A Secure and Trustworthy Biometric Data Ecosystem for Cross-border Suspect Identification

Katerina Kyriakou[†,‡], Apostolos Apostolaras[†,‡], Polychronis Velentzas[†], Georgios Benos[⊖], Konstantinos Koutsoukos[⊖], Chrysostomos Symvoulidis[⊖], Kaitai Liang[⊕], Zeshun Shi[⊕], Asterios Leonidis[◇], Kyriaki Miniadou[◇], Eleni Veroni[*,§], Spyridon Evangelatos[*,§], Georgios Th. Papadopoulos[◇,≀] and Thanasis Korakis[†,‡]

[†]*Department of Electrical and Computer Engineering, University of Thessaly*, Volos, Greece
[‡]*CERTH, The Centre for Research & Technology, Hellas*, Volos, Greece
[⊖]*Research & Development Thridium, Ltd.*, London, United Kingdom
[⊕]*Delft University of Technology*, Delft, The Netherlands
[◇] *Institute of Computer Science (ICS) Foundation for Research and Technology - Hellas (FORTH)*, Crete, Greece
[*]*Research & Innovation Development, Netcompany-Intrasoft S.A.*, Luxembourg, Luxembourg
[§]*Dept. of Electronic Engineering, Hellenic Mediterranean University*, Crete, Greece
[≀]*Department of Informatics and Telematics, Harokopio University of Athens*, Athens, Greece

2024 IEEE International Conference on Big Data (BigData) | 979-8-3503-6248-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/BigData62323.2024.10826113

***Abstract*—This paper introduces the Biometrics Data Space framework, which is a secure ecosystem built on Data Spaces technology and it is designed to address the challenges of suspect identification during cross-border crime investigation. Apart from Data Spaces technology, the proposed framework innovates by leveraging also Privacy Enhancing Technologies (PETs) and blockchain to enable secure, trustworthy, and sovereign data exchange between Law Enforcement Agencies (LEAs) across borders. Specifically, it utilizes advanced PETs, including Large-Scale Biometric Data Indexing based on deep hashing techniques and Homomorphic Encryption to allow for suspect identification without disclosing sensitive information of personal biometric data. Thus, it enables LEAs to securely compare and exchange encrypted sensitive biometric data, including facial images, fingerprints and voiceprints, while maintaining data privacy and data sovereignty. LEAs define the usage rules for the biometic data they own and these rules are enforced to and respected by the other LEAs participating in the Biometrics Data Space. The proposed architecture is designed to be scalable, allowing the incorporation of additional biometric modalitiies and the easy expansion and integration with new participant LEAs.**

***Index Terms*—Biometrics Data Space, Cross-Border Suspect Identification, Privacy Enhancing Technologies, Homomorphic Encryption, Biometrics Data Sovereignty**

I. Introduction

Over the past few decades, there has been considerable interest in the adoption of biometric technologies for suspect identification systems, driven by the uniqueness of a person's biological characteristics, such as the facial features, voiceprints or fingerprints. These technologies have a wide application area such as border control, police investigations, and the cross-organization sharing of information to combat terrorism, global crime, and illegal migration. In tackling crime and terrorism across several countries, the necessity for cross-border police cooperation has become significantly critical. Cross-border suspect identification systems in forensic investigations require several key criteria to be met, among which are the security of the exchanged results and the participants' sensitive information, the accuracy and speeding up of obtaining the identification results, and the improvement of automated exchange for biometric data.

As digital biometric technologies become more prevalent for enhancing identification and identity verification processes, concerns about privacy, ethics, and the requirement to comply with EU data protection regulations are growing. Such risk-related issues may involve misuse of personal data or ethical issues arising from Artificial Intelligence (AI) driven decisions. To address the limitations on cross-border data exchange for police cooperation, this paper proposes the Biometrics Data Space, leveraging the power of strong privacy safeguards introduced by the International Data Spaces (IDS) [1]. The suggested framework incorporates Privacy Enhancing Technologies (PETs), such as Encryption and Large-Scale Data Indexing, along with the Blockchain Smart Contracts for the orchestration of a trustworthy, secure, and self-sovereign data-sharing ecosystem.

II. Literature Review & State-of-the Art Analysis

A. Prüm Framework

The Prüm framework has established a roadmap to enhance police authorities cooperation against fighting global crime, with a focus on the facilitation of cross-border information sharing. Initially introduced in 2005 as the Prüm Treaty [2], which provides a signed legal framework for the data exchange of DNA, fingerprints, and vehicles registration data for law enforcement purposes. In 2008, the European Commission proposed the incorporation of Prüm Treaty into the EU legislation and in 2012, the implementation of the Prüm Decision [3] was initiated with several EU member states working towards the connection of their national databases for automated data exchange. The European Commission introduces Prüm II [4] regulation in 2021, with measures for improved cooperation in data exchange to combat serious crime and terrorism. The

Prüm II regulation is enhanced in 2023 by adding facial images and police records, and by centralising data flows, to make them faster and more effective.

B. Data Spaces

The role of data has become of significant importance in characterizing several areas of interest ranging from businesses to governments. Among the several valuable roles of data are the facilitation of operational optimization, the data-driven decision making towards innovation and evaluation, the establishment of transparency and traceability, as well as the enhancement of tackling global crime. To harness the value of data, sharing in a secure manner inside an ecosystem of trusted participating actors is considered crucial. Nonetheless, in cases of data sharing, it is essential to align and respect the European Union's regulations such as General Data Protection Regulation (GDPR) [5] and principles like Findable, Accessible, Interoperable, and Reusable (FAIR) [6]. For a data sharing scheme that aims at the exploitation of the value of data, it is crucial to guarantee the protection of personal data of users and to prohibit the unregulated use of digital data in ways that compromise the ethical rights of citizens.

Emphasizing on achieving a common purpose of sharing for growth, innovation and security, the notion of data sovereignty is considered a fundamental characteristic of this process. These considerations have been focused by the European Union in terms of providing a strategy for data [7] that exploits the power of preserving owners' control over their own resources. The European strategy for digital data aims at the formation of strong safeguards over data protection while prioritizing privacy and trust. In response to these specifications, Data Spaces have been gaining attention by the European Union as an infrastructure that provides the ability for storage, accessing, exchanging and processing digital data in a privacy-preserving, self-sovereign, and interoperable manner.

In 2015, Fraunhofer research project launched the initiative of the IDS [1] aiming to establish the distributed software architecture of Data Spaces for security, interoperability and sovereignty. Moreover, the IDS Association (IDSA) extended the research on IDS by introducing the IDS Reference Architecture Model (IDS RAM) [8], which serves as the main architecture for the software design of a Data Space. IDS RAM provides several layers of the architecture including the system layer, which describes the primary and secondary Data Spaces components, and the process layer, which defines the necessary processes taking place for the enabling of sharing data and enforcing access and usage control policies. These layers ensure the privacy preservation and the ethical use of data for the purposes of sharing, while also adhering to the common European regulations.

