48
3.2. Functional requirements 52
4. Preliminary choice of technologies 59
5. Conclusion 61
6. References 62

Document versioning

Date (MM.YYYY)	Version	Author	Changes
06.2025	0.1	Tuuli Lõhmus	Initial draft
09.2025	1.0	Maili Kark	First release
09.2025	1.1	Maili Kark	Short general introductions to technologies added, referencing updated.

1. Introduction

1.1. Purpose and scope

The purpose of this Study Report is to provide an in-depth overview of the disruptive technologies considered relevant for ensuring EO data provenance and integrity. The Study Report describes each technology in detail and where possible, real-life use cases are included. In addition to the technology overview, the Study Report contains user and system requirements.

1.2. Relevant requirements

The following requirements from [RD-01] have been considered in the analysis provided in this document.

Requirement ID	Description
REQ-1	**The Contractor shall identify disruptive and innovative technologies, which are expected to dramatically change EO with respect to traditional methods in the short and medium term. [OBJ-1]**
REQ-2	The Contractor shall build a SW prototype for the selected topic. [OBJ-2]
REQ-3	The Contractor shall demonstrate the added-value of the developed methodologies both in terms of quality performance but also perspectives for user adoption [OBJ-3]
REQ-4	The Contractor shall engage with various communities, including EO, and digital technologies practitioners. [OBJ-4]

1.3. Structure of the document

The document consists of the following chapters:

- **Introduction** – Overview of the document's objectives, references and terms

- **Disruptive technologies for data provenance** - Description of disruptive technologies, referencing contemporary academic studies and real-world applications from recent years
- **System requirements** - User/business requirements and functional requirements derived from Use Case definition document
- **Preliminary choice of technologies** - Reasoning for the selection of previously surveyed technologies based on how the technologies are able to fulfil the system requirements
- **Conclusion** – Takeaway from the study report's

1.4. Reference documents

Reference	Document title	Document reference	Issue	Date
[RD-01]	Statement of Work. ESA Express Procurement Plus - EXPRO+. Disruptive Technologies for Decentralized Storage and Provenance of EO data	ESA-EOPΦL-SOW-2024-0471	1	20/01/2025
[RD-02]	Disruptive Technologies for EO Data Provenance. Use Case Definition	ESA-TRACE4EO-UC-0001	1.0	09/09/2025

1.5 Acronyms

Term	Definition
ABE	Attribute-Based Encryption
AI	Artificial Intelligence
API	Application Programming Interface
CCC	Confidential Computing Consortium
CPU	Central Processing Unit
DC	Data Capsule
DID	Decentralized Identifiers

DIF	Decentralized Identity Foundation
DKG	Decentralized Knowledge Graph
DLT	Distributed Ledger Technology
dMRV	decentralized Measurement, Reporting, and Verification
DOI	Digital Object Identifier
EBSI	European Blockchain Services Infrastructure
ECDSA	Elliptic Curve Digital Signature Algorithm
EO	Earth Observation
ERP	Enterprise Resource Planning
EU	European Union
FL	Federated Learning
GBBC	Global Blockchain Business Council
GDPR	General Data Protection Regulation
HE	Homomorphic Encryption
IdP	Identity Provider
IETF	Internet Engineering Task Force
IIoT	Industrial IoT
IoT	Internet of Things
IPFS	InterPlanetary File System
ISO	International Organization for Standardization
ISS	International Space System
LWE	Learning with Errors
M2M	Machine-to-machine
NASA	National Aeronautics and Space Administration
NIST	U.S. National Institute of Standards and Technology
OIDF	OpenID Foundation
OWL	Web Ontology Language
PCCS	Provenance and Context Content Standard
PKI	Public Key Infrastructure
PoA	Proof of Authority

PoS	Proof of Stake
PoW	Proof of Work
PQC	Post-quantum cryptography
RDF	Resource Description Framework
SEC	Space Edge Computing
SDK	Software Development Kit
SMPC	Secure Multi-Party Computation
SSI	Self-Sovereign Identity
SVDO	Self-Verifiable Data Object
TEE	Trusted Execution environment
TFF	TensorFlow Federated
TPM	Trusted Platform Module
URI	Uniform Resource Identifier
VC	Verifiable Credentials
VON	Verifiable Organisations Network
VP	Verifiable Presentation
W3C	World Wide Web Consortium
ZKP	Zero-Knowledge Proof
ZKRP	ZK Range Proof
ZKSM	ZK Set Membership
ZK-SNARK	Zero-Knowledge Succinct Non-Interactive Argument of Knowledge

2. Disruptive Technologies for Data Provenance

Given the transnational nature of Earth Observation (EO) data generation and processing—encompassing stakeholders operating under diverse legislative frameworks, hardware infrastructures, and network conditions—it is prudent to architect a system grounded in standardized, disruptive technologies that have demonstrated reliability in large-scale deployments. To ensure practical viability, the proposed solution must balance stringent requirements for data privacy, processing efficiency, and verifiability, while avoiding excessive architectural complexity. Given that many disruptive technologies address diverse and nuanced privacy and confidentiality requirements for both data and stakeholders, it is essential to clearly define these needs prior to selecting appropriate technologies for system architecture design.

2.1. Blockchain-based solutions

Blockchain technology—generally a decentralized digital ledger maintained across a network of computers—has seen increasing adoption across diverse sectors over the past decade. Its defining characteristics, e.g. transparency, security, decentralization, and verifiability—offer significant benefits to industries such as finance, healthcare, and supply chain management. In addition, blockchain provides a robust framework for ensuring the trustworthiness of data, including the tracking of Earth Observation (EO) data provenance.

The basic logic of blockchain involves participants submitting transactions that are validated by a network of computers (nodes). Once verified, transactions are grouped into blocks, each containing a cryptographic hash that links it to the previous block. This structure ensures immutability, making recorded data resistant to tampering or deletion. Because the ledger is distributed across multiple nodes rather than controlled by a single authority, blockchain eliminates reliance on centralized intermediaries and protects against single points of failure or potential manipulation.

Blockchains can be classified according to their governance model: public blockchains are open and transparent to all users; private blockchains are controlled by a single organization; consortium blockchains are governed by a group of organizations; and hybrid blockchains combine features of both public and private systems. Regardless of type, every transaction is time-stamped and

permanently recorded, allowing participants with appropriate access rights to track data provenance from its origin. To ensure integrity, data entities are cryptographically hashed, while unique block identifiers guarantee continuity across the chain.

Furthermore, smart contracts—programs stored in blockchain, that self-execute when predefined conditions are fulfilled by participants—enable automation of processes, accelerating data exchange and enforcing predefined rules regarding who can view, modify, or contribute provenance data. Encrypted records of transactions are then shared with participants making any intermediate parties unnecessary.

However, blockchain technology has its challenges, often caused by the same features that contribute to the security and trust in blockchains. As in some blockchains every node stores and validates each transaction, scalability becomes a problem. Consensus mechanisms needed to validate each block are energy-intensive. The cryptographic puzzles solved in the blockchain to secure and validate Proof-of-Work are slow by design to ensure security, but the same feature also causes delays in transaction processing. Each block can only hold a limited amount of data, that may not be sufficient for some use cases. While nothing can be deleted from the blockchain and each node stores entire blockchain history, the data storage is ever growing and running full nodes becomes even harder. As public blockchains are transparent by design, it is not suitable for storing confidential or private information without encryption. While blockchain can be private or public, a careful analysis must be conducted to find the best trade-off for each solution's pros and cons.

There are several initiatives to find solutions for the challenges that using a blockchain brings. A research referenced here[1] on food and pharmaceutical food chains conducts a systematic literature review on blockchain applications in food and pharmaceutical supply chains. Deriving from the analysis of current blockchain configurations, a general framework is proposed to foster research and support practitioners with a reference design suited for the food and pharmaceutical blockchains are distinguished into public, consortium, and private blockchains.

Public blockchains allow everyone to join, participate, and validate blocks. They are often permissionless but could also require permission. This would support the idea that each party can read the data but not add new data without any

[1] https://www.sciencedirect.com/science/article/pii/S0924224425000287

control. In contrast to them, access to private blockchains is controlled by a single entity. This single entity decides on participation, consensus, and responsibilities regarding the maintenance of the blockchain. Consortium blockchains are similar to private blockchains but controlled by a consortium of entities. Additionally, hybrid blockchains are a mix of private and public blockchains. They usually consist of two blockchains that interact: a public chain to manage participants' identities and a private chain to store confidential data. Nevertheless, data could also be stored confidentially in public blockchains without providing access to all participants, e.g., by encrypting the data, i.e., the data is transformed to a cipher which cannot be read without access to a key.

According to the analysis, the majority of both supply chains used permissioned networks, containing 59.0% of applications in food supply chains and 74.4% in pharmaceutical supply chains. Further, 2.6% and 10.3% of applications are permissionless networks in food supply chains and pharmaceutical supply chains, respectively. 1 application in food supply chains and 2 applications in pharmaceutical supply chains (5.1% and 2.6%) followed a hybrid approach, combining a permissionless and a permissioned network. Private blockchains are applied most frequently in both supply chains. The examined years 2018-2022 showed the overall decrease in use of private blockchains, while the application of public blockchains increased from 0.0% in 2018 to 22.2% in 2021. Consortium blockchains and hybrid approaches also showed an increasing trend in the later years.

Private blockchain network types were selected primarily due to high security and confidentiality requirements but public blockchains were used to incorporate public participation in the model to allow all members to check and verify the data.[2]

A variety of blockchain platforms exist, each providing the infrastructure, services, and development tools necessary for the implementation of blockchain-based solutions. Prominent examples include Ethereum, Hyperledger Fabric, Solana, Corda, and Polygon, among others. These platforms offer diverse features and configurations, enabling their application across a wide range of use cases.

Regarding blockchain platforms, the most used platform is Ethereum, with 50.0% and 38.5% of applications in food supply chains and pharmaceutical supply chains, respectively. The second largest share of applications deployed Hyperledger Fabric at 22.5% in the food supply chain and 25.6% in the pharmaceutical supply

[2] https://www.sciencedirect.com/science/article/pii/S0924224425000287

chains. Further, 5.0% and 5.1% applied Hyperledger Sawtooth, and 5.1% Hyperledger Besu in the pharmaceutical supply chain only.

The reasoning to choose Ethereum and Hyperledger Fabric was the integrated deployment of smart contracts. In addition, Hyperledger Fabric provides high modularity, enabling the development of individual distributed ledger-based supply chains and the flexibility of implementing the models into practice. For similar reasons, Hyperledger Sawtooth was deployed, which was additionally supplemented by its novel consensus protocol that was particularly suitable for tiny devices and the written permissions and rules to access the supply chain.

The paper proposes a general framework for blockchain implementation, recommending a permissioned, consortium blockchain network using Hyperledger Fabric, PoA or similar consensus protocols, integrating combining off-chain data storage, using smart contracts to increase efficiency, tracing data and products, and providing real-time information.[3]

Another research[4] focuses on possibilities to make only specific datasets on blockchain accessible for each stakeholder, providing greater privacy for sensitive information. The paper proposes BSTProv, a blockchain-based system for secure and trustworthy decentralized data provenance sharing. It enables secure and trustworthy provenance sharing by partitioning a large provenance graph into multiple small subgraphs and embedding the encrypted subgraphs instead of raw subgraphs or their hash values into immutable blocks of a consortium blockchain; it enables decentralized and flexible authorization by allowing each peer to define appropriate permissions for selectively sharing some sets of subgraphs to specific requesters; and it enables efficient cross-domain provenance composition and tracing by maintaining a high-level dependency structure among provenance graphs from different domains in smart contracts, and by locally storing, decrypting, and composing subgraphs obtained from the blockchain. A prototype is implemented on top of an Ethereum-based consortium blockchain. With this proposed BSTProv system, a provenance owner can selectively share useful and insensitive provenance subgraphs to specific requesters to prevent sensitive information leakage. Furthermore, a provenance requester can obtain trustworthy provenance subgraphs from blockchain, and can decrypt and compose them into a partially complete but useful provenance graph for further cross-domain provenance tracing and trustworthy provenance-based data trustworthiness

[3] https://www.sciencedirect.com/science/article/pii/S0924224425000287
[4] https://www.mdpi.com/2079-9292/11/9/1489

enhancement. The proposed system enables different peers to define authorization policies for provenance subgraphs shared on-chain in an autonomous and flexible way.

According to W3C, data provenance records the evolutionary history of a data object, and is usually structured into an append-only directed acyclic graph (see below). A provenance graph contains three types of nodes: Entity, Activity, and Agent. Entity represents intermediate artifacts involved in producing a data object. Activity represents processes that manipulate entities. An agent represents persons or organizations that control activities. An edge is a directed dependency among nodes and is actually a provenance record with a timestamp that indicates an event from the past. An edge can be created and appended into a provenance graph whenever a concerned event happens. Along the direction of edges in a provenance graph, one can trace and scrutinize the historical nodes and events that directly or indirectly influenced a node. This process is called provenance tracing, which is the main usage of a provenance graph. Provenance tracing is the prerequisite for data trustworthiness verification, wrong data attribution, and accountability.

