Artificial Intelligence and Interoperability for Solving Challenges of OSINT and Cross-Border Investigations

Amr el Rahwan
International Hellenic University[1]

Abstract

The major investigation challenges are summarised as multiple-identity, fraudulent actions, lack of interoperability and absence of an effective technical solution for exchanging Cross-Border information, and complexity of OSINT investigations.
The EU published Regulations (EU) 2019/817 and 2019/818 for establishing a framework for EU interoperability between information systems in the field of borders and visa information systems, police and judicial cooperation, asylum, and migration. Existing systems such as EURODAC, SIS / SISII, and VIS must share data, and new systems such as ECRIS-TCN, EES, and ETIAS also need to follow these guidelines. Although the eu-LISA will implement the interoperability framework in 2023, new challenges will emerge, such as investigating multiple-identity and identity frauds due to the different formats and structures of data, low quality of biographic and biometric data, and low accuracy of matching algorithms.
Furthermore, the Open Source Intelligence (OSINT) investigation process is not automated, consumes a lot of time, and is overwhelming. When border security and law enforcement officers use methods of OSINT to investigate terrorism and serious crime, it is very difficult to match and link the identity-related data and facial images of the suspects stored in the EU systems, Cross-Border systems, and open sources.
The paper argues different Artificial Intelligence (AI) methods and algorithms and interoperability could be the optimum solution for the challenges mentioned above. The paper highlights a Person-Centric approach using Artificial Intelligence and interoperability to solve the challenges that emerge during investigations, such as multiple-identity, identity frauds, exchanging Cross-Border information, and the complexity of OSINT investigations.

Keywords: Multiple-Identity, OSINT, Interoperability, Cross-Border, Artificial Intelligence

1 Author's email: amr.rahwan@secureidentityalliance.org

Introduction

The paper highlights using artificial intelligence and interoperability for solving the challenges of OSINT and Cross-Border investigations. The four major challenges are multiple-identity, fraudulent identity, cross-border investigations, and OSINT complexity. The multiple-identity and fraudulent identity challenges exist on the national level of the European Member States, and more challenges will emerge between the national level and the central EU level after implementing the new interoperability architecture. The newly established central and national ETIAS "European Travel Information and Authorisation System" units will face the challenge of confirming or rejecting the relations and links between the different encounters of the multiple-identity and fraudulent identity. The main challenge for cross-border investigation is the difficulty of exchanging cross-border information and the non-existence of a proper interoperable information system or a technical solution for exchanging information related to the cross-border investigation. The complexity of OSINT is challenging because only officers with strong information technology skills and background can obtain optimum results from OSINT investigations. In contrast, detectives and investigators with basic IT skills can't obtain good results from OSINT investigations, either for investigations for solving national or cross-border crimes.

Furthermore, the paper highlights the relevant technologies that could be used for solving the mentioned challenges, especially using interoperability and pre-trained Artificial Intelligence algorithms. Moreover, understanding the existing technology limitations is essential for obtaining good results and recommending the best practice for achieving optimal results. Furthermore, introducing a new Person-Centric OSINT approach complies with the UMF "Universal Message Format" standard of European interoperability. The newly introduced Person-Centric OSINT approach will allow the detectives and investigators with basic IT skills to achieve good results in identifying suspects and victims of terrorism and serious crimes without being overwhelmed with learning advanced IT or OSINT.

Moreover, the paper presents three hypothetical cases, recommends the HORUS system for SSI "Single Search Interface" as a practical technical solution for cross-border interoperability and exchanging of cross-border information, and simulating an automated search scenario for identifying an unknown terrorist.

Finally, the paper describes the required training for law enforcement officers in each Member State, and it concludes the required training for compliance with EU interoperability standards, the required support for purchasing and implementing AI, interoperability, and Single Search Interface, the required capacity building for technical, functional, and operational officers, and essential AI training on Facial Recognition and Person-Centric OSINT for cross-border investigations.

Challenges

Multiple-Identity

The central EU information systems were implemented in silos, creating information gaps due to a lack of interoperability. Implementing the information systems in silos has created challenges for detecting incorrect, incomplete, or fraudulent identities.

Subsequently, on the 22nd of May 2019, the EU published two new regulations. Regulation (EU) 2019/817: establishing a framework for EU interoperability between information systems in the field of borders and visa information systems (Council Regulation (EC) 817/2019). Regulation (EU) 2019/818: establishing a framework for EU interoperability between information systems in the field of police and judicial cooperation, asylum, and migration (Council Regulation (EC) 818/2019). Article (38) of the regulations established the UMF "Universal Message Format" standard to achieve interoperability.

The need to improve EU interoperability is clear. Existing systems such as EURODAC, SIS / SISII, and VIS must share data, and new IT systems such as ECRIS-TCN (Council Regulation (EC) 816/2019), EES (Council Regulation (EC) 2226/2017), and ETIAS also need to follow these guidelines. That must be done without adding new databases or changing access rights to existing systems.

The components needed as part of the move towards EU interoperability include the following: the European Search Portal (ESP) for fast and seamless simultaneous searches in EU information systems, in addition to Europol and Interpol data; the Shared Biometric Matching Service (sBMS) (European Union Agency for the

Operational Management of Large-Scale IT Systems in the Area of Freedom, Security and Justice, 2018) that searches and compares biometric data (fingerprints and facial images), linking this data to other systems; the Common Identity Repository (CIR) (European Union Agency for the Operational Management of Large-Scale IT Systems in the Area of Freedom, Security and Justice, 2018) to increase the accuracy of identification through automated comparison and matching, and the Multiple Identity Detector (MID) (European Council: Council of the European Union, 2019) for automatic detection of multiple identities linked to the same set of biometric data.

However, many common challenges will emerge due to the different formats and structures of data, low quality of biographic and biometric data, low accuracy of matching algorithms, errors in data entry, and fraudulent actions. For example, when the border authorities receive the Advance Passenger Information (API) and the Passenger Name Record (PNR) of air and sea passengers, it is difficult to exchange and match the identity of one passenger with his/her records stored in the EES, ETIAS, SIS, and VIS due to lack of interoperability and different data structures and formats. Another example of these challenges is the car license plate number. The license plate number has different formats and structures that vary from one Member State to another, creating difficulties in searching and finding the correct license plates and linking them with individuals, such as owners or suspects.

TCNs, or Third Country Nationals, are mainly the persons of interest stored in the EU information systems for different purposes, except for SIS, which also stores information about European citizens. The SIS / SIS-II (Council Regulation (EC) 1862/2018) stores security alerts on persons wanted by the Member States, and