In addition to those contributions, the GAIA-X initiative [9] aims at the extension of the IDSA goals on data sharing and sovereignty. The goal is the formation of an architecture that provides services for data exchange as well as for data storage and management on cloud infrastructures. The goals of GAIA-X architecture are focused on the formation of a distributed services catalog, the secure and sovereign data exchange, the trusted identity management, and the compliance services.

The key component to the Data Space functionalities is the Connector, which acts as a secure communication tunnel for exchanging information securely while preserving the owner's control policies over the data. There are several approaches in the implementation of the Connector component that are certified by IDSA and comply with IDS and GAIA-X protocols and requirements. The updates and additions to Connectors implementation is disseminated through the periodic IDSA Connector Report [10]. The Connector focused on in this paper is the TRUE (TRUsted Engineering) Connector [11] by Engineering which is a compatible with the FIWARE [12] technical blocks catalog.

C. Blockchain

Over the years, Blockchain technology due to its decentralized and immutable nature and security, has found applications in various fields such as finance, supply chain management, and healthcare. To begin with, there exist three major categories of blockchain types; (*i*) public blockchains, (*ii*) private blockchains, and (*iii*) consortium blockchains. Public blockchains such as Ethereum [13], [14], are decentralized platforms, open to the public in which, multiple nodes contribute to the consensus process and maintain the integrity of the blockchain. Public blockchains are appropriate for applications that need trust and censorship resistance since they are usually transparent, unchangeable, and secure. On the other hand, private blockchains, such as Hyperledger Fabric [15], [16], and Corda [17], [18], are exclusive to a certain member group and have permission-based access controls. In corporate context, when a consortium or a collection of organizations operate together on a blockchain network, private blockchains are frequently utilized due to their increased control over privacy, scalability, and governance. The third category, *i.e.*, Consortium blockchains are in essence a hybrid solution between public and private ones, where they involve organizations or entities collaborate in order maintain the network. Consortium blockchains are typically preferred when a more decentralized approach is required.

Blockchain technology, can also be used in order to provide transparent, secure and decentralized ways for sharing and accessing data, as in the case of the Biometrics Data Space. Such solutions include the InterPlanetary File System (IPFS) [19], a distributed file system that utilized blockchain technology in order to generate a decentralized and peer-to-peer network for storing and sharing files. Ocean Protocol is yet another platform that allows individuals and organizations to securely share their data by providing a marketplace where data providers can publish and sell their data while maintaining control over their access and usage rights. Similar solutions also include the ones provided by IOTA [20], [21] which allows IoT and machine-to-machine communication, and Sovrin [22] which allows sharing of the citizen's digital identity information with trusted parties.

D. Privacy Enhancing Technologies

1) Large-Scale Indexing: Indexing [23] broadly refers to the use of some indicator or measure as a reference. It

adds structure to an existing body of data points significantly enhancing the storage, retrieval and comparison processes. Hashing [24] is a particularly effective technique for indexing since it converts high-dimensional data into more efficient, compact representations called hashcodes. This mitigates the complexity of operations performed on high-dimensional data reducing both search complexity and storage demands.

With the growing popularity of Deep Learning as a cutting-edge field and its increasing use in academic research, Deep Hashing [25] has gained prominence. This prominence is accentuated by its ability to optimise the differentiation and grouping of similar data which in turn reduces computational cost and enhances retrieval performance. Deep hashing, utilises advanced hashing functions that move beyond exact matching to introduce the concept of approximate matching, which is especially important for managing noisy biometric data with subtle differences. This approach effectively handles biometric data that are inherently high-dimensional while accommodating minor variations, such as slight changes in facial expressions or fingerprint positioning, ensuring biometric systems remain accurate and efficient.

The increasing demand for efficient and accurate biometric data search across large-scale databases has driven the development of sophisticated multimodal frameworks that combine more that one type-modality of biometric data. Examples include a multi-biometric algorithm [26] that integrates palmprints and dorsal hand veins using a combination of deep learning and graph matching techniques at score level. Efficient deep palmprint recognition has also been achieved [27] through a combination of hash coding and knowledge distillation techniques, which aim to enhance the overall efficiency and effectiveness of palmprint recognition systems. Furthermore, the Regularised Adversarial Domain Adaptive Hashing (R-ADAH) method [28] has been introduced for cross-domain palmprint recognition, enabling the adaptation of palmprint recognition systems across various domains, thereby improving performance in diverse operational settings.

Deep hashing techniques have been widely applied in facial recognition systems, where they enhance the speed and accuracy of searches within large-scale databases. For instance, a fast face search method proposed by [29] integrates deep convolutional neural networks (CNNs) and semantic hashing to improve the efficiency of facial data retrieval. The system uses a ResNet model to extract deep face features, followed by Principal Component Analysis (PCA) and binarization to convert real-valued features into compact hash codes, significantly accelerating the face search process. Furthermore, the work of [30] focuses on person re-identification (re-id) by combining appearance, posture, and emotion cues in a portrait interpretation framework, which leverages deep learning and hashing for accurate and real-time face-based searches.

2) Homomorphic Encryption: Homomorphic encryption (HE) is a powerful cryptographic technique that enables computations to be performed directly on encrypted data without requiring decryption. This feature offers substantial advantages for ensuring data privacy and security, particularly in scenarios where sensitive information must remain confidential while undergoing processing. Over the past decades, HE has matured significantly, with notable advancements in both theory and application [31]–[33]. These developments have unlocked numerous applications across sectors such as cloud computing, healthcare, and finance [34], [35]. For instance, companies leveraging cloud services can outsource data processing tasks without exposing their raw, sensitive data to cloud providers. In healthcare, HE allows secure querying of encrypted databases, facilitating medical research while preserving the privacy of patient data [36], [37].

In the field of biometrics, homomorphic encryption offers a robust solution for secure biometric identification and authentication, providing privacy-preserving methods for individuals to verify their identity [38], [39]. This is especially critical in sectors like healthcare, where the confidentiality of biometric data is paramount. Biometric features, such as encrypted fingerprints, can be matched with encrypted databases without disclosing any raw biometric information [40]. Within the context of the proposed framework, we aim to combine HE with deep hashing techniques to efficiently compare biometric data against large-scale databases. This approach focuses on selectively encrypting key features of the biometric data, such as facial or fingerprint features, to optimize computational efficiency. By minimizing the data involved in the encryption process, the system ensures privacy while reducing the overhead typically associated with HE [34], [37], [39]. This approach allows for secure and scalable biometric comparison while maintaining cost-effective encryption standards.