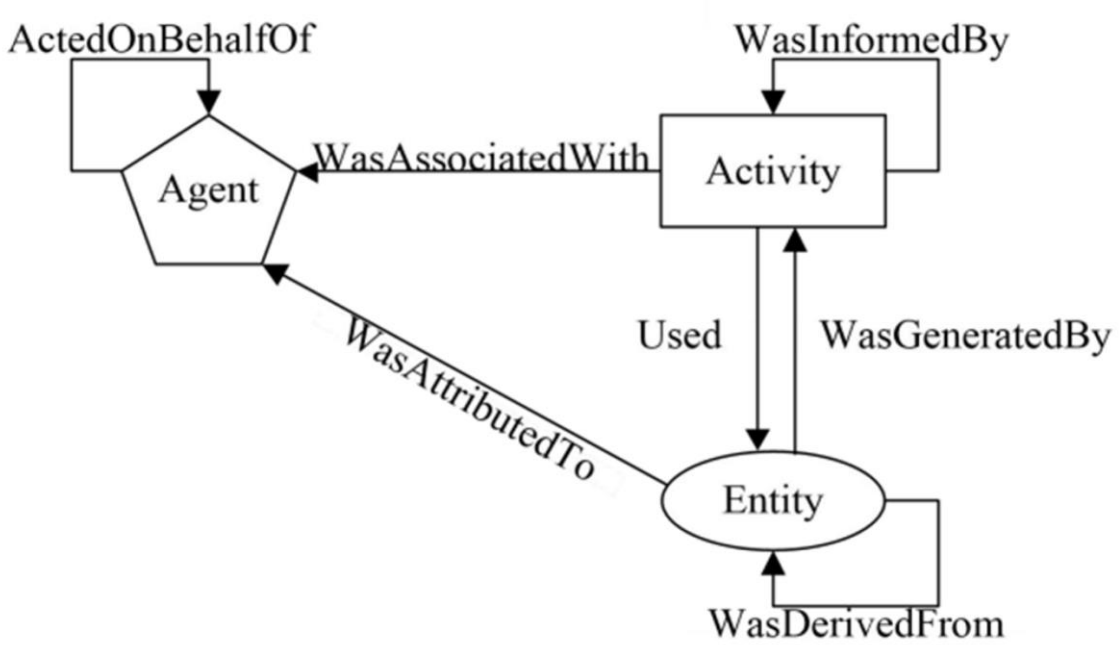


Figure 1 Core structure of a provenance data model[5]

[5] https://www.mdpi.com/2079-9292/11/9/1489 (Original of the graph from: Wood, G. Ethereum: A secure decentralized generalized transaction ledger. Ethereum Proj. Yellow Pap. 2014, 151, 1–32)

With the BSTProv system proposed by the research[6], a provenance owner first works locally to partition a provenance graph into multiple subgraphs, to encrypt each subgraph using a unique symmetrical encryption key, to embed each encrypted subgraph instead of its hash value into a blockchain transaction, and to submit the transactions onto the blockchain network. Second, the blockchain network spreads the transactions. Third, the miner elected according to some consensus protocol, such as PoS, packages these transactions into a block and spreads the block across the network. Fourth, provenance owners can independently define and submit policies onto the blockchain network to authorize a requester to access a set of subgraphs. Meanwhile, provenance requesters can independently define and submit requests onto the blockchain network to request a set of subgraphs.[7]

The comparison between two most used blockchain platforms for supply chains are compared in a research that reviews literature on food supply chains has also indicated Hyperledger and Ethereum as two most suitable blockchain platforms for food supply chains. Ethereum is an open-source distributed public blockchain network that uses Smart Contract technology to allow decentralized applications to be built on top of it. Hyperledger Fabric, an open-source project like Ethereum, is a widely accepted platform for enterprise blockchain platforms with its modular structure. Designed to develop enterprise-grade applications and professional solutions, the convenient, modular architecture uses "plug and play" components to adapt to many use cases.

The most important point of the referenced research is to create intersectoral cooperation by enabling blockchain-based projects to interact with each other. Hyperledger hosts several enterprise-grade blockchain-based software projects. Projects are designed by the developer community for vendors, organizations, service providers, and academics to build and deploy blockchain networks or commercial solutions. Each peer in Ethereum has a role, which means that whenever a transaction occurs, numerous nodes must participate in order for it to be completed, which causes scalability, privacy, and efficiency difficulties.

Hyperledger, on the other hand, is a distributed ledger technology (DLT) that does not require each peer in the network to be informed in order to complete a transaction. The anonymity of users within the system is one of the most emphasized issues in crypto money projects. However, this is not always required.

[6] https://www.mdpi.com/2079-9292/11/9/1489
[7] https://www.mdpi.com/2079-9292/11/9/1489

Keeping data on a public network and making it accessible to everyone can cause issues in some projects. Hyperledger is a permissioned blockchain that uses an identity management module to enable authentication. For this reason, it can store some information specific to a certain user group by using Hyperledger due to the private structure.

Feature	Ethereum	Hyperledger
Confidentiality	Public blockchain	Private blockchain
Purpose	Client-side B2C apps	Enterprise-level B2B apps
Governance	Ethereum Developers	Linux Foundation
Participation	Anyone	Organisations having Certificate of Authorisation
Programming language	Solidity	Node, Go, Java
Consensus mechanism	Proof of Stake (since 2022, PoW before)	Byzantine Fault Tolerance
Speed of transactions	Lower	Higher
Cryptocurrency	Ether or Ethereum	-

Discussing the input data for blockchain, the research addresses its possible connection to Internet of Things (IoT), capable of monitoring and collecting additional information. Wireless communication technologies (such as Bluetooth and Wi-Fi) are used in the connection layer to transmit data between sensor nodes and relay nodes, while machine-to-machine (M2M) communication technologies are used to transmit data between relay nodes and specified IoT platforms. IoT development platforms are used to develop and manage applications at the application layer, and application programming interfaces are used to connect external systems and databases (APIs). It should also be incorporated with ERP for things like managing and controlling internal resources and expenses. In terms of decentralized control, data transparency, auditability, distributed information, decentralized consensus, and high security, blockchain may currently bridge the gap in IoT systems.

The research concludes that blockchain and IoT integration ensures secure, immutable data sharing across all stakeholders, reducing reliance on human input and improving traceability, efficiency, and trust. IoT devices collect real-time environmental and logistical data, while blockchain guarantees its integrity and

transparency. Smart contracts automate processes and enforce consistency without requiring trust between parties. Overall, this model enhances food safety, supports informed decision-making for producers and consumers, and improves operational and cost efficiency across the supply chain.[8]

Blockchain-based solutions in use in other domains

Finance/Supply Chain - MasterCard

MasterCard launched its first provenance use case using blockchain in 2021 in Zimbabwe.[9] In 2021, they used ConsenSys Quorum which is an open-source protocol layer that enables enterprises to leverage Ethereum for their private or public production blockchain applications.10 Since then, it has focused on building the Mastercard Multi-Token Network™ (MTN), a project with strategic partners focusing on crossborder payments and asset tokenisation as of 202511. MTN uses a private permissioned blockchain. The MTN aims to make transactions within digital asset and blockchain ecosystems more secure, scalable, and interoperable. Its 4 main pillars are:

- Enabling trusted identities and reinforcing compliance
- Enabling stable and regulated payment tokens
- Powering secure cross-chain interoperability
- Creating a comprehensive governance framework[12]

There are a number of MTN users for payments, e.g. Vodafone and Coadjute[13]. In addition, the Fresh Supply Co has partnered with MasterCard for tracking

[8] https://dergipark.org.tr/en/pub/ejosat/issue/70985/1131779

[9] https://www.mastercard.com/news/eemea/en/newsroom/press-releases/en/2021/june/e-livestock-global-launch-mastercard-blockchain-based-solution/

[10] https://www.mastercard.com/news/press/2021/april/partnership-with-consensys-supports-the-future-of-multi-blockchain-commerce/

[11] https://www.ethnews.com/mastercard-advances-blockchain-strategy-with-multi-token-network-to-bridge-crypto-and-traditional-finance/

[12] https://www.mastercard.com/news/media/5zmixdjy/unlocking-the-potential-of-digital-asset-innovation-building-a-mastercard-multi-token-network-1-1.pdf

[13] https://www.mastercard.com/us/en/news-and-trends/press/2024/may/the-flexibility-of-crypto-the-convenience-of-fiat-bringing-blockchain-to-banking.html

avocado supply chain in Queensland, Australia. For provenance solutions, MasterCards focus is on supply chains and commerce[14].

Blockchain-backed satellite images

Thales Alenia Space and 3IPK have deployed the first blockchain network in space as part of the IMAGIN-e demonstration mission (ISS Mounted Accessible Global Imaging Nod-e). This pioneering technology is already aboard the International Space Station (ISS), where it will be updated regularly by IMAGIN-e's innovative concept of operations implementing Space Edge Computing (SEC). Through Microsoft's Azure Space SDK (software development kit), the mission provides application developers with an infrastructure to deploy their own data processing applications directly in space. The application developed by Thales Alenia Space in Italy uses multiple AI models onboard satellites to rapidly analyze sensor data, such as detecting clouds or water bodies. It efficiently switches between models to balance speed and resource use, allowing the satellite to discard low-value data (e.g., cloudy images) and focus processing power on useful imagery.

The application developed by 3IPK synchronizes blockchain nodes between satellites and Earth, enabling real-time recording of Earth-observation data on the blockchain at the source or early processing stage. This enhances data authenticity and traceability, addressing the growing risk of AI-generated fake data.

In the near future, new AI-based applications—including terrain classification and anomaly detection—will be tested on the IMAGIN-e payload as part of the OrbitalAI challenge, led by Thales Alenia Space, Microsoft, and ESA's Φ-lab under the AI4EO program. These tools, developed by international teams and Thales' in-house experts, will be deployed in orbit to demonstrate real-time data processing.[15]

[14] https://www.mastercard.com/global/en/business/large-enterprise/mastercard-enterprise-partnerships/global-trade-freight-solutions.html

[15] https://www.thalesaleniaspace.com/en/news/thales-alenia-space-and-3ipk-deploy-first-blockchain-network-space-imagin-e-payload-aboard

Carbon credits - Coorest project

A piloting project, featured in Global Blockchain Business Council (GBBC)'s real world blockchain use cases handbook (2025), that uses satellite data and environmental monitoring to calculate carbon credits, leveraging blockchain technology and smart contract is the Coorest project. As traditional carbon credit registries have suffered from fragmentation, lack of transparency, and susceptibility to fraud and in response to these challenges, Coorest has embarked on building a blockchain-based registry to address the shortcomings of traditional registries. Coorest leverages blockchain and decentralized finance (DeFi) to improve transparency and trust in the voluntary carbon market by creating a fully on-chain registry for carbon offset projects. By standardizing processes like project registration, measurement, reporting, verification, and tokenization through decentralized Measurement, Reporting, and Verification (dMRV), Coorest aims to ensure that carbon credits are based on accurate and verifiable environmental data.

Coorest stores all project and MRV data on Filecoin, a decentralized storage network, which guarantees data immutability and transparency. Real-time environmental data, such as biomass changes, is sourced via Chainlink's Floodlight Oracle, and smart contracts on the Polygon blockchain automate the issuance and retirement of carbon tokens—preventing double counting and enabling full on-chain carbon accounting.

Despite the system's lean development and technical strengths, improvements in user experience and the inclusion of financial safeguards, such as insurance, are seen as crucial to scaling the platform and attracting outside investment.[16] Currently the project is piloting in Spain, Argentina, Angola and Kenya.[17]

[16] https://downloads.ctfassets.net/so75yocayyva/1QeBFAtvDyoHK6pJgG5ITU/63401e1dfda1d22e4a5172cc33cb424d/GBBC-s_101_Real-World_Blockchain_Use_Cases_Handbook_digital.pdf

[17] https://coorest.io/onboarding/

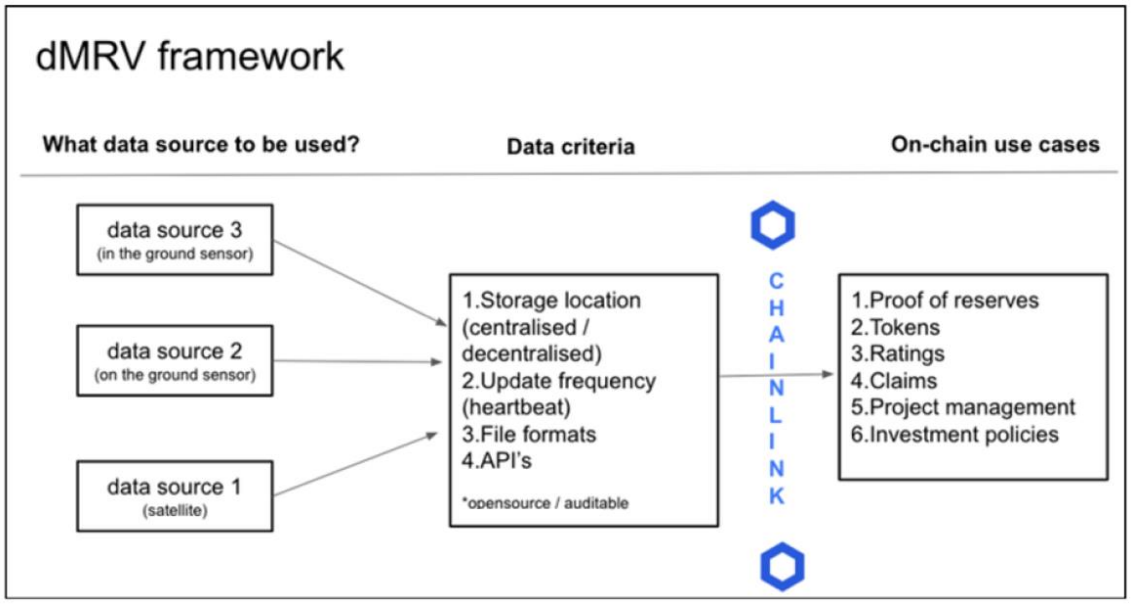


Figure 2 Digital measurement, reporting, and verification (dMRV) in COOREST solution[18]

2.2. Quantum-Resistant Cryptographic Timestamping

Timestamping, a method of providing cryptographically secured proofs ensuring that an entity of data existed or was processed at a certain point of time, is one of the most critical points in the EO data chain. The main components of timestamping are generating a digital fingerprint of data via hashing and then digitally signing it. Digital signature enables the authentication of a timestamp.

As quantum computers could break traditional digital signatures and undermine hash functions, quantum-resistant timestamping should be utilised to minimise the risks. There are several examples of quantum-resistant digital signatures, using specific cryptography. Some examples include lattice-based, hash-based, code-based and multivariate cryptography.

The U.S. National Institute of Standards and Technology (NIST) has initiated an evaluation for public-key digital signature algorithms for potential standardization and recommends two primary algorithms to be implemented for most use cases: CRYSTALS-KYBER (key-establishment) and CRYSTALS-Dilithium (digital signatures). In addition, the signature schemes FALCON (lattice-based) and

[18] https://downloads.ctfassets.net/so75yocayyva/1QeBFAtvDyoHK6pJgG5ITU/63401e1dfda1d22e4a5172cc33cb424d/GBBC-s_101_Real-World_Blockchain_Use_Cases_Handbook_digital.pdf

SPHINCS+ (hash-based) will also be standardized.[19] The CRYSTALS-Kyber algorithm has been chosen by NIST for general encryption, which is utilized when accessing secure websites. Kyber is based on lattice cryptography and benefits include the speed of operation and very minimal encryption keys that two parties can simply exchange. The algorithm offers a variety of hybrid settings and overall excellent performance on hardware and software. CRYSTALS-Dilithium is suggested by NIST as the main algorithm. The digital signature algorithm is based on the Fiat-Shamir paradigm. Dilithium is a signature method that has a strong theoretical security foundation and offers a simple implementation with great efficiency. The FALCON signature has the smallest bandwidth and is fast. The algorithm was chosen for its solid security and due to its low bandwidth, which may be necessary for applications that require smaller signatures. SPHINCS+ is slightly larger and slower than the other algorithms but is useful as a backup for one main reason: it does not use lattice-based cryptography. The algorithm uses a stateless hash-based signature scheme and offers a different arithmetic methodology than the other three selected algorithms.[20]

Facts	**FALCON**	**CRYSTALS-Dilithium**	**SPHINCS+**
Purpose	Digital signatures	Digital signatures	Digital signatures
Based on	Lattice-based	Lattice-based	Hash functions
Advantages	Scalable Simple Fast	Efficient	Efficient
Disadvantages	Smaller signature, still at its early stage	Difficult to implement, slow, bigger signature	Larger keys and signatures, slower

Comparison of quantum-resistant algorithms[21]

Quantum resistant hashing functions include widely recognised SHA-2 family (SHA-256 being the default for most applications), SHA-3 family (Keccak), BLAKE2 and BLACKE3 families and also SPHINX+ hash functions. The most used blockchain platforms for supply chains leverage quantum resistant hashing functions in their processing and signatures. Ethereum blockchain uses an instance of SHA-3 family, Keccak-256, to hash the transaction data, creating

[19] https://csrc.nist.gov/news/2022/pqc-candidates-to-be-standardized-and-round-4
[20] https://link.springer.com/article/10.1007/s11128-024-04272-6
[21] https://link.springer.com/article/10.1007/s11128-024-04272-6

Merkle trees, in smart contracts and cryptographic proofs. Hyperledger Fabric blockchain utilizes SHA-256 algorithm in consensus policies, ensuring transaction integrity and signatures, but SHA-3, while not used by default, can be integrated modularly by sophisticated modifications.

Some researchers have proposed post-quantum blockchains or modifications of current blockchains to tackle the quantum threat while other authors have suggested the implementation of quantum-safe blockchains.[22] Some examples are referenced below.

Research has been conducted to propose a condensed system architecture for a file transfer system that leverages post-quantum cryptography and blockchain technology. The architecture of the system integrates CRYSTALS-Kyber for encryption and CRYSTALS-Dilithium for digital signatures with an immutable blockchain ledger to provide an auditable, decentralized storage solution. Modular design comprises a Sender module for secure file encryption and signing, a central User Storage module that manages decryption, re-encryption, and blockchain logging, and a Requestor module for authenticated data retrieval. Considering performance insights, the findings on the proposed system demonstrated the practicality of post-quantum algorithms in low-resource environments.[23]

Another research focusing on performance explores integrating post-quantum signatures with the InterPlanetary File System (IPFS) in a blockchain system to enhance security and efficiency. By storing only hash values of signatures and public keys on the blockchain and placing their full content on IPFS, the system reduces blockchain load while ensuring long-term data storage. A comparison between NIST-recommended post-quantum signatures - Dilithium, FALCON, and SPHINCS+ - and Elliptic Curve Digital Signature Algorithm (ECDSA) in a Bitcoin exchange scenario demonstrates the approach's effectiveness in resisting quantum attacks while maintaining strong performance. The research concluded that the FALCON and Dilithium-based systems are recommended for applications that prioritize strong performance in key generation, signing, and verification times, especially when utilizing the suggested IPFS for managing large keys. If the IPFS is not preferred, then Falcon was found to be a suitable choice. Block capacity as well as the issue of quantum attack are both resolved this way. Overall,

[22] https://doi.org/10.1109/ACCESS.2020.2968985

[23] https://www.researchgate.net/publication/390671281_Development_of_a_Quantum-Resistant_File_Transfer_System_with_Blockchain_Audit_Trail

the proposed IPFS-based approach successfully reduces the signature/public key sizes for all signature schemes evaluated, improving the efficiency of blockchain systems.[24]

The ideal characteristics for quantum resistant blockchain schemes listed in a research focusing on post quantum blockchains are: small key sizes, small signature and hash length, fast execution, low computational complexity and low energy consumption. A list of schemes has been proposed for key and signature sizes, comparison of the average execution times and performance, Dilithium schemes being the fastest on average.[25]

Quantum-Resistant Timestamping use cases in other domains

Company	Action Taken	Result
DigiCert	Upgraded to quantum-resistant algorithms in 2022	99.9% uptime for time-stamping services (DigiCert says 100% since 1999)
Sectigo	Implemented blockchain-based time-stamping in 2021	Reduced fraud attempts by 75%
GlobalSign	Added AI-powered anomaly detection in 2023	Detected and prevented 150 potential breaches

Reference for table: https://www.scoredetect.com/blog/posts/time-stamping-digital-content-technical-guide

The European Commission has issued a recommendation to prepare a comprehensive strategy for the adoption of Post-Quantum Cryptography, to ensure a coordinated and synchronised transition among the different Member States.[26] The European Union's DIGITAL Europe Programme launched a quantum-resistant blockchain initiative in late 2023, coordinated via the European Blockchain Services Infrastructure (EBSI) to protect critical infrastructure. The project employs a hybrid approach, combining classical ECDSA signatures with

[24] https://www.mdpi.com/2227-7390/11/18/3947
[25] https://ieeexplore.ieee.org/document/10288193
[26] https://digital-strategy.ec.europa.eu/en/library/recommendation-coordinated-implementation-roadmap-transition-post-quantum-cryptography

lattice-based quantum-resistant signatures to ensure security against future quantum attacks while remaining compatible with existing systems. Initial results published in January 2024 demonstrated successful processing of over 100,000 transactions using these hybrid signatures. Although transaction sizes grew by about 300%, processing speeds stayed manageable through optimized implementation and network design. This deployment offers valuable insights into the practical challenges and solutions of implementing quantum-resistant blockchain technology.[27]

A global digital security company DigiCert provides public key infrastructure (PKI) and validation required for issuing digital certificates or TLS/SSL certificates, acting as a certificate authority (CA). DigiCert enables PQC algorithms in its private CA and Trust Lifecycle Manager environments— but not its public CAs. In these private environments, users can issue timestamping or signing certificates using ML-DSA (CRYSTALS-Dilithium) — versions MLDSA-44, 65, 87 and SLH-DSA (SPHINCS+) — SLHDSA-128, 192, 256.[28]

2.3. Self-Verifiable Data Objects

Self-Verifiable Data Objects (SVDOs) are data structures designed so that anyone can independently verify their authenticity and integrity without needing to trust a third party. Each data object contains a cryptographic proof or digital signature generated by a trusted source, which links the data to a secret key or specific validation criteria. Any alteration to the data invalidates this proof, making tampering easily detectable. Upon receiving the data object, a verifier can use the embedded proof along with public information (such as a public key or hash) to verify that the data is authentic and unchanged. Since the proof is embedded within the data itself, there is no need to consult a central authority or external database for validation.

Self-Verifiable Data Objects come with several challenges. Implementing them can be complex, requiring specialized cryptographic knowledge to correctly create and verify proofs. The process of generating and validating these proofs often introduces computational overhead, which may slow down processing, especially for large or frequent data exchanges. Additionally, effective key management is

[27] https://beyondthehype.terrencegatsby.com/blockchain/quantum-resistant-blockchain-protocols-preparing-for-the-future/

[28] https://docs.digicert.com/en/platform-overview/digicert-private-ca-services/ca-manager/post-quantum-cryptography--pqc--support.html?

crucial; if private keys are compromised or lost, the integrity of the data can be compromised. Embedding proofs also increases the size of data objects, impacting storage and transmission efficiency. Another challenge is handling updates or revocations, as SVDOs are self-contained and lack external references for easy modification. Finally, although SVDOs reduce reliance on intermediaries, they still depend on trusting the original source or key holder who creates the proofs.

Ensuring that the quality of dataset inputs can be trusted, research concerning scientific information, proposes mechanisms by which scientists and the organisations they represent can certify the authenticity of characteristics and provenance of any scientific datasets they publish so that secondary users can inspect and gain confidence in the qualities of data they source. Existing data sharing practices often lack mechanisms for verifying dataset quality while only a small subset of generated data is effectively utilized due to quality assessment challenges. To add a level of confidence to the quality of data in space projects, The Digital Object Identifier (DOI) System is widely used for identification and management of intellectual content and metadata. Different domains have their own specifications and requirements for metadata within DOI datasets - Earth Science Information Partners, founded by NASA, recommends the use of the Provenance and Context Content Standard (PCCS) matrix to perform identification, capturing and tracking of all metadata that can be used to validate the data and to facilitate efficient scientific reproducibility. The existing methods lack the ability to revoke or update the digital assets and can be vulnerable to certain types of attacks.

The research proposes using Self-Sovereign Identity (SSI) and verifiable credentials (VCs) as secure, decentralized mechanisms for certifying dataset quality and provenance. The approach aims to enhance confidence in shared datasets by issuing signed credentials linked to the data and its owners, supported by cryptographic methods like hash values for integrity.[29]

Self-Sovereign Identity (SSI) Technology comes with its own complexities and interoperability issues that are addressed in research focusing on document verification use case where the document must pass certain requirements. SSI identity management builds on core concepts of decentralization, distributed ledger technology and cryptography, and holds the potential to make the existing systems more secure, efficient, interoperable, and user-centric. In essence, SSI allows individuals to manage their own digital documents and credentials. It also

[29] https://arxiv.org/abs/2004.02796

allows organizations to define their own business processes and workflows without having to rely on third-parties and central authorities. The sole ownership over the ability to control the user's personal data is handed to the user in SSI. The SSI space is growing exponentially and there are different groups and standardization agencies working to develop new standards and protocols which could be the base of the SSI model, such as the Decentralized Identity Foundation (DIF), the European Blockchain Services Infrastructure (EBSI), the Internet Engineering Task Force (IETF), Sovrin, OASIS, the OpenID Foundation (OIDF), and the World Wide Web Consortium (W3C). The two fundamental base standards for self-sovereign identities are decentralized identifiers (DIDs) and verifiable credentials (VCs) by the W3C. The DID and VC standards propose a common data model for unique identifiers and credentials for self-sovereign identity solutions.

A DID is a new type of identifier that is decentralized, globally unique, resolvable, and cryptographically secure. It differs from other types of identifiers in that it can exist without the involvement of any certificate authorities, third parties, providers, or centralized identity registers. A DID is expressed as a URI scheme; an example of a DID is "did:example:12345". A DID is made up of three parts that are separated by colons. The "did" part of this DID represents that it is a DID, "example" is the DID method, and "12345" is the method-specific identifier that is used to distinguish this DID from other DIDs with the same method. The DID can be stored as a DID document on a blockchain or other storage system. The DID document contains all the information required to authenticate, authorize, or interact with the subject of the DID, such as the cryptographic material and public keys.

It may also contain service endpoints that describe a mechanism on how the DID subject is reached and establishes trusted communication. A DID document can be serialized in either the JSON or JSON-LD format. The location of where the document is stored depends on the used DID method and may be stored either on-chain, meaning that the document is written to a blockchain, or off-chain, meaning that the document is not written to the blockchain and stored somewhere else.