3) Fuzzy Extractor and AES Encryption: Fuzzy extractors are cryptographic primitives that are crucial for generating stable cryptographic keys from noisy inputs, such as biometric data. These extractors work by employing a reconciliation mechanism that minimizes the influence of noise while deriving a consistent secret key. This mechanism is particularly useful for biometric data, which naturally introduces variability due to environmental factors and sensor inconsistencies [41], [42]. Biometric-based fuzzy extractors offer several key benefits compared to traditional cryptographic hashing functions like SHA-256 or MD5, including improved security, flexibility, and privacy preservation [43], [44]. For example, a fuzzy extractor can securely derive a key from a fingerprint scan by removing noise and variations in the scanned data, generating a key that is resilient to errors introduced during data capture.

The AES encryption algorithm is employed to safeguard the underlying biometric data [45], [46]. AES is known for its lightweight and computationally efficient nature, making it well-suited for real-time biometric matching processes, such as secure fingerprint or iris-based authentication. In the proposed system, fuzzy extractors and AES encryption are combined to ensure both privacy and security in biometric authentication. In our design, fingerprints or facial templates are extracted from noisy inputs, transformed into a stable cryptographic key using a fuzzy extractor, and then utilized as the AES key to encrypt the raw biometric data. By using the fuzzy extractor's output as the AES key, we create a direct link between the biometric data and the encryption process. This approach enhances the overall security and privacy of the system, as the encryption key is inherently tied to the specific biometric input, making

unauthorized access significantly more difficult [41], [45].

III. User requirements

Forensic agencies, police authorities and cross-border control investigators encounter a variety of challenges when addressing the process of suspect identity verification through cross-border data exchange. Among the potential complications of biometric data sharing through LEAs are the differentiations in each country's national legislation for protecting privacy of personal data, the concerns regarding ensuring security and trust, as well as the existing technological tools and their extension capabilities of each member and non-member EU countries. The specific obstacles may hinder the seamless sharing of biometric data for the purpose of cross-border identification. Another aspect that is considered as a predicament for the international cooperation among LEAs involves the concerns related to potential misuse of the sensitive data and compromisation of the personal privacy.

The confrontation of the existing complications towards the formation of a cooperative investigation system is intented to rely on the adaptation of the data sharing system to respect specific design principles. The design requirements must ensure trust and security among participating actors through certification and validation, and data sovereignty that enforces access and usage control regulations strictly determined by the data holder. Additionally, interoperability that ensures data exchange capability between remote actors with differentiations on their legislation on existing systems, and the formation of an ecosystem of data where the necessity for a centralized storage unit is omitted and data remains at their original resources instead.

Along with the privacy preservation and ethical use of data concerns, such a system is obligated to respect the requirements of the new Prum II regulation [4] on the police cooperation on automated data exchange. As a consequence, the alignment of the data sharing system should be enforced by allowing the automation of the exchange of criminal records with biometric data, the establishment of central components to enable the automation process, the acquisition of results within 48 hours, and the enhancement of data protection with corresponding technologies.

IV. Architecture

A. Data Model

The objective of suspect identification requires strong evidence in order to guarantee accuracy in the obtained results. For that cause, the Biometrics Data Space for identity verification leverages the power of biometric characteristics. The proposed data model for a potential suspect's profile is based on the ANSI/NIST-ITL 1-2011 [47] structure for criminal record exchange. The suggested suspect record format harnesses the distinctiveness of biometric features, specifically, facial images, fingerprints and voiceprints, along with additional record transaction information. The specific model is targeted around strong biometric characteristics that aim towards high accuracy of identification. The proposed model can be extended to support other biometric features for recognition.

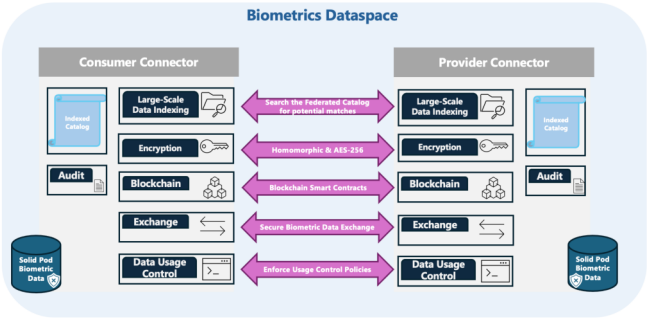


Fig. 1: Biometrics Data Space Reference Architecture diagram

Apart from the biometric characteristics, the potential suspect's profile utilized in the Biometrics Data Space will store additional information regarding the record registration details, such as date of record creation, LEA officer who executed the process, etc., and potential suspect-related information from notes taken by the Police Officer, such as height, nationality, gender, name, or any other descriptive text. Based on the selected data model the process of trustworthy cross-border suspect recognition will be enhanced towards accurate, fast and efficient results.

B. Reference Architecture for Biometrics Data Space & Design Principles

The design of the Biometrics Data Space relies on a high-level reference architecture model that involves the key technology components for the establishment of a trustworthy, independent and interoperable data exchange framework. The core functionalities and technologies that are incorporated are presented in Figure 1. The integrated Data Space components involve one Connector component per each LEA which is responsible for the secure biometric data exchange, the communication with the key technologies, the retrieval of the biometric data from the local storage of each police authority, and the enforcement of usage control policies. In order to ensure privacy preservation and data integrity, the Biometrics Data Space leverages thw power of Encryption and Large-Scale Indexing. Moreover, transparency and traceability is handled by the Blockchain Smart Contracts.

The main design principles that characterize the Biometrics Data Space are the following:

Trustworthiness: Trust is significantly important in the Biometrics Data Space. Every participant is obliged to possess a certified Connector with a unique identifier in order to exchange data. Thus, only trusted organizations are allowed to request or provide data and trustworthy usage of resources is ensured since the usage policies are respected.

Security: The biometric data sharing should be performed under increased security through the establishment of a communication channel among only trusted Connectors. Security is enhanced by the integration of Privacy Enhancing Technologies which enable the anonymization and pseudonymization of data, thus, augmenting the sharing process with an additional layer of protection.

Data Sovereignty: The capability of access and usage policies self-determination over the data is a fundamental aspect

of the Biometrics Data Space. Data owners define the control policies on their own data including who has access to the data, the duration for their availability, location-based filtering, and more.

Interoperability: The data exchange process in the Biometrics Data Space is considered as type agnostic, since different LEAs from different countries own different formats and protocols of biometric data.

Transparency & Traceability: When it comes to sharing biometric data between cross-border LEAs, it is important to maintain recording of data exchanges through Blockchain Smart Contracts, safeguarding both parties (data requestor and data owner) against unauthorised data access requests and premature or unjustified access revocations, respectively.