The VC data model was adopted as a standard in 2019 by the W3C. It is used to build trust between the involved parties in an SSI ecosystem, which often includes an issuer, holder, verifier, and verifiable data repository. A common procedure among the roles is that the issuer first offers the holder a VC. The credential is used by the issuer of a credential to make claims about a credential subject. A credential can hold many claims about a subject. The issuer is responsible for

creating and specifying the credential's content as well as the verification method. The verifiable credential is typically held by the credential subject, who then stores it in a digital wallet and is referred to as the holder of the credential. The credential subject can then present these claims to the verifier upon request to prove something about themselves. Lastly, the verifier then validates that the credential has not been tampered with and was issued by a trustworthy issuer, in addition to its own policy, to determine the credentials validity. The verification process can be carried out without involving the issuer directly.

A VC is made up of three main parts. First, there are the credential metadata, which consist of information that describes the credential such as credential type, who issued the credential, when it was issued, and when it expires, as well as a context property that permits an agreed-upon understanding of the credential and its structure and can be processed by JSON-LD. Second, the credential can contain statements about the credential subject in the form of one or more claims expressed as property–value pairs in the credential. Last but not least, it contains proof(s) that enable(s) the credential to be cryptographically verifiable using digital signatures. A verifiable credential can be serialized in JSON or JSON-LD, with the proof format being JWT or Linked Data.

The proposed architecture leveraging these techniques for a process of online loan application involves three key actors: issuers (such as identity providers, employers, and tax authorities), holders (loan applicants), and verifiers (banks). Holders request VCs from trusted issuers, which include tamper-evident claims and are linked to issuers' public DIDs stored on the blockchain for transparent verification. Holders store their credentials and private keys in digital wallets and, when applying for a loan, present verifiable presentations (VPs) containing the required credentials to the bank. The bank verifies the authenticity, validity, and status of these credentials by checking the public key information on the blockchain without contacting issuers directly, ensuring a secure, efficient, and decentralized verification process.[30]

Another example of SVDO is a research paper[31] introducing a Data Capsule (DC): self-contained, privacy-preserving data sharing model that lets individuals control and securely share their data without depending on centralised service providers. It combines Self-Sovereign Identity (SSI), Attribute-Based Encryption (ABE), and blockchain to enforce fine-grained, transparent access policies. A data capsule is

[30] https://www.mdpi.com/1424-8220/22/21/8408
[31] https://ieeexplore.ieee.org/document/9569788?

a digital container that includes encrypted personal data, an access policy and the cryptographic key for decryption (bound to recipient attributes). The capsule is self-verifiable and self-enforcing: it travels with all necessary metadata and rules for secure use. Attribute-Based Encryption ensures that only entities with specific attributes can decrypt data, Decentralized Identifiers (DIDs) control identity of users/service providers without central authority. The claims are cryptographically signed with Verifiable Credentials (VCs) and stored in Hyperledger Indy blockchain.

The case study presents the following flow for the tax submission system - service provider (SP) registers its identity (DID) and access policy on the blockchain. The user creates a Data Capsule with encrypted data and access policy, using ABE to encrypt the capsule, generating keys bound to SP's attributes. The capsule is then sent to the SP. SP uses its key to decrypt the capsule only if its attributes satisfy the embedded access policy.

The user needs to apply for a VC from a trusted identity provider (IdP), which allows them to authenticate themselves. The VC contains a set of tamper-evident claims about a user and metadata that cryptographically prove who issued it. When an organization issues a VC, they attach their public DID to that credential. That same public DID is also stored on the blockchain. When someone wants to verify the authenticity of the credential, they can check the public DID on the blockchain to see who issued it without having to contact the issuing party. The blockchain acts as a verifiable data registry.[32]

As shown in previously referenced research, SVDOs allow various types of business processes to be run while utilizing a specific instance or a combination of SVDO methods. As one of the main objectives of SVDOs is privacy, the requirements of end-users as well as all stakeholders must be specified with precision.

Self Verifiable Data Objects use cases in other domains

The Verifiable Organizations Network (VON) - a collaboration between the Government of British Columbia, the Government of Ontario, and the Government of Canada - has aimed to create a unified and trusted network of organizational data simplifying interactions between businesses and government

[32] https://ieeexplore.ieee.org/document/9569788?

services by utilizing decentralized identity technologies.[33] VON employs Verifiable Credentials (VCs) and Decentralized Identifiers (DIDs) to issue SVDOs. These credentials are cryptographically signed and can be independently verified without the need to contact the issuing authority.[34]

2.4. Zero-Knowledge Proofs

Zero-Knowledge Proofs (ZKPs) are a cryptographic technique that allows one party (the prover) to prove to another party (the verifier) that they know a value or a secret without revealing the value itself or any additional information. This is particularly valuable in situations where sensitive information must remain confidential to protect its privacy and security. For example, ZKPs can be valuable for secure authentication, privacy-preserving identity verification, access control systems and confidential data sharing, but for using ZKPs for chained verification of data that has been through multiple processing steps, they may be used in some combination with blockchain.

Research focusing on provenance trail of a product presents a solution for cases when participants must be verifiable, but the data might involve business secrets. To address this data privacy problem, data can be encrypted before sharing it on the ledger. However, encrypted data cannot be used in systems which use blockchain technology for instant verification of provenance. Hyperledger Fabric proposes a solution with its private channels but it is inflexible to support selective data protection by a peer and involves additional overheads. This work proposes privacy preservation and incentive enforcement mechanisms on permissioned blockchain based on Zero Knowledge Proofs (ZKPs) and commitment schemes. ZKPs help to prove possession of a secret without revealing the actual information whereas commitment schemes can be used to verify a committed secret later in time. Supply chain participants can provide ZKP proofs and get reciprocated by the committed incentive amounts for utilizing their resources. The blockchain can verify these proofs, initiate an off-chain payment mechanism and log the results in an immutable way.

The proposed PrivChain framework enhances data privacy in blockchain-enabled supply chains by allowing participants to share zero-knowledge proofs (ZKPs) instead of raw IoT data (e.g., location, temperature). These proofs verify product provenance and quality without revealing sensitive information. A smart contract

[33] https://www.lfdecentralizedtrust.org/blog/2019/03/11/reducing-government-red-tape-british-columbia-creates-new-business-identity-model-with-hyperledger-indy?
[34] https://www.weboftrust.org/project/orgbook_bc-79?

on the blockchain automates proof verification and rewards participants with payments for verified claims. This method protects trade privacy while maintaining trust and accountability. The framework is implemented on Hyperledger Fabric and is also applicable to sectors like energy trading and healthcare. Key contributions include:

- A ZKP-based privacy-preserving mechanism for supply chains.
- Smart contracts for automated proof verification and incentives.
- Demonstrated low overhead of proof verification in real-world deployment.

ZKPs are cryptographic methods in which a prover can attest to a verifier that some secret information is true without revealing any details about the secret. Depending on the requirement of interactions between the prover and the verifier, ZKPs are interactive or non-interactive. For a distributed network such as blockchain, interactive ZKPs are impractical due to the communication overhead for verification. Zero-Knowledge Succinct Non-Interactive Argument of Knowledge (ZK-SNARKs) are an efficient non-interactive variant of ZKPs. In addition, the proofs generated are of constant size despite the complexity of proof generation. The steps involve a key setup by a trusted third party, generating a key for proving and a key for verifying; proof generation by Prover, using a proving key, common input and a secret; and verification, where verifier uses the proof, common input and verification key to confirm whether the proof is correct.

Variants of ZKPs - such as ZK Set Membership (ZKSM) or ZK Range Proofs (ZKRPs) can have more specific properties. The ZKSM scheme allows anyone to prove that a secret lies within a given set. ZK-Range Proofs (ZKRPs) are a special instance of ZKSM, which prove that a certain secret number lies within a specific numeric range. The ZKSM algorithms can be easily adapted to carry out ZKRP computation as it is specific to numeric intervals as compared to ZKSM. When compared to a generic ZKP, ZKRP is proven to reduce the computation overhead of proof verification by an order of 10. Hence, ZKRP can be applied to a range of decentralised applications where numerical data needs to be proven without actually revealing it such as account balances (e-finance), sensor readings (IoT), electronic auctions (e-commerce), rating/trust scores (supply chains), and distribution of sales (energy trading). Bulletproofs are short and efficient ZKRPs designed specifically for blockchain. They have a performance advantage as the size of proofs are logarithmic to the input size and they do not require a trusted setup. Note that, the choice of the range proof methods differs based on the computation power of the prover and the verifier. Bulletproofs with a smaller

proof size reduce the transaction size when compared to other variants of ZKRP such as signature based methods. However, signature based methods incur lower verification complexity only when the range size is less than 102 bits.

For the proposed privacy preservation scheme, the non-interactive signature based range proofs were chosen based on 1) reliance on permissioned blockchains for trusted setup and verification whereby the computation overhead for setup and verification is delegated to the blockchain provider; 2) the lower proof verification complexity of non-interactive signature based range proofs given the high transaction send rate in typical supply chain use cases[35].

Another research addressing the need for restricting access to some parts of product's provenance, a blockchain-based model for the supply chain is presented that provides privacy and traceability with efficient contamination tracing and enables the consumer to be convinced about a product to be unaffected by identified faulty upstream processes (contamination) without access to the whole history of the product. The solution uses zero-knowledge proofs and cryptographic accumulations to provide these guarantees. Provenance tracking is leveraged though constant sized on-chain commitment for supply chain documents. This enables stakeholders to confirm if a product uses an identified contaminated lot of products in its processing, taking into account that a product lot can be split and merged in supply chain operations.

The paper has defined a set of document types that have specific requirements based on their purpose and position in the supply chain. For example, an Entry document type represents a product's entry into the supply chain and this document is not previously linked to any other document. Other document types are Exit, Ship, Merge, Split and Process. This forms a chain of documents through smart contracts and a product's provenance can be established tracking back all its upstream documents.

The main objectives of that solution are privacy of business-sensitive data in the documents, unlinkability for obscuring the relationships of the supply chain participants from the publicly available data; traceability of the products and contamination tracking to confirm if a product has used an identified contaminated lot in its processing. Hyperledger Fabric has been chosen as blockchain technology for its scalability and practicality.

[35] https://arxiv.org/abs/2104.13964

The client's authentication is required to be anonymous, to achieve unlinkability. The solution uses Hyperledger Fabric's Identity Mixer certificates with Zero-Knowledge proofs to prove knowledge of signature. It is assumed that the organisations introducing items to blockchain are trusted, a participant "consumes" a document only if the initial document's hash matches the delivery it received off-chain and that exit participants only transfer the product to end customers if the document hash is consistent with off-chain receiving document.[36]

Research addressing large data amounts to be recorded in blockchain discusses ZKPs as verifiable proofs for energy consumption and proposes a solution that includes a robust event logging mechanism to ensure the public verifiability of energy certificates. As the energy space is truly data intensive, especially with production and consumption certificates at hourly or 15 min intervals and to address this issue it is proposed to first decentralize the registries for granular certification by distributing the workload to a size where the services can run reliably; secondly to devise a strategy for each registry to be auditable and verifiable by means of applying Blockchain and Merkle trees validation (cf. Use of Blockchain) and ensuring confidentiality using Zero-Knowledge proofs and commitments (cf. Confidential Information) and thirdly to ensure that registries have a common public key infrastructure to interact with each other and validate their soundness so they can interact with stateful certificates—so the system can scale into adoption. The research focuses on the second principle. To ensure transparency in verifying certificates (such as energy claims), each registry keeps an event log recording all certificate changes. The validity of a certificate or energy claim can be checked against this log. Blockchain is used to guarantee the log's immutability, transparency, and public verifiability. However, to protect privacy and competitive information, details about energy amounts are kept confidential using commitments, and zero-knowledge proofs verify the consistency of this hidden information across the log. The paper also discusses options for managing scalability and costs associated with the blockchain solution, proposing the optimization of the size of Merkle trees as a strategy.[37]

[36] https://www.slideshare.net/slideshow/enabling-privacy-andtraceabilityinsupplychainsusingblockchainandzeroknowledgeproofs/250842942

[37] https://energyinformatics.springeropen.com/articles/10.1186/s42162-023-00283-2

Zero-knowledge proofs use cases in other domains

ING Bank, based in Amsterdam, has used Zero-Knowledge proofs (ZKPs) to enhance privacy in financial services by allowing clients to prove facts about their data without revealing the data itself. Zero-Knowledge Range Proofs (ZKRP) solution was introduced in 2017, that allowed for example a mortgage applicant to prove their income is within a required range without disclosing the exact amount. Similarly, the ZKRP could prove that a payment amount is within a limit, without showing the exact amount.[38] The developed ZKFlow consensus protocol enables private transactions on Corda blockchain for arbitrary smart contracts using Zero Knowledge Proofs, featuring a range of privacy for transactions.[39]

Another innovation for data privacy by zero-knowledge proofs was introduced in 2018, when ING developed open-source Zero-Knowledge Set Membership (ZKSM) that allowed other types of data besides numerical to be checked. For instance, banks could validate that a new client lives in a country that belongs to the European Union, without revealing the country.[40] In 2019 ING Bank implemented Bulletproofs for more efficient, scalable ZKP solutions, improving both privacy and performance.[41]

2.5. Federated Learning and AI-powered anomaly detection

Federated learning and AI-powered anomaly detection will be discussed together in this paragraph as these technologies can provide greater benefits when some of the features or techniques are combined. While federated learning provides a decentralized framework for data sharing, enabling privacy preservation at the same time, AI-anomaly detection finds and flags irregularities in the training data.

Federated Learning (FL) presents a novel technology for harnessing distributed computational and data resources in a privacy-preserving manner. It enables multiple decentralized entities—such as devices, institutions, or data custodians—to collaboratively train a machine learning model without transferring raw data to

[38] https://www.ing.com/Newsroom/News/Blockchain-transactions-just-got-a-whole-lot-safer.htm

[39] https://github.com/ing-bank/zkflow?tab=readme-ov-file

[40] https://www.ing.com/Newsroom/News/Blockchain-innovation-improves-data-privacy-for-clients.htm?

[41] https://www.ledgerinsights.com/ing-blockchain-privacy-bulletproofs-zkp/

a central repository. In this framework, each participating client performs local training on its private dataset and shares only model parameters or gradients with a central aggregator. The server then applies a federated aggregation algorithm (e.g., Federated Averaging) to update the global model, which is subsequently redistributed to clients for the next training iteration. This decentralized approach significantly mitigates privacy and data governance concerns, especially in contexts governed by strict data protection regulations (e.g., GDPR).

The federated learning workflow typically involves a coordinating server initializing a global model, which is dispatched to client devices or nodes. Each client trains the model locally and submits encrypted or compressed updates to the server. After secure aggregation, the global model is refined and redistributed. This cyclical process continues until the model reaches a satisfactory performance threshold, at which point it can be deployed either centrally or locally, depending on the application.

AI-powered anomaly detection systems complement federated learning by employing machine learning algorithms to identify deviations from established patterns in data streams. These models are trained to discern "normal" behaviour using historical datasets. Supervised approaches require labelled data indicating both normal and anomalous cases, whereas unsupervised methods rely on pattern recognition to detect statistical outliers without labelled examples. Semi-supervised techniques offer a hybrid strategy, typically training on predominantly normal data and flagging deviations as potential anomalies. Once trained, the AI model continuously analyses incoming data and flags irregularities based on its learned representations.

Several software frameworks support the integration of federated learning and AI-powered anomaly detection. Tools such as TensorFlow Federated (TFF), developed by Google, enable the training of custom anomaly detection models (e.g., autoencoders or recurrent neural networks) across decentralized data silos. PySyft, maintained by OpenMined, provides a privacy-preserving infrastructure for distributed AI, supporting secure multi-party computation (SMPC) and differential privacy. FedML and Flower are also widely used for scalable federated learning experiments, offering cross-platform support and modular APIs. These can be combined with geospatial data processing libraries like GDAL or Rasterio for effective integration with satellite image workflows.

In the context of Earth Observation (EO), federated learning offers a scalable and privacy-respecting solution for training machine learning models on satellite

imagery sourced from multiple ground stations or organizations. Satellite data—often sensitive and voluminous—can remain locally stored while contributing to a shared global model. This is especially beneficial for applications such as crop classification, environmental monitoring, and disaster detection, where data sovereignty and real-time responsiveness are critical. By avoiding the need for centralized image collection, FL preserves both data privacy and bandwidth.

However, federated learning is not without limitations. The frequent transmission of model updates between clients and servers can lead to increased communication overhead, particularly in bandwidth-constrained or heterogeneous environments. Variability in data distributions, device capabilities, and data quality across clients can adversely affect model convergence and overall accuracy. Model updates themselves can inadvertently leak sensitive information if not adequately protected via differential privacy or homomorphic encryption. FL systems also require complex orchestration infrastructure to ensure synchronization, scalability, and robustness against client dropouts or adversarial behaviour.

As the majority of FL works manage supervised tasks where clients' training sets are labelled. To leverage the enormous unlabelled data on distributed edge devices, a study has been conducted addressing the problem of anomaly detection in decentralized settings, studying the possibilities of unsupervised FL. The proposed method groups clients into communities in a preprocessing phase, each having similar majority (i.e., inlier) patterns. Subsequently, each community of clients trains the same anomaly detection model in a federated fashion. The resulting model is then shared and used to detect anomalies within the clients of the same community that joined the corresponding federated process. Experiments show that this method is robust, and it can detect communities consistent with the ideal partitioning in which groups of clients having the same inlier patterns are known. The performance is significantly better than those in which clients train models exclusively on local data and comparable with federated models of ideal communities' partition.[42]

Leveraging IoT for gathering and registering data and then using the data as input, a study proposes a clustered federated learning architecture for unsupervised network intrusion detection in large-scale IoT/Industrial IoT (IIoT) environments. It highlights that IoT devices often encounter diverse and distributed intrusion data that cannot be centralized due to privacy and logistical constraints. The study

[42] https://www.computer.org/csdl/proceedings-article/msn/2022/645700a495/1LUtH1eeYak

addresses this by enabling local model training on each node's data and aggregating updates through federated learning clusters. This approach enhances detection accuracy across distributed datasets while preserving data ownership and confidentiality.[43]

Effective deployment of federated learning and AI-powered anomaly detection in EO systems hinges on the availability of high-quality, verifiable data. Ensuring the provenance, integrity, and authenticity of satellite imagery is therefore a foundational requirement. Technologies such as blockchain, Zero-Knowledge proofs (ZKPs), and digital watermarking can be leveraged to establish data trustworthiness before federated learning is applied. In this context, federated learning and AI-driven analytics represent the next logical progression in the architectural evolution of secure, privacy-aware, and intelligence-driven EO platforms.

Federated Learning and AI-powered anomaly detection use cases in other domains

An FL model, called "EXAM" (EMR CXR AI Model) was developed in collaboration between Massachusetts General Brigham Hospital and NVIDIA, resulting in large, diverse federated learning initiative with 20 hospitals from around the world. Within weeks the global collaboration achieved a model with excellent prediction capabilities for the level of oxygen required by incoming patients. Using the NVIDIA Clara Federated Learning Framework, researchers at individual hospitals were able to use a chest X-ray, patient vitals and lab values to train a local model and share only a subset of model weights back with the global model in a privacy-preserving federated learning technique.[44]

By leveraging federated learning, the participating institutions would not have to transfer data to a central repository, but rather leverage a distributed data framework. The EXAM model was trained using a cohort of 16,148 cases, making it not only among the first FL models for COVID-19 but also a very large and multicontinent development project in clinically relevant AI. The locally trained models were compared with the global FL model on each client's test data. During the FL training task, each client site selects its best local model by tracking the model's performance on its local validation set. At the same time, the server

[43] https://www.sciencedirect.com/science/article/pii/S0167404823002092?via%3Dihub
[44] https://blogs.nvidia.com/blog/federated-learning-covid-oxygen-needs/?

determines the best global model based on the average validation scores sent from each client site to the server after each FL round. After FL training finished, the best local models and the best global model were automatically shared with all client sites and evaluated on their local test data. The stability of results was validated by repeating three runs of local and FL training on different randomized data splits.

In this study, federated learning facilitated rapid data science collaboration without data exchange and generated a model that generalized across heterogeneous, unharmonized datasets for prediction of clinical outcomes in patients with COVID-19.[45]

2.6. Confidential computing & homomorphic encryption

Confidential computing technology protects data while it's being processed. Data is normally encrypted when stored or transferred but becomes vulnerable when it's actively used in memory or CPU. Confidential computing uses hardware-based trusted execution environments (TEEs)—secure areas within a processor—to isolate data and code during processing, shielding it from the rest of the system. This means data remains encrypted and protected even while computations are running, preventing unauthorized access by cloud providers, admins, or malware.

While Confidential computing can fill a gap in secure data processing, the cons of Confidential computing include its reliance on specialised hardware but this is currently only available on selected CPUs (e.g., Intel SGX, AMD SEV) limiting deployment options and possibly increasing costs. Users need to trust the hardware manufacturer. Also, integrating Confidential computing into existing systems can be complex, needing developers to implement software to work within restricted secured enclaves. Due to encryption and limited resources, running workloads inside TEEs can cause performance issues.

Homomorphic encryption is a type of encryption that allows computations to be performed directly on encrypted data without needing to decrypt it first. The result of these computations remains encrypted, and when decrypted, it matches the result as if operations had been done on the raw data. This enables secure

[45] https://www.nature.com/articles/s41591-021-01506-3

data processing and sharing, letting third parties analyze or compute on sensitive data without ever exposing the actual information. This can be used in privacy-preserving Machine Learning, for example, as it enables training on encrypted data.

Homomorphic encryption schemes use various mathematical foundations and some homomorphic encryption schemes use lattice-based schemes that are believed to be secure against quantum attacks. While NIST has not yet standardized homomorphic encryption schemes, it has standardized and recommended foundational primitives (like LWE and NTRU) that most lattice-based homomorphic encryption schemes rely on.

In EU Horizon programs that fund research and innovation, homomorphic encryption plays a key role in projects focusing on privacy-preserving AI, secure data sharing, and digital sovereignty.[46] Also, the International Organization for Standardization (ISO) is working to standardize homomorphic encryption to ensure interoperability, security, and global trust in cryptographic systems.[47] Homomorphic encryption is not widely used yet due to several significant challenges. One of the primary barriers is its high computational overhead—operations on encrypted data are often hundreds of times slower than plaintext computations, making homomorphic encryption impractical for real-time or resource-constrained applications. Additionally, homomorphic encryption is complex to implement, requiring deep cryptographic expertise and lacking mature tooling. The technology is also use-case specific; while powerful for secure data aggregation or privacy-preserving machine learning, it is not well-suited for general-purpose encryption needs. Homomorphic encryption solutions typically require extensive system redesign and are not easily compatible with existing infrastructure or mainstream cloud platforms. However, as standards mature and tools improve, homomorphic encryption is expected to see broader adoption.

A critical review of Confidential computing technology and related technologies such as Trusted Execution Environments (TEEs), homomorphic encryption (HE), Trusted Platform Modules (TPMs), argues that the Confidential computing Consortium (CCC) definition and comparison of these technologies are ambiguous, incomplete, and sometimes conflicting, which can mislead users, create regulatory uncertainty, and make vendor comparisons unreliable. As key security properties—including data confidentiality, code integrity, attestation, and

[46] https://cordis.europa.eu/project/id/101070214?
[47] https://www.iso.org/standard/87638.html?

programmability—are analysed, the review emphasizes the need for formal, precise definitions and rigorous evaluation frameworks.[48]

Confidential computing and homomorphic encryption in other domains

Confidential computing in financial fraud detection

Financial fraud detection company based in New York, MonetaGo, addressed the need to prevent duplicate financing fraud schemes. Using Google Cloud Confidential computing solution, which leverages AMD Secure Encryption Virtualization on AMD EPYC™ CPUs, MonetaGo built Secure Finance, a platform that enables banks to share sensitive lending documents with MonetaGo, which authenticates the information and flags duplicates across multiple institutions in real time, while remaining in compliance with local regulations and without breaching privacy. This helps protect data in process, a crucial feature for MonetaGo as it processes sensitive data from multiple institutions.

To cover both security concerns while maintaining the performance, MonetaGo is using Confidential Google Kubernetes Engine (GKE) from Google Cloud, part of the Confidential computing platform. While data is currently stored with MonetaGo, which protects it at rest with a multi-key encryption process to ensure no single individual can breach security, the company is working on introducing Confidential Space into Secure Finance. Confidential Space is a service from Google Cloud that's integrated to Secure Finance and allows customers to store their own data, which will then be pulled into a permissioned confidential workload for advanced analysis. Financial institutions retain full control over their data with a greatly minimized risk of data breach, even as it is aggregated and analysed by MonetaGo.

Traditionally, while cloud data is encrypted while at rest or in transit, it is not encrypted while in use. Google Cloud Confidential computing fills this gap by encrypting data in use, a crucial feature for MonetaGo as it analyses and compares data across competitive institutions with adherence to strict regulations. To accomplish this, Google Cloud uses AMD Secure Encryption Virtualization on AMD EPYC™ CPUs. This hardware-accelerated memory encryption uses keys to

[48] https://cybersecurity.springeropen.com/articles/10.1186/s42400-023-00144-1

keep data in use encrypted and shielded against rootkits, rogue admins, or infrastructure vulnerabilities. These keys are held within the AMD Secure Processor on an EPYC™ CPU so not even Google can read them.

MonetaGo's Secure Finance system analyses financial data from collaborating institutions by generating digital fingerprints of documents using hashing and storing them in a global hash registry, which acts as a secure unified repository. This allows the system to detect duplicate submissions from different lenders without exposing confidential data and affected parties are notified without any confidential information being revealed. It also verifies document authenticity by comparing them with trusted data sources to identify fraud.[49]

Homomorphic encryption in ElectionGuard

While there are no widely deployed industry-scale enterprise systems using homomorphic encryption in production, there are several prototypes and pioneering applications leveraging this technology. Microsoft in collaboration with Galois has developed ElectionGuard - open-source cryptographic SDK, designed to enhance existing voting systems (like paper ballots, touchscreen voting, and scanners) with verifiable encrypted vote tracking. It operates in parallel with existing voting systems, allowing vote recoding and tabulation via secure, auditable, and privacy-preserving processes.

Each vote is encrypted with homomorphic encryption and given a unique identifier at the time it is cast. The voter is given a tracking code that lets them check that their vote goes through the system unchanged and ends up in the final tally, while independent auditors can verify the integrity of the election using publicly available encrypted data and cryptographic proofs. A quorum of trusted guardians collaboratively decrypts only the final tally, ensuring both vote privacy and tamper detection. The first pilot has already been successfully carried out in an election in Fulton, Wisconsin.[50]

[49] https://www.amd.com/content/dam/amd/en/documents/resources/case-studies/monetago-case-study.pdf
[50] https://news.microsoft.com/on-the-issues/2020/03/27/what-is-electionguard/

2.7. Decentralised Knowledge Graphs

Decentralized Knowledge Graphs (DKGs) work by distributing the creation, storage, and management of knowledge graphs across multiple independent nodes or participants, removing reliance on a central authority. At their core, knowledge graphs organize information as entities (nodes) and relationships (edges) using semantic web standards such as RDF (Resource Description Framework) or OWL (Web Ontology Language). In a decentralized setup, each participant can contribute data or knowledge triples, which are cryptographically signed to prove authenticity and origin.