Ecosystem of Data: In the Biometrics Data Space, there is no demand for a central storage and management unit to successfully conduct data sharing. On the other hand, the databases of biometric data remain decentralized on the local premises of each LEA.

C. Case Management System

The Case Management System (CMS) allows central handling of cases and data management. Its scope is to bridge the gap between the various components participating in the use case scenario by streamlining and monitoring the entire process as well as orchestrating the performed operations. The architecture of the Case Management System will be flexible and scalable to accommodate the various case management needs of LEAs in various contexts. The primary functionalities of the CMS component are (*i*) workflow automation, (*ii*) case data management, (*iii*) component orchestration, and (*iv*) case progress tracking. The above functionalities are all supported by robust data security protocols to protect sensitive information and adhere to rigorous user privacy policies.

This will be achieved by exploiting a Business Process Management Tool (BPMN), such as Camunda [48] to model the flow of the use case and illustrate the involvement of the various components. An external NoSQL database (MongoDB [49]), connected to the CMS will hold information about both active and closed cases, along with an integrated file server that is responsible for the secure storage of any kind of file evidence associated with the case (e.g. case videos, audio files, suspect or incident photos, or PDF reports). Simultaneously, the communication between the CMS and the other components happens through an API that provides the users with the required endpoints. Eventually, the CMS brings to the table facilitated communication between the scenario stakeholders, clear visualization of complex workflows and transparent process modelling. Through utilizing this approach, users can monitor the progress of each use case, share information between components, trigger actions, and ensure that everything is carried out as it should.

D. Data Spaces

The core architecture of Data Spaces relies on the IDS-RAM [8]. The formation and operation of data exchange systems that maintain sovereignty, preserve privacy of sensitive information, and respect European regulations regarding the ethical use of data is the main objective of IDS-RAM. The accomplishment of these purposes is based on the concept that data holders strictly determine the access and usage rules on their data. Additionally, the maintenance of decentralized storage at the origin of the data sources reduces the need for implementation of central storage and management units. Instead, the involved functionalities for data exchange are formed through the IDS-RAM Components, while data owners maintain their resources and services in their local premises or cloud infrastructures.

The Biometrics Data Space proposed in this paper leverages several components of the IDS-RAM including the basic component for data exchange, which is the Connector, as well as additional components, such as the Identity Provider, the Metadata Broker, and the Clearing House:

Connector: The component that serves as the fundamental element for the data sharing process in a Data Space is the Connector. For the purpose of data exchange, each organization is assigned one or multiple Connectors, through which it is able to request or provide data to the other trusted participants. The Connector is responsible for acting as a secure gateway, while ensuring privacy-preserving artifacts exchange throughout cross-border organizations, and strictly adhering to the data owner's usage policies and the international standards for personal data protection. Data holders possess the capability of determining regulations on the accessibility of their resources in terms of usage, process or sharing. The internal architecture of the Connector constitutes of separate containers that aim at the creation of isolated and secure environment for the secure exchange functionalities. These containers are separated per functionality and they include the data exchange container, the message router and the usage control enforcement application. Moreover, the Connector is responsible for communication with the following components in order to serve basic functionalities, such as certification, registration, or transaction logging.

Identity Provider: In order to maintain reliability throughout the data sharing in a Data Space, the Identity Provider component is responsible for the generation, management, and validation of each corresponding participant's certifications. Unauthorized request or provision of data is prohibited to guarantee trust among the actors. The core role of the Identity Provider is the issuance of certificates and the authorization of Connectors to allow accessing and sharing of data. Its internal architecture relies on three components. Firstly, a Certificate Authority (CA) which is involved the creation and handling of the certificates for each Connector in the Data Space. The authorization is performed through obtaining tokens with short validity range in order to perform access requests through the Connector. These tokens are Dynamic Attribute Tokens (DATs) and they are provided through the Dynamic Attribute Provisioning Service (DAPS), which provides up-to-date information about the Connectors and their assigned participants. Finally, the necessary organization-related information of the participants is registered and maintained in the Participant Information Service (ParIS). Altogether, CA, DAPS, and ParIS fulfill the compulsory

requirements of the Identity Provider in the Data Space that handles identification, authentication, and authorization aspects.

Metadata Broker: The corresponding functionality of a resource and participants catalog in Data Spaces is offered through the Metadata Broker component. Its role is related to data resources management in terms of listing artifacts to a shared registry, and facilitating access to the self-description of the resource, while maintaining anonymization of the owner. A Self-Description is a document that incorporates all the required information of a resource or a Connector registered in the Data Space, to facilitate the data exchange process. During the registration of a Connector in the Data Space, its self-description is registered in the Metadata Broker. In a similar manner, when the Data Offering process of an artifact takes place, its self-description is registered in the Metadata Broker's catalog.

Clearing House: In terms of traceability and transparency, it is of essential importance to register all attempts in the Data Space through the Clearing House component. The main objective of the Clearing house is the provision of a functionality for the registration of access transactions that take place among the trusted participants. The logging of such attempts includes history records with information related to the requestor, the owner, the resource and the usage control policies related to the requested artifact.

E. Blockchain

The proposed blockchain-based data sharing platform aims to facilitate in a fast, secure and valid exchange of sensitive and confidential biometric data between different law enforcement agencies (LEAs) across Europe. The platform can handle different types of data for access, storage and transfer management, due to its "data-agnostic" nature. Its main functionalities focus on (*i*) the management of datasets, image files (suspect photos), and other biometric modalities like voice and fingerprints, as well as (*ii*) enabling interoperibility and integration between different components, participating in the same use case scenario. Its primary goal is to offer a data layer on which, the various connections are built to form the data space. The Blockchain component has two main dependencies in terms of underlying technology. The *Ethereum-based blockchain*, which manages auxiliary storage and offers the trust, immutability, and non-repudiation to the process required to achieve the proper "Solid specified" permissions needed to access the actual data; and the *Solid Community server*, which in turn implements the Solid protocol for safe, private, and decentralized data access [50]. At its core, the data sharing platform is made up of three separate but connected components.

- **Data storage component**: Every data expert / security officer utilizes the data storage, access, and retrieval mechanism, which is based on a customized deployment of the Solid data server for each LEA and runs locally in each organization's premises.
- **Permissions and Accounting Component**: The permissions and accounting component is based on a proof-of-authority blockchain network owned by a private consortium. It can be deployed in a fully decentralized manner on the local premises of each organization, or in a semi-decentralized manner that is more user-friendly and hosted on a cloud-based consortium server.
- **API Component**: The API componentn is a gateway, implemented in tandem with the data storage mechanism. Through this gateway, all parts are combined into a single service, standardizing the authorized actions on the service and streamlining all interactions.

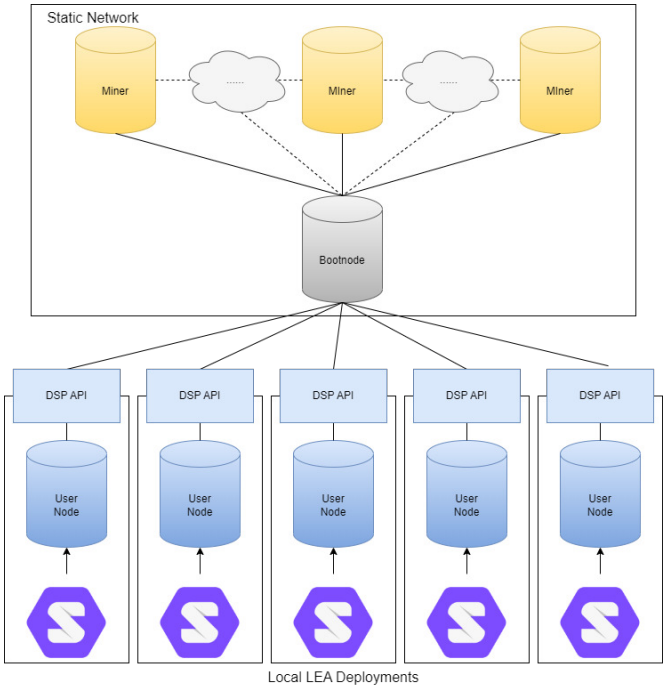


Fig. 2: Blockchain Architecture Diagram

F. Privacy Enhancing Technologies

1) Large-Scale Indexing: The Large-Scale Biometric Data Indexing, is a crucial tool for enabling the efficient comparison and retrieval of biometric data using deep learning techniques, specifically designed to support cross-border collaboration between Law Enforcement Agencies (LEAs).Through the application of deep hashing, it transforms high-dimensional biometric data, such as facial images, fingerprints, and voice samples, into compact, efficient representation for storage, retrieval, and comparison. Its primary goal is to facilitate the identification of suspects by enhancing the speed and accuracy of biometric data processing. As the volume of available biometric data grows, manual comparison by LEAs has become increasingly impractical; this shift to data-driven, algorithmic methods reflects a broader trend in law enforcement, where technologies like deep learning are reshaping policing practices to improve efficiency and accuracy in suspect identification and criminal investigations [51].

The architecture of the Biometric Indexer is built around three critical phases, each of which interdependent, meaning that the successful operation of each subsequent phase depends on the preceding one. Specifically, the Training Phase lays the foundational models required for the generation of hash representations in the Data Indexing Phase. Without the

proper training of these specialised deep learning models, accurate and efficient indexing of biometric data would not be possible. Similarly, the Data Searching Phase depends on the accurate generation and storage of hash representations during the Data Indexing Phase. Without a well-structured and comprehensive index, the search and comparison functions would be ineffective. Therefore, the architecture's functionality is cumulative, where each phase builds upon the last, ensuring that all components work in a cohesive and interconnected manner. The absence or failure of any phase would disrupt the entire system's ability to operate, making it essential that all phases function in sequence for optimal performance.

The Training Phase, involves training three specialised deep learning models, one for each biometric data type. These models are shared with LEAs, allowing them to independently generate hash representations from their original biometric data without the need to transfer this data to a central repository. This decentralised approach enhances security by eliminating the need for data transmission over networks, and it also allows for future expansion. Both new data types and additional LEAs can be seamlessly integrated without disrupting the system's operation.

Once the models are deployed in a distributed manner, the Data Indexing Phase begins. During this phase, each LEA converts its biometric data into collections of hash representations. These collections are stored with pseudonymised suspect IDs, ensuring that personal information is never shared across LEAs. Each LEA maintains a secure internal mapping of suspect IDs to actual individuals, which allows the system to operate without access to sensitive personal data.

The final phase, known as the Data Searching Phase, handles queries from an LEA containing biometric data to be compared against the indexed hash representations. The query data is encrypted using homomorphic encryption and distributed to the other LEAs, where it remains encrypted during the comparison process. Each LEA compares the query against its own stored data, compiles a list of results sorted by similarity, and returns the results to the querying LEA. These results are grouped by LEA and displayed to the requesting authority, ranked according to a predefined similarity scale: "Highly Likely Suspect," "Probable Suspect," "Unlikely Suspect," "Doubtful Suspect," and "Highly Doubtful Suspect."

This three-phase architecture ensures the system's efficiency, scalability, and security. Homomorphic encryption plays a key role by allowing comparisons to occur while data remains encrypted, thereby preserving the privacy of biometric information. This decentralised, secure, and scalable design makes the system highly effective for large-scale, cross-agency biometric identification.

2) Homomorphic Encryption: In the Biometrics Data Space architecture, after extensive research into different homomorphic encryption (HE) libraries, we selected Microsoft's SEAL library as the core HE tool. SEAL was chosen due to its robust implementation of the CKKS (Cheon-Kim-Kim-Song) scheme, which supports approximate arithmetic on encrypted data—ideal for our deep hashing similarity computations. The integration of SEAL allows the system to compute similarity scores between encrypted biometric feature hashes without needing to decrypt them, preserving the privacy of the underlying biometric data at all times.

The technical process begins with the extraction of feature vectors from biometric data through deep hashing algorithms. These vectors, typically representing high-dimensional feature space, are then encrypted using SEAL's CKKS scheme. The advantage of CKKS is its ability to handle real numbers, which is essential for calculating similarity metrics such as cosine similarity or Euclidean distance between hashed features. Once encrypted, these feature vectors are passed to the homomorphic computation module, where similarity calculations are performed directly on the ciphertexts. This approach ensures that even during computation, sensitive biometric data remains protected from potential leakage.

The decision to use SEAL was further justified by its efficient performance and support for customizable encryption parameters, allowing us to balance security and computational overhead. By adjusting the parameters—such as polynomial modulus degree and coefficient modulus—we optimized the encryption process for real-time requirements. This makes SEAL a crucial component in securely comparing biometric feature hashes across jurisdictions, ensuring compliance with cross-border data protection laws while maintaining high accuracy in matching results.

Figure 3 illustrates the user interface (UI) for the homomorphic encryption component in functional testing. The UI consists of three main sections: 1) the data input and processing section (top panel), where fingerprint data is entered in JSON format, including details such as image ID, minutiae type, and coordinates, and processed by clicking "Process Data" to initiate encryption; 2) the encryption details and quality metrics section (middle panel), which displays the encryption and decryption results, including the encrypted data location, execution time, and quality assessment of the fingerprint data based on homomorphic calculations, demonstrating the capability for cross-border data comparison by allowing similarity computations across encrypted data from different jurisdictions; and 3) the encrypted ciphertexts display section (bottom panel), which shows the encrypted data as polynomial expressions, representing the feature values in ciphertext form, ready for similarity computations without decryption, thereby preserving data privacy.