To ensure trust and integrity, DKGs leverage decentralized identifiers (DIDs) and verifiable credentials. These allow nodes to authenticate contributors and validate the provenance of the data, making it tamper-evident and auditable. The underlying infrastructure often involves distributed ledger technology (e.g., blockchain) or peer-to-peer networks, which maintain immutable records of transactions and proofs related to data updates or contributions. This ledger acts as a verifiable registry of data provenance and ownership without exposing the entire dataset.

When querying a DKG, the process is often federated or distributed, meaning queries aggregate data from multiple nodes or shards while respecting access controls and privacy settings. Nodes collaborate by sharing subsets of the graph or responding to query requests with verifiable data. Cryptographic proofs and consensus mechanisms ensure that responses are trustworthy and consistent despite the absence of a central authority.

Moreover, smart contracts or programmable logic on the blockchain can enforce policies, automate data sharing agreements, or trigger updates, enhancing transparency and governance. Overall, DKGs combine semantic web principles with decentralized trust frameworks and cryptography to enable secure, transparent, and collaborative knowledge representation and reasoning across diverse participants.

Several tools and platforms support the development and use of decentralized knowledge graphs by combining blockchain, distributed storage, and decentralized identity technologies. For example, Ocean Protocol enables decentralized data sharing with built-in provenance and access control, while The Graph Protocol allows decentralized querying of blockchain data that can be

adapted for knowledge graph applications. IPFS provides distributed storage for the data objects underlying these graphs, and Hyperledger Indy focuses on decentralized identity management with verifiable credentials and DIDs to anchor entities within the graph. Additionally, protocols like KILT offer verifiable claims that enhance trust and provenance in the data. Although fully integrated decentralized knowledge graph platforms are still evolving, many projects leverage these interoperable tools and frameworks to build transparent, tamper-resistant semantic data networks.

A study on distributed knowledge graphs explores how they can be constructed using distributed ledger technologies in decentralized computing environments. It proposes a framework that integrates resource orchestration, information acquisition, and event streaming to enable dynamic knowledge graph creation. Distributed ledgers (e.g., blockchains) are used to maintain immutable logs of events and provenance, ensuring data integrity and trust. This system empowers decentralized ecosystems—such as edge computing networks—by enabling nodes to collaboratively ingest data, process events, and build verifiable knowledge graphs without central coordination.[51]

Concerning data privacy on multi-stakeholder, decentralized environments, a study proposes verifiable querying techniques, incorporating provenance, cryptographic proofs, and trusted decentralized protocols to ensure that both the source and integrity of query results can be independently audited. It is explored whether technical methods can be used by application providers to ensure transparent and responsible data processing. Focusing on personal data stores, especially within the Solid framework that uses knowledge graphs, the authors propose four key features to support auditable and trustworthy data usage: (1) Traces – logs of inputs, outputs, and code execution; (2) Policy validation – ensuring user-defined data policies are enforced; (3) Verification – using tamper-evident data and verifiable code; and (4) Anonymity – implementing measures to prevent unnecessary data exposure. These mechanisms would generate receipts to inform users when their data is queried, though the paper notes that translating such mechanisms into user-friendly formats is a complex, separate challenge.[52]

[51] https://dl.acm.org/doi/10.1145/3627673.3679644
[52] https://dl.acm.org/doi/fullHtml/10.1145/3543873.3587635

Decentralised knowledge graphs use cases in other domains

A digital infrastructure developed and maintained by TIB – Leibniz Information Centre for Science and Technology in Germany - aims to describe research papers in a structured manner and make them easier to find and compare, leveraging knowledge graph technology in a scientific field. Open Research Knowledge Graph (ORKG) has two main components: the back end, which contains the logic to handle requests by client applications and the front end through which users create, curate or explore scholarly knowledge. ORKG centres on the concept of a ResearchContribution, structuring key scholarly elements — ResearchProblem, ResearchMethod, and ResearchResult — into a machine-actionable format. Using a graph-based data model similar to RDF but enhanced with annotated edges and provenance metadata (e.g., creator and timestamp), ORKG captures scholarly knowledge in a detailed, traceable way. Users interact via a step-by-step UI to create these structured contributions, with advanced features like automated discovery of similar works to support efficient comparisons and literature reviews.[53]

2.8. Two-Dimensional Hash Trees for Image Certification

The following section on using two-dimensional hash trees for EO image certification is written by Ahto Buldas, who is also a full professor in the Centre for Digital Forensics and Cyber Security in Tallinn University of Technology. As known today, the two-dimensional hash trees have not previously been used in the manner described below.

Satellite produced images have to be certified for data provenance. Such images are often large, but sometimes, only a small sub-image of the full image is needed. The goal is to use a certification scheme so that a certificate for any sub-image:

- is easy to extract from the certificate of the full image
- is verifiable against the sub-image without using the full image

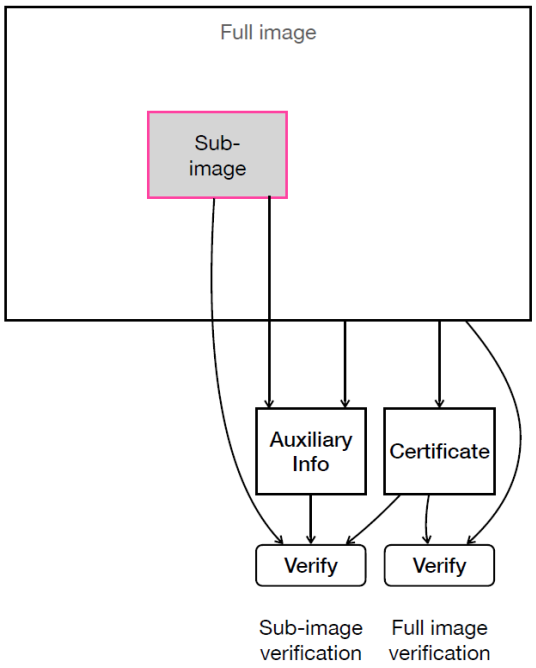
[53] https://arxiv.org/abs/2206.01439

Some auxiliary information (e.g. authentication paths) might be necessary to make the certificates of sub-images verifiable.

Merkle tree certification

Cryptographic hashing combined to digital signatures is a commonly used way of certification, so that:

• The full image is divided into 2^{2k} much smaller base-rectangles



• Merkle tree is used to compute the hash of the full image, so that the hashes of the base-rectangles are the leaves of the Merkle tree

• Signature algorithm is applied to the root hash

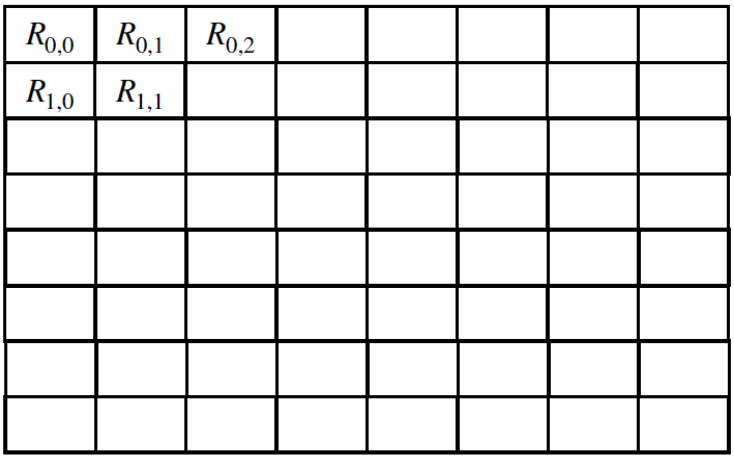
• The certificate of any base rectangle is an authentication path containing $\log_2 2^{2k}$ = 2k sibling hashes of the corresponding leaf

Merkle tree is computed layer by layer so that at each layer:

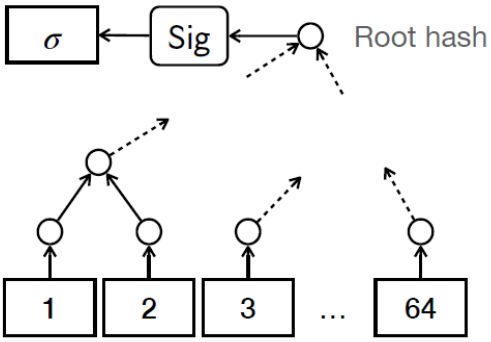
1. Remaining 2^{ℓ} hashes are paired by using a pairing rule (operator)

2. Pairs are hashed together to produce $2^{\ell-1}$ remaining hashes

These two steps are repeated until a single hash remains.



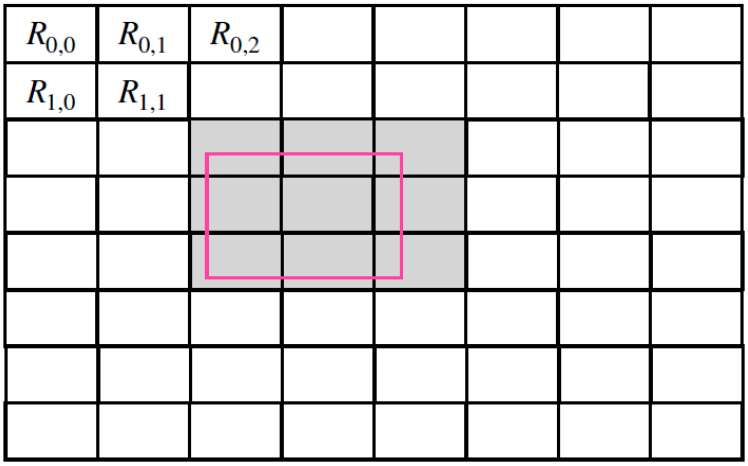
Full image split into $2^6 = 64$ base-rectangles ($k = 3$)



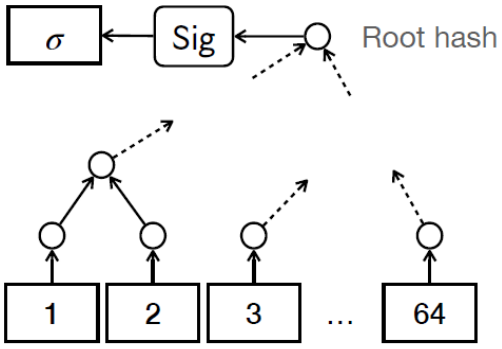
Base-rectangles are numbered and hashed via Merkle tree

Sub-image certificate

To certify a sub-image (pink rectangle in the Figure), we first extend the subimage to a rectangle of base-rectangles (grey area)



Full image split into $2^6 = 64$ base-rectangles ($k = 3$)



Base-rectangles are numbered and hashed via Merkle tree

The certificate for the extended sub-image contains the minimal set of hashes of the Merkle tree that enables to compute the root hash given the extended sub-image.

The size of the sub-image certificate depends on the pairing scheme.

Let the sub-image be of size m×m′ base-rectangles, where m is the vertical size (number of rows) and m′ is the horizontal size.

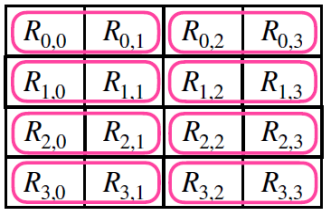
If a random hashing order is used, all m×m′ base-rectangles of the extended sub-image are randomly distributed on the leaves of the Merkle tree and that is why we may need m×m′ almost complete authentication paths (each containing 2*k* hashes) to certify the extended sub-image. Therefore, the certificate size is $O(k \cdot m \cdot m')$, i.e. the size is proportional to the area of the sub-image.

We will show that for a carefully chosen pairing scheme that uses two-dimensional hashing, we are able to reduce the size to $O(k \cdot m)$ or $O(k \cdot m')$.

2-d hash trees

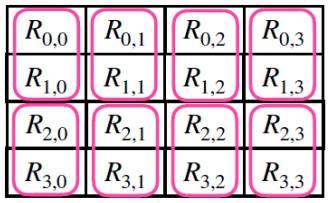
Combine two pairing operators:

H — horizontal pairing, that creates pairs $(R_{i,2j}, R_{i,2j+1})$



Horisontal pairing

V — vertical pairing, that creates pairs $(R_{2i,j}, R_{2i+1,j})$



Vertical pairing

If there are 2^k rows and 2^k columns, then both H and V must be used exactly k times.

Examples of total pairing rules:

V^kH^k — k rounds of H followed by k rounds of V

H^kV^k — k rounds of V followed by k rounds of H

$(HV)^k$ — alternating application of V and H (starting from V)

$(VH)^k$ — alternating application of V and H (starting from H)

Certificate size: interval certificates

Let $x_0, \ldots, x_{n-1}$ (where $n = 2^k$) be the leaf hashes of a Merkle tree.

Assume that for $0 \le i \le j < n$ we need to certify that $(x_i, x_{i+1}, \ldots, x_j)$ is a sub-interval of the leaf hashes.

The certificate for $(x_i, x_{i+1}, \ldots, x_j)$ consists of:

1. The left sibling hashes of x_i
2. The right sibling hashes of x_j

It can be shown that for any interval and $k > 0$ and $i < j$, the total number of sibling hashes in the certificate does not exceed $2(k - 1)$.

Indeed, we can apply induction on k. For $k = 1$ the statement clearly holes, because then $x_i = x_0$ and $x_{i+1} = x_1$ are the only leaves.

Assuming that $k \ge 2$ and the statement holds for $k - 1$, we observe two cases:

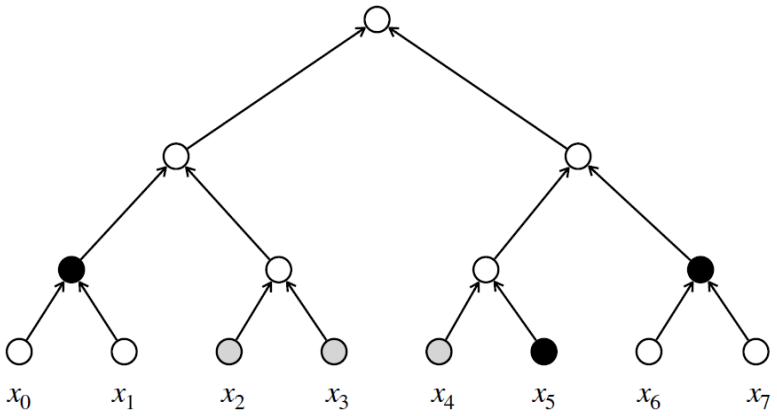
$j < n/2 = 2^{k-1}$ — in this case we need no more than

$2((k-1)-1)+1 = 2(k-1) - 1 < 2(k-1)$ hashes

$i < n/2 \le j$ — in this case we need no more than

$(k-1) + (k-1) = 2(k-1)$ hashes

Example with $n = 8$ and $k = 3$



Black dots — the sibling hashes of the certificate for (x_2, x_3, x_4)

Certificate size: for pairing scheme $(VH)^k$

Let the sub-image be of size $m \times m'$ base-rectangles, where m is the vertical size (number of rows) and m' is the horizontal size.

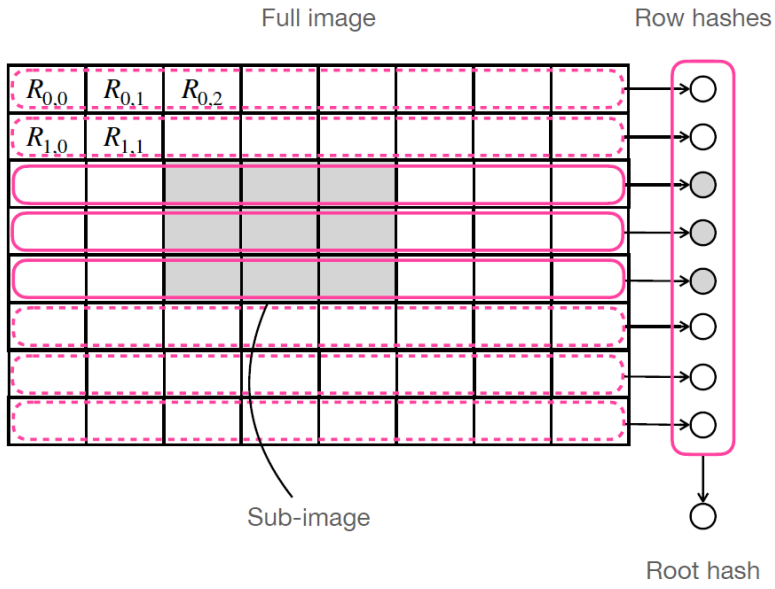
In the case of the pairing scheme $(VH)^k$ we first apply horizontal pairing to compute the 2^k row hashes, and then apply vertical pairing to compute the root hash from the row hashes.

For every row, the certificate size (relative to the corresponding row hash) does not exceed 2(k − 1).

In addition, we need up to 2(k − 1) hashes to extend the certificate for the total root hash.

Hence, the total number of hashes in the certificate does not exceed:

$$m \cdot 2(k-1) + 2(k-1) = 2(k-1)(m+1)$$



3. System Requirements

3.1. Business/User requirements

Requirement ID	**GT-DT-BUS-01**
Subject	System interface
Requirement	User is able to use a GUI or CLI tool to access the system
Explanation	User is able to use System for verifying both single data entity's provenance and batch data's provenance
Input	N/A
Output	Access to systems data query interface
Verification method	Inspection, Test

Requirement ID	**GT-DT-BUS-02**
Subject	Provenance verification
Requirement	User is able to verify the whole provenance trail of data entity in sufficient detail
Explanation	User is confirmed that every event in provenance trail meets the requirements of legitimate data modifications (e.g. the data modification is immutable, identifiable and authenticated). System enables access to both whole provenance trail and a verification certificate proving that the provenance trail was verified in a certain time.

Input	Data query
Output	Provenance record, verification certificate for queried data
Verification method	Inspection, Test

Requirement ID	**GT-DT-BUS-03**
Subject	Provenance authentication
Requirement	User is able to authenticate all events in data entity's lifecycle
Explanation	Authentication of the events enables the user to have confirmed information on the origin of data modifications, what organisation, process or model issued them.
Input	Data query
Output	Provenance record with authenticated data entries and events, verification certificate for queried data
Verification method	Inspection, Test

Requirement ID	**GT-DT-BUS-04**
Subject	Provenance components
Requirement	User is able to get confirmation what data entities were used to compile the output

Explanation	User is informed of every data entity, that has been used for compiling the end result (e.g. models, statistics, sensors)
Input	Data query
Output	Provenance record, verification certificate for queried data
Verification method	Test

Requirement ID	**GT-DT-BUS-05**
Subject	Provenance trail focus
Requirement	User is able to get confirmation if data input by specific organisation/generator was used to generate data output
Explanation	User can indicate, if e.g. a specific satellite produced the images used to train a model. This may be relevant, if an entity in data provenance trail (a specific satellite image, process, model, organisation) has been identified as defective, outdated or untrustworthy and its contributions to the ends result must be identified
Input	Data query on specific organisation/generator
Output	Confirmation/Rejection
Verification method	Test

Requirement ID	**GT-DT-BUS-06**
Subject	System product
Requirement	User is able to download data provenance record
Explanation	The user can pick a file format for record download suitable for their needs (human readable or machine readable)
Input	Data query
Output	Signed JSON, XML, txt, csv, pdf file presenting provenance record with all details, verification certificate
Verification method	Test

Requirement ID	**GT-DT-BUS-07**
Subject	User authentication
Requirement	User making modifications to data must authenticate themselves
Explanation	User making modifications to data (e.g. decisions on deleting irrelevant data) must be identifiable at least by organisation and role
Input	User credentials
Output	Access to system

Verification method	Inspection, Test

Requirement ID	**GT-DT-BUS-08**
Subject	System configuration
Requirement	Authenticated user is able to update the schema of data processing and initial data while ensuring backwards compatibility
Explanation	User inserting initial data, making the modifications or processing original data can update the schema of data and its modifications to ensure that the latest relevant information set about the data is captured. However, some attributes of the previous and updated schema must remain the same or be linked, so backwards compatibility is ensured.
Input	Schema configuration
Output	Updated configuration
Verification method	Test

3.2. Functional requirements

Requirement ID	**GT-DT-FUN-01**
Subject	System function

Requirement	The system shall capture, register and store all events in data provenance
Explanation	Every event in output data's provenance is captured by the system from raw satellite image download, every processing step, ingestion into ML/AI models and output generation
Input	Satellite images, processing parameters, indices, models, user decisions
Output	Provenance record, verification certificate
Verification method	Inspection, Test

Requirement ID	**GT-DT-FUN-02**
Subject	Event detail registration
Requirement	The system shall register all relevant information about the data entries and events uniquely (e.g. raw data provenance, event details)
Explanation	The details of data inputs and events, such as ID of data or event, data source source, process description, timestamp, actor organisation. location of the event (if relevant) are registered uniquely, so they can be traced and verified
Input	Data entities, events
Output	Provenance record, verification certificate
Verification method	Inspection, Test

Requirement ID	GT-DT-FUN-03
Subject	Data immutability
Requirement	All data provenance events shall be captured, registered and stored immutably
Explanation	Immutability proves that all data modifications are captured in a way that they can be traced later
Input	Data entry or event
Output	Data entry or event indicated in provenance record
Verification method	Test

Requirement ID	GT-DT-FUN-04
Subject	Data signing
Requirement	Each entry and event in data provenance shall be digitally signed in a cryptographically verifiable way
Explanation	The entities making modifications to data in the provenance trail must be identifiable by their signatures and time of signatures
Input	Data query
Output	Provenance record with identifiable stakeholders

Verification method	Inspection, Test

Requirement ID	**GT-DT-FUN-05**
Subject	Data traceability
Requirement	The system shall enable the traceability of each data entry and event
Explanation	Each data entry and event is recorded in the provenance trail
Input	Data query
Output	Provenance record, verification certificate
Verification method	Inspection, Test

Requirement ID	**GT-DT-FUN-06**
Subject	Data verification
Requirement	The system shall enable the verification of each data entry and event
Explanation	The verification proves the immutable and traceable nature of data entries and events
Input	Data query
Output	Provenance record, verification certificate

Verification method	Inspection, Test

Requirement ID	**GT-DT-FUN-07**
Subject	Data output
Requirement	The system shall enable the export of data provenance record in both machine readable and standard file formats (e.g. JSON, XML, txt, csv, pdf) and a verification certificate (signed hash of the verified provenance record)
Explanation	The provenance record is downloadable in a file format preferred by the user, indicating the size of file before download. Both the provenance record and verification certificate will be signed by the system when downloaded.
Input	Data query, specified download preference
Output	Signed provenance record in preferred file format, file size indication, verification certificate
Verification method	Test

Requirement ID	**GT-DT-FUN-08**
Subject	Automatic verification
Requirement	Adding new entry or event to data provenance trail triggers automatic verification of the data's previous trail

Explanation	No entries or events can be added to the data that's previous provenance trail is not verified, confirming that only verified data is registered in provenance trail
Input	Data entry, event addition
Output	Provenance trail verification
Verification method	Test

Requirement ID	**GT-DT-FUN-09**
Subject	User management
Requirement	The system allows differentiating users with entry and event adding rights and users with verification rights
Explanation	The output of data, queried by user, has clear information on which permissioned entities have made modifications to the data
Input	User role definitions, credentials
Output	User roles and accounts
Verification method	Test

Requirement ID	**GT-DT-FUN-10**
Subject	Authentication

Requirement	The system allows only authenticated entities with corresponding permissions to add traceable events to data provenance
Explanation	The output of data, queried by user, has clear information on which permissioned entities have made modifications to the data
Input	User credentials with according permissions
Output	Access to system's data modification functionalities
Verification method	Test

Requirement ID	**GT-DT-FUN-11**
Subject	System configuration
Requirement	The system allows updates by authenticated users to initial data schemas, modification and processing schemas that are being conducted on satellite data, ensuring backward compatibility
Explanation	The schemas describing modification and processing steps or initial data must be possible to update, so new attributes or structures can be added or existing attributes changed. Backward compatibility must be ensured to enable end-to-end data tracking without interruptions
Input	User credentials with corresponding permissions
Output	Access to system's data modification functionalities
Verification method	Test

4. Preliminary choice of technologies

To fulfil the system requirements stated above, some disruptive technologies are more suitable than others, taking into account the complexity of setting up such a system, the level of usage for these technologies, and compatibility with existing systems and solutions.

As the solution should be suitable for a number of different stakeholders, possibly situated in various countries with diverse legislation, and technology maturity, one of the first criteria for choosing the suitable technology for the described use cases [RD-02] is that the technology should be established and already in use in a large-scale project. Though some disruptive technologies described above might see a wider range of applications in the future, they could introduce complexities to the development of the system that are yet to be solved to a satisfactory level. This criterion makes the possible usage of confidential computing and homomorphic encryption questionable because of the lack of standardization and incompleteness.

Another principle for finding the suitable technology or set of technologies is the requirement of privacy: does any entity in the traceable lifecycle of the data need to be hidden. According to the use cases, the traceability service could be used by anybody to trace data, but generating and modifications with the data are conducted by known actors that will be traced. For cases when an organisation uses traceable data to develop a model or conduct other processes with the initial data that would be usable only inside the organisation and not for others, the system should also allow different privacy set ups. So the usage of zero knowledge proofs, mainly being concerned with the privacy preservation for all processing, also coming with its own complexities, would be hard to justify in this case.

Federated learning and AI powered anomaly detection technologies would already be used in the system, but mostly as traceable processing steps in the lifecycle of the initial data. The main objective of the system is to create a trail of processes while federated learning and AI powered anomaly detection - although powerful tools for other goals - are not suitable for this.

For complying with the traceability system's main purpose that is in accordance with stated system requirements, a combination of blockchain, quantum resistant timestamping, decentralised knowledge graphs, and two-dimensional hash-trees seems most feasible. While the definition of Self Verifiable Data Objects is vague,

allowing different solutions with Verifiable Credentials (VC) and Decentralized Identifiers (DID), this may be used for the systems user management and privacy settings, if the goals for proper user rules could not sufficiently be fulfilled with other technologies. Quantum resistant timestamping is able to create verifiable steps in a data trail, also attributing any step securely to its conductor - an organisation, an AI model used by an organization, etc. Blockchain enables solid traceability and verification for the steps and allows diverse management of stakeholders (i.e. contributors and verifiers) with different rights to contribute to the chain. Some aspects of the decentralised knowledge graphs, that combine distributed ledger (blockchain), distributed storage, and decentralized identity technologies would likely be used in the traceability system too, but the benefits of decentralised knowledge graphs, such as traceability, policy enforcement and verification are likely also to be covered with other aforementioned technologies.