3) Fuzzy Extractor and AES Encryption: In the described architecture, the combination of a Fuzzy Extractor (FE) and AES-256 encryption provides robust protection for biometric data. The Fuzzy Extractor is responsible for transforming noisy biometric inputs, such as fingerprints or facial templates, into stable cryptographic keys. We implemented the FE using BCH (Bose-Chaudhuri-Hocquenghem) coding along with Hamming distance metrics to effectively handle the inherent variability in biometric data. The use of BCH coding allows for efficient error correction, ensuring that slight variations in biometric readings do not result in key mismatches, while Hamming distance ensures accurate comparison between stored templates and new inputs.

The cryptographic seed derived from the FE is then used to generate the secret key for AES-256 encryption, which secures the raw biometric data. By encrypting this data with AES-256,

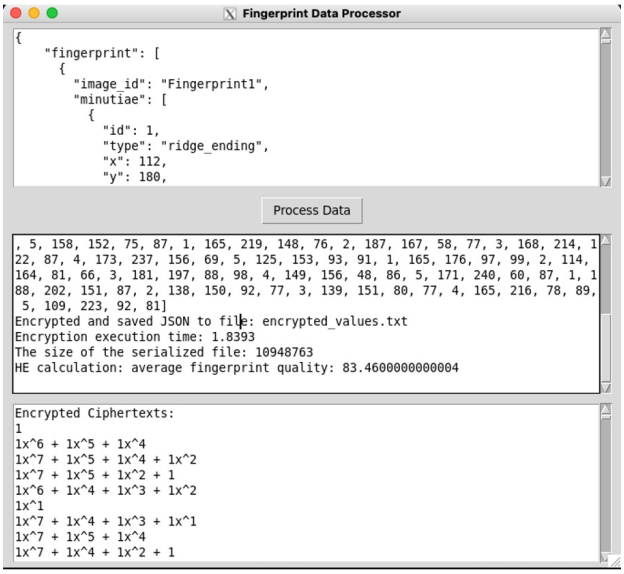


Fig. 3: Homomorphic Encryption Component UI

the architecture ensures that even if intercepted, the biometric data remains protected and unreadable without the correct key. This combined approach mitigates key leakage risks, as the key is generated from the biometric data itself, removing the need for separate key management. Additionally, the error-correction capability of BCH coding enables consistent key generation despite biometric noise, enhancing system usability and security.

G. Scalability

In the context of the Biometrics Data Space architecture, the ability to scale up is provided both in terms of utilized biometric technologies and in cases of expansion with multiple participants. The suggested framework is currently designed to evaluate facial, voice, and fingerprint recognition for the purpose of suspect identification. Nonetheless, the potentiality for extension with more biometric features is feasible.

Regarding scalability by network effects, it is significantly crucial that the proposed model can be extended to support new participants in the automated biometric data exchange framework. The current scenario being considered encompasses two organizations, one that acts as a data access requestor and the other one as a data holder. The notion of a Data Space can be extended to support the participation of more than two organizations by easily scaling up through the assignment and initialization of the Connector, which is the core component of the Data Space and is responsible for handling the data sharing. Each trusted participating organization is assigned to one certified Connector, which is able to operate either in Consumer mode or in Provider mode. The Consumer Connector is responsible for initiating a data access or usage request, while the Provider Connector handles the biometric data offering, the usage control enforcement, and the authorization of data exchange.

In order to scale up and become a trusted participant to the existing Data Space, a new LEA is assigned a new Connector instance operating in the preferred mode. Based on the International Data Spaces Reference Architecture Model [8], the Connector follows predefined procedures including the Onboarding, where they communicate with the central Identity Provider to get certification and authorization in the Data Space, and the Data Offering of the Connector's self-description document to the Metadata Broker, for it to be discoverable by the other trusted participants. In the selected Connector implementation of "FIWARE TRUE (TRUsted Engineering)" Connector by Engineering [11], such procedures can be easily initiated by REST API endpoints. Afterwards, the Connector constitutes a part of the existing Biometrics Data Space and is ready for data offering, contract negotiation and data exchange.

V. Suspect Identification Application Scenario

The Biometrics Data Space framework for the facilitation of the cross-border suspect identification scenario combines the technologies mentioned in the architecture in a way that ensures trust, sovereignty and interoperability throughout biometric data sharing. The processes that comprise the suspect identification workflow are divided in four main steps:

1) The LEA aiming to conduct cross-border suspect identification through the Biometrics Data Space initiates the flow by requesting the comparison process to begin. During the step depicted in Figure 4a, the LEA officer makes available to the Data Space the captured data of the potential suspect to be identified. These biometric samples are forwarded to the Connector of the Data Space which is responsible for enabling the Large-Scale Indexing component for the hash creation. Once the biometric data is hashed, it is forwarded to the Homomorphic Encryption component for encryption. The hashed and homomorphically encrypted data are provided to the Connector of the Consumer for the process of matching to take place.
2) The next phase of the suspect identity verification scheme involves the distributed comparison of the biometric data for the acquisition of the matches on existing databases, as depicted in Figure 4b. The specific process begins with the retrieval of participants in the Biometrics Data Space through the Metadata Broker component. Once the list of the participants and their metadata is retrieved, the Consumer Connector requests the distributed comparison of the captured biometric data with the existing storage of each trusted participant through its own assigned Connector. Eventually, the aggregated list of matches that exceed a predefined threshold is returned to the LEA officer.
3) Before proceeding to the actual data exchange, the LEA officer is able to initiate the request for access to selected suspect profiles. The completion of this step requires the creation of a blockchain transaction related to the access request through an external Smart Wallet service. The outcome of this interaction equips the Consumer with the necessary information to manually complete the access request signing. This process is analytically described in Figure 4c.

4) The final step facilitates the process of the data exchange between the Data Consumer and the Data Provider LEAs and is depicted in Figure 4d. During this process, the request for suspect profiles access is forwarded from the Consumer Connector to the Provider Connector. The Data Provider gets notified about the request and determines the response to access and usage policies. Upon approval of access, the Data Provider follows the process for signing the response transaction to the Blockchain through an external Smart Wallet. Once this process is accomplished, the data is securely exchanged from the Provider Connector to the Consumer Connector. Since the exchanged suspect profile is encrypted with AES-256, the Consumer Connector requests its decryption in order to display the decrypted biometric data and suspect's information to the requestor LEA officer for further offline analysis.

VI. Conclusions

Secure systems for suspect identification in cross-border scenarios are significantly important towards the confrontation of terrorism and global crime. Existing approaches face several challenges in the case of police cooperation through data exchange due to the necessity for compliance with EU regulations on personal data privacy protection. The proposed Biometrics Data Space framework of this paper addresses these limitations by (1) enabling trustworthy suspect profile exchange between cross-border LEAs, (2) integrating strong safeguards on data protection through the utilized PETs (Encryption & Large-Scale Indexing), and (3) ensuring traceability on resource sharing history for immutability through the Blockchain. Altogether, the proposed framework achieves self-sovereign, trustworthy and interoperable data exchange between cross-border LEAs for the purpose of suspect identification.

VII. Acknowledgments

This work has received funding from the EU Horizon Europe research and innovation programme through TENSOR project under GA: 101073920.

References

[1] H. Pettenpohl, M. Spiekermann, and J. R. Both, *International Data Spaces in a Nutshell*. Cham: Springer International Publishing, 2022, pp. 29–40. [Online]. Available: https://doi.org/DOI:10.1007/978-3-030-93975-5_3

[2] P. Luif, "The treaty of prüm: A replay of schengen?" 2007.

[3] "Stepping up cross-border cooperation – the prüm decision — eur-lex." [Online]. Available: https://eur-lex.europa.eu/EN/legal-content/summary/stepping-up-cross-border-cooperation-the-pr-m-decision.html

[4] "EUR-Lex - 52021PC0784 - EN - EUR-Lex, "Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on automated data exchange for police cooperation ("Prüm II"), amending Council Decisions 2008/615/JHA and 2008/616/JHA and Regulations (EU) 2018/1726, 2019/817 and 2019/818 of the European Parliament and of the Council"." [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2021%3A784%3AFIN

[5] "EUR-Lex - 32016R0679 - EN - EUR-Lex, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)" ." [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679

[6] M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, p. 160018, 2016.

[7] "Member States and Commission to work together to boost artificial intelligence "made in Europe"." [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/IP_18_6689

[8] B. Otto, S. Steinbuss, A. Teuscher, and S. Lohmann, "IDS Reference Architecture Model," Jul. 2021. [Online]. Available: https://doi.org/DOI:10.5281/zenodo.5105529

[9] Gaia-X Association, "Overview - Gaia-X Architecture Document - 22.10 Release," 2022. [Online]. Available: https://docs.gaia-x.eu/technical-committee/architecture-document/22.10/overview/

[10] M. H. G. Giussani, S. Steinbuss and N. Gras, "Data Connector Report," Jan. 2024. [Online]. Available: https://doi.org/DOI:10.5281/zenodo.10591027

[11] "FIWARE TRUE (TRUsted Engineering) Connector: easing data sharing in Gaia-X." [Online]. Available: https://www.eng.it/en/case-studies/true-connector-per-facilitare-la-condivisione-di-dati-in-gaiax

[12] "Fiware – foundation." [Online]. Available: https://www.fiware.org/foundation/

[13] D. Puthal and S. P. Mohanty, "Proof of authentication: Iot-friendly blockchains," *IEEE Potentials*, vol. 38, no. 1, pp. 26–29, 2018.

[14] S. Yu, Y. Huang, L. Wang, Y. Makihara, S. Wang, M. A. Rahman Ahad, and M. Nixon, "Hid 2022: The 3rd international competition on human identification at a distance," in *2022 IEEE International Joint Conference on Biometrics (IJCB)*, 2022.

[15] A. Al-Khazzar and N. Savage, "Graphical authentication based on user behaviour," in *2010 International Conference on Security and Cryptography (SECRYPT)*. IEEE, 2010, pp. 1–4.

[16] O. Protocol, "Data: The new asset class — ocean protocol," 2024. [Online]. Available: https://oceanprotocol.com/

[17] IOTA, "Iota," 2024. [Online]. Available: https://iota.org/

[18] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song, "Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication," *IEEE transactions on information forensics and security*, vol. 8, no. 1, pp. 136–148, 2012.

[19] IPFS, "Interplanetary file system," 2024. [Online]. Available: https://ipfs.tech/

[20] R. V. Yampolskiy and V. Govindaraju, "Behavioural biometrics: a survey and classification," *International Journal of Biometrics*, vol. 1, no. 1, pp. 81–113, 2008.

[21] Y. Dodis, L. Reyzin, and A. Smith, "Fuzzy extractors: How to generate strong keys from biometrics and other noisy data," in *Advances In Cryptology-EUROCRYPT 2004: International Conference On The Theory And Applications Of Cryptographic Techniques, Interlaken, Switzerland, May 2-6, 2004. Proceedings 23*. Springer, 2004, pp. 523–540.

[22] R. Oak and M. Khare, "A novel architecture for continuous authentication using behavioural biometrics," in *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*. IEEE, 2017, pp. 767–771.

[23] A. Gani, A. Siddiqa, S. Shamshirband, and F. Hanum, "A survey on indexing techniques for big data: taxonomy and performance evaluation," *Knowledge and information systems*, vol. 46, pp. 241–284, 2016.

[24] L. Chi and X. Zhu, "Hashing techniques: A survey and taxonomy," *ACM Computing Surveys (Csur)*, vol. 50, no. 1, pp. 1–36, 2017.

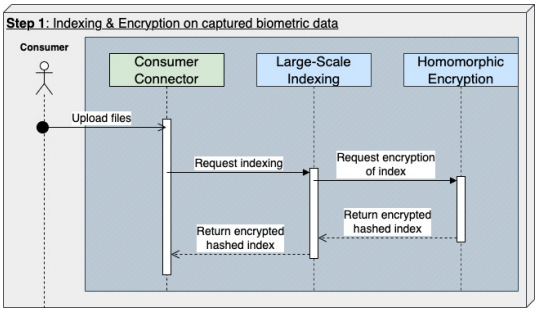
[25] X. Luo, H. Wang, D. Wu, C. Chen, M. Deng, J. Huang, and X.-S. Hua, "A survey on deep hashing methods," *ACM Trans. Knowl. Discov. Data*, vol. 17, no. 1, Feb. 2023. [Online]. Available: https://doi.org/10.1145/3532624

[26] D. Zhong, H. Shao, and X. Du, "A hand-based multi-biometrics via deep hashing network and biometric graph matching," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 12, pp. 3140–3150, 2019.

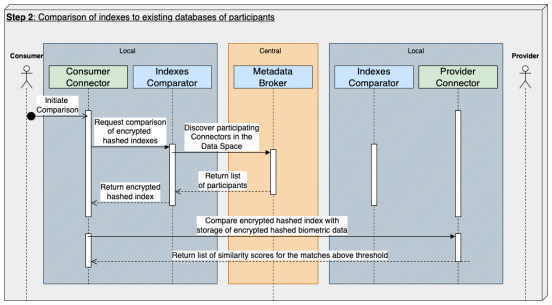
[27] H. Shao, D. Zhong, and X. Du, "Efficient deep palmprint recognition via distilled hashing coding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[28] X. Du, D. Zhong, and H. Shao, "Cross-domain palmprint recognition via regularized adversarial domain adaptive hashing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 6, pp. 2372–2385, 2020.