A technology requiring further analysis is the two-dimensional hash trees solution proposed in chapter 2.8. While it allows the certification of sub-images from a large-scale satellite image, creating a secure step in the provenance trail of the initial image, this technology has not been used in that way before as known.

5. Conclusion

In the study report, a list of disruptive technologies were analysed for the possible strengths and weaknesses they could bring to the traceability system. Relevant real life use cases for these technologies were surveyed, with the emphasis on how well established and widely used these solutions are. On the basis of the possible use cases for the traceability system, system requirements describing the main functionalities of the system to be developed were formed, stating the needs for user access, authorisation, provenance tracking and verification of data's life cycle.

When combining the system requirements, use cases and analysed technologies, the implementation of some technologies does not seem feasible, because of the lack of standardisation, not having large scale real life use cases or being focused on purposes that are not the main objectives of the traceability system, so Zero Knowledge Proofs, Confidential computing and homomorphic encryption were discarded from next development phases.

The technologies that are chosen - based on system requirements - are the blockchain technology, quantum resistant timestamping that fulfil the main objective of the system - tracing all processes being conducted to an initial data(set) and allowing the verification and authentication of every step in its provenance. Other technologies, such as federated learning and AI powered anomaly detection, Self Verifiable Data Objects, decentralized knowledge graphs and two-dimensional hash trees have promising features that can be useful to perform some required functions in the system.

In conclusion, using the blockchain in combination with quantum resistant timestamping has been found to be most suitable for the traceability system to be developed, while the various functionalities of the remaining disruptive technologies are acknowledged and will be implemented, if necessary.

6. References

1. AMD. (2023). MonetaGo uses confidential computing to detect duplicate financing fraud. https://www.amd.com/content/dam/amd/en/documents/resources/case-studies/monetago-case-study.pdf
2. Barclay, I., Radha, S., Preece, A., Taylor, I., & Nabrzyski, J. (2020, April 6). Certifying provenance of scientific datasets with self-sovereign identity and verifiable credentials (arXiv:2004.02796 [cs.CR]). arXiv. https://doi.org/10.48550/arXiv.2004.02796
3. Cebeci, U., & Arat, E. (2022). Establishing Agri and Food Supply Chain Provenance Based on Blockchain: Literature Review. Avrupa Bilim ve Teknoloji Dergisi, (37), 59–64.https://dergipark.org.tr/en/pub/ejosat/issue/70985/1131779
4. Coorest. (n.d.). Onboarding partners. Coorest. https://coorest.io/onboarding/
5. Dayan, I., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., Liu, A., Costa, A. B., Wood, B. J., Tsai, C.-S., Wang, C.-H., Hsu, C.-N., Lee, C. K., Ruan, P., Xu, D., Wu, D., Huang, E., Kitamura, F. C., Lacey, (...), Mazzulli, T. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. Nature Medicine, 27(10), 1735–1743. https://doi.org/10.1038/s41591-021-01506-3
6. Dhamodharan, R. (2024, May 14). The next generation of blockchain – and banking: Fusing the flexibility of crypto with the convenience of fiat. Mastercard News & Trends. https://www.mastercard.com/us/en/news-and-trends/press/2024/may/the-flexibility-of-crypto-the-convenience-of-fiat-bringing-blockchain-to-banking.html
7. DigiCert, Inc. (n.d.). Post-quantum cryptography (PQC) support. DigiCert Documentation. Retrieved in July 2025, from https://docs.digicert.com/en/digicert-private-ca-services/ca-manager/post-quantum-cryptography--pqc--support.html
8. European Commission. (2022). TRUSTEE: Trust and privacy preserving computing platform for cross-border federation of data (Grant Agreement No. 101070214). CORDIS. https://cordis.europa.eu/project/id/101070214
9. European Commission. (2024, April 11). Commission recommendation (EU) 2024/1101 on a Coordinated Implementation Roadmap for the transition to post-quantum cryptography (C/2024/2393). Shaping Europe's digital future. Retrieved in July 2025, from https://digital-

strategy.ec.europa.eu/en/library/recommendation-coordinated-implementation-roadmap-transition-post-quantum-cryptography
10. Fernández-Caramés, T. M., & Fraga-Lamas, P. (2020). Towards post-quantum blockchain: A review on blockchain cryptography resistant to quantum computing attacks. IEEE Access, 8, 21091–21116. https://doi.org/10.1109/ACCESS.2020.2968985
11. Flores, M. (2020, October 5). Triaging COVID-19 patients: 20 hospitals in 20 days build AI model that predicts oxygen needs. NVIDIA. https://blogs.nvidia.com/blog/federated-learning-covid-oxygen-needs/
12. Gatsby, T. (2025, March 7). Quantum-resistant blockchain protocols: Preparing for the future. Beyond the Hype. Retrieved in July 2025, from https://beyondthehype.terrencegatsby.com/blockchain/quantum-resistant-blockchain-protocols-preparing-for-the-future/
13. Global Blockchain Business Council. (2025). 101 Real-World Blockchain Use Cases Handbook [PDF].https://downloads.ctfassets.net/so75yocayyva/1QeBFAtvDyoHK6pJgG5ITU/63401e1dfda1d22e4a5172cc33cb424d/GBBC-s_101_Real-World_Blockchain_Use_Cases_Handbook_digital.pdf
14. Henrichs, E., Boller, M. L., Stolz, J., & Krupitzer, C. (2025). Quantum of trust: Overview of blockchain technology for product authentication in food and pharmaceutical supply chains. Trends in Food Science and Technology, 157, Article 104892. https://www.sciencedirect.com/science/article/pii/S0924224425000287
15. ING Bank. (2017, November 16). Blockchain transactions just got a whole lot safer. ING. https://www.ing.com/Newsroom/News/Blockchain-transactions-just-got-a-whole-lot-safer.htm
16. ING Bank. (2018, October 21). Blockchain innovation improves data privacy for clients. ING. https://www.ing.com/Newsroom/News/Blockchain-innovation-improves-data-privacy-for-clients.htm
17. ING Bank. (n.d.). zkflow. GitHub. https://github.com/ing-bank/zkflow
18. International Organization for Standardization. (2025). ISO/IEC DIS 28033-1: Information security — Fully homomorphic encryption — Part 1: General (ISO/IEC DIS 28033-1). https://www.iso.org/standard/87638.html
19. Jokumsen, M., Pedersen, T. P., Daugaard, M. S., Tschudi, D., Madsen, M. W., & Wisbech, T. (2023). Verifiable proofs for the energy supply chain:

Small proofs bring you a long way. Energy Informatics, 6(Suppl 1), Article 28. https://doi.org/10.1186/s42162-023-00283-2

20. Kolade, T. (2025, April 2). Mastercard advances blockchain strategy with multi-token network to bridge crypto and traditional finance. ETHNews. https://www.ethnews.com/mastercard-advances-blockchain-strategy-with-multi-token-network-to-bridge-crypto-and-traditional-finance/
21. Ledger Insights. (2019, May 2). ING's privacy blockchain breakthrough makes it into financial results announcement. https://www.ledgerinsights.com/ing-blockchain-privacy-bulletproofs-zkp/
22. LF Decentralized Trust. (2019, March 11). Reducing government red tape: British Columbia creates new business identity model with Hyperledger Indy. https://www.lfdecentralizedtrust.org/blog/2019/03/11/reducing-government-red-tape-british-columbia-creates-new-business-identity-model-with-hyperledger-indy
23. Malik, S., Dedeoglu, V., Kanhere, S. S., & Jurdak, R. (2021). PrivChain: Provenance and privacy preservation in blockchain-enabled supply chains (arXiv:2104.13964 [cs.CR]). https://doi.org/10.48550/arXiv.2104.13964
24. Mastercard. (2021, April 15). Partnership with ConsenSys supports the future of multi-blockchain commerce [Press release]. Mastercard Newsroom. https://www.mastercard.com/news/press/2021/april/partnership-with-consensys-supports-the-future-of-multi-blockchain-commerce/
25. Mastercard. (2021, June 17). E-Livestock Global launches first Mastercard blockchain-based solution to bring visibility to the cattle industry in Zimbabwe. Mastercard Newsroom. https://www.mastercard.com/news/eemea/en/newsroom/press-releases/en/2021/june/e-livestock-global-launch-mastercard-blockchain-based-solution/
26. Mastercard. (2023, July). Unlocking the potential of digital asset innovation: Building a Mastercard Multi-Token Network™ [White paper]. Mastercard. https://www.mastercard.com/news/media/5zmixdjy/unlocking-the-potential-of-digital-asset-innovation-building-a-mastercard-multi-token-network-1-1.pdf
27. Mastercard. (n.d.). Global Trade & Freight Solutions. Mastercard. https://www.mastercard.com/global/en/business/large-enterprise/mastercard-enterprise-partnerships/global-trade-freight-solutions.html

28. Nardi, M., Valerio, L., & Passarella, A. (2022). Anomaly detection through unsupervised federated learning. In 2022 18th International Conference on Mobility, Sensing and Networking (MSN) (pp. 495–501). https://www.computer.org/csdl/proceedings-article/msn/2022/645700a495/1LUtH1eeYak
29. National Institute of Standards and Technology. (2022, July 5). Announcing PQC candidates to be standardized, plus fourth round candidates. National Institute of Standards and Technology. https://csrc.nist.gov/news/2022/pqc-candidates-to-be-standardized-and-round-4
30. Sahai, S., Singh, N., & Dayama, P. (2019, March 11). Enabling privacy and traceability in supply chains using blockchain and zero-knowledge proofs [SlideShare presentation]. https://www.slideshare.net/slideshow/enabling-privacy-andtraceabilityinsupplychainsusingblockchainandzeroknowledgeproofs/250842942
31. Sardar, M. U., & Fetzer, C. (2023). Confidential computing and related technologies: A critical review. Cybersecurity, 6, Article 10. https://doi.org/10.1186/s42400-023-00144-1
32. Satybaldy, A., Subedi, A., & Nowostawski, M. (2022). A framework for online document verification using self-sovereign identity technology. Sensors, 22(21), Article 8408. https://doi.org/10.3390/s22218408
33. Sinai, N. K., & In, H. P. (2024). Performance evaluation of a quantum-resistant Blockchain: A comparative study with Secp256k1 and Schnorr. Quantum Information Processing, 23(3), 99. https://doi.org/10.1007/s11128-024-04272-6
34. Sola-Thomas, E., & Imtiaz, M. H. (2025). Development of a quantum-resistant file transfer system with blockchain audit trail. Clarkson University. https://www.researchgate.net/publication/390671281_Development_of_a_Quantum-Resistant_File_Transfer_System_with_Blockchain_Audit_Trail
35. Soltani, R., Nguyen, U. T., and An, A., "Data Capsule: A Self-Contained Data Model as an Access Policy Enforcement Strategy," 2021 3rd Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS), Paris, France, 2021, pp. 93-96, https://ieeexplore.ieee.org/document/9569788
36. Sumner, M. (2024, August 16). Time-Stamping digital content: Technical guide. ScoreDetect. Updated May 6, 2025. Retrieved in July 2025, from

https://www.scoredetect.com/blog/posts/time-stamping-digital-content-technical-guide

37. Sun, L.-S., Bai, X., Zhang, C., Li, Y., Zhang, Y.-B., & Guo, W.-Q. (2022). BSTProv: Blockchain-Based Secure and Trustworthy Data Provenance Sharing. Electronics, 11(9), 1489.https://www.mdpi.com/2079-9292/11/9/1489
38. Sáez-de-Cámara, X., Flores, J. L., Arellano, C., Urbieta, A., & Zurutuza, U. (2023). Clustered federated learning architecture for network anomaly detection in large-scale IoT and IIoT systems. Computers & Security, Volume 131, August 2023, 103299. https://www.sciencedirect.com/science/article/pii/S0167404823002092?via%3Dihub
39. Thales Alenia Space. (2025, June 25). Thales Alenia Space and 3IPK deploy first blockchain network in space on IMAGIN-e payload aboard the International Space Station. Thales Alenia Space. https://www.thalesaleniaspace.com/en/news/thales-alenia-space-and-3ipk-deploy-first-blockchain-network-space-imagin-e-payload-aboard
40. Thanalakshmi, P., Rishikhesh, A., Marion Marceline, J., Joshi, G. P., & Cho, W. (2023). A quantum-resistant blockchain system: A comparative analysis. Mathematics, 11(18), Article 3947. https://doi.org/10.3390/math11183947
41. Third, A., & Domingue, J. B. (2023).Decency and Decentralisation: Verifiable Decentralised Knowledge Graph Querying. Companion Proceedings of the ACM Web Conference 2023, 1432–1434. https://dl.acm.org/doi/fullHtml/10.1145/3543873.3587635
42. Thornton, A. (2020, March 27). What is ElectionGuard? Microsoft. https://news.microsoft.com/on-the-issues/2020/03/27/what-is-electionguard/
43. Web of Trust. (n.d.). OrgBook BC 79. https://www.weboftrust.org/project/orgbook_bc-79
44. Yang, Z., Alfauri, H., Farkiani, B., Jain, R., Di Pietro, R., & Erbad, A. (2024). A survey and comparison of post-quantum and quantum blockchains. IEEE Communications Surveys and Tutorials, 26(2), 967–1002. https://ieeexplore.ieee.org/document/10288193
45. Zaarour, T., Khalid, A., Pradeep, P., & Zahran, A. Using distributed ledgers to build knowledge graphs for decentralized applications. Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM 2024). https://doi.org/10.1145/3627673.3679644