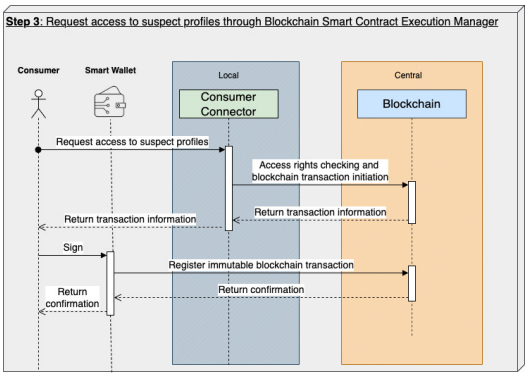
[29] F. Zou, F. Yang, W. Chen, K. Li, J. Song, J. Chen, and H. Ling, "Fast large scale deep face search," *Pattern Recognition Letters*, vol. 130, pp. 83–90, 2020.



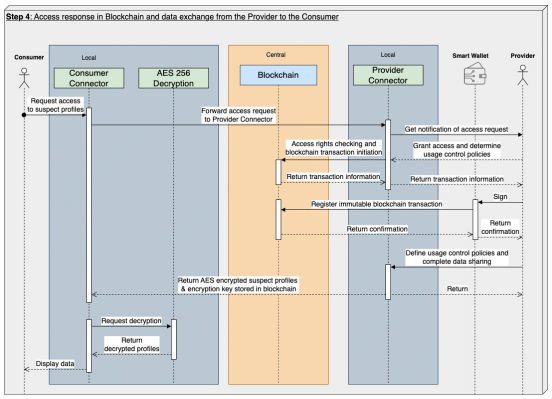
(a) Step 1 of cross-border data exchange: Large-Scale Indexing on the biometric data artifacts and Homomorphic Encryption of the produced hashed indexes.



(b) Step 2 of cross-border data exchange: Distributed comparison of the encrypted hashed indexes of Step (1) to determine the list of similarity scores that exceed a predefined threshold.



(c) Step 3 of cross-border data exchange: Blockchain Smart Contracts Execution Manager creation of transaction for access request by the Consumer.



(d) Step 4 of cross-border data exchange: Blockchain creation of transaction for access response by the Provider for the Consumer and data exchange facilitation between the assigned Connectors

Fig. 4: Suspect Identification workflow main steps through the utilization of the Biometrics Data Space.

[30] F. Zhu, X. Kong, L. Zheng, H. Fu, and Q. Tian, "Part-based deep hashing for large-scale person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4806–4817, 2017.

[31] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *Proceedings of the 41st annual ACM symposium on Theory of computing*. ACM, 2009, pp. 169–178.

[32] J. Fan and F. Vercauteren, "Somewhat practical fully homomorphic encryption," *IACR Cryptology ePrint Archive*, vol. 2012, p. 144, 2012.

[33] Z. Brakerski and V. Vaikuntanathan, "Lattice-based fhe as secure as pke," *IACR Cryptology ePrint Archive*, vol. 2014, p. 46, 2014.

[34] S. Costanzo, D. Laudenbach, and L. Schrettner, "Efficient privacy-preserving biometric identification," in *International Conference on Information Security and Cryptology*. Springer, 2015, pp. 266–284.

[35] Z. Brakerski and V. Vaikuntanathan, "Efficient fully homomorphic encryption from (standard) lwe," in *2014 IEEE Symposium on Security and Privacy*. IEEE, 2014, pp. 578–594.

[36] C. Juvekar, V. Vaikuntanathan, and A. Chandrakasan, "Gazel: A secure and efficient scheme for cloud-assisted medical diagnosis," in *Proceedings of the 27th USENIX Security Symposium*, 2018, pp. 5–21.

[37] D. Froelicher, J.-L. Raisaro, J. R. Troncoso-Pastoriza, and J.-P. Hubaux, "Unlynx: A decentralized system for privacy-conscious data sharing," in *Proceedings of the 26th USENIX Security Symposium*, 2017, pp. 311–328.

[38] M. Blanton, M. Aliasgari, and A. Steele, "Secure and efficient protocols for iris and fingerprint identification," in *Proceedings of the 7th ACM workshop on Privacy in the electronic society*. ACM, 2011, pp. 175–184.

[39] Z. Erkin, E. Franz, J. Guajardo, S. Katzenbeisser, I. Lagendijk, and T. Toft, "Privacy-preserving face recognition," *International Journal of Information Security*, vol. 12, pp. 19–32, 2009.

[40] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Order-preserving encryption for numeric data," in *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*. ACM, 2004, pp. 563–574.

[41] Y. Dodis, L. Reyzin, and A. Smith, "Fuzzy extractors: How to generate strong keys from biometrics and other noisy data," in *International conference on the theory and applications of cryptographic techniques*. Springer, 2008, pp. 523–540.

[42] A. Juels and M. Wattenberg, "Fuzzy commitments," in *Proceedings of the 6th ACM conference on Computer and communications security*. ACM, 1999, pp. 28–36.

[43] X. Boyen, "Reusable fuzzy extractors for low-entropy distributions," in *Proceedings of the 11th ACM conference on Computer and communications security*. ACM, 2004, pp. 82–91.

[44] R. Renner and A. Smith, "Security of fuzzy extractors," in *IACR International Conference on Theory of Cryptography*. Springer, 2008, pp. 499–517.

[45] C. Rathgeb and A. Uhl, "Survey of biometric template protection," vol. 14, no. 3, 2011, pp. 760–780.

[46] T. C. Clancy, N. Kiyavash, and D. J. Lin, "Secure smartcard-based fingerprint authentication," in *Proceedings of the ACM SIGMM workshop on Biometrics methods and applications*. ACM, 2003, pp. 45–52.

[47] K. C. Mangold, "Data format for the interchange of fingerprint, facial & other biometric information ansi/nist-itl 1-2011 nist special publication 500-290 edition 3," 2016-08-21 04:08:00 2016. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=921456

[48] CAMUNDA, "Camunda: The universal process orchestrator," 2024. [Online]. Available: https://camunda.com/

[49] MongoDB, "Mongodb: The developer data platform," 2024. [Online]. Available: https://www.mongodb.com/

[50] S. Project, "Solid project," 2024. [Online]. Available: https://solidproject.org/

[51] A. G. Ferguson, "The rise of big data policing: Surveillance, race, and the future of law enforcement," in *The Rise of Big Data Policing*. New York University Press, 2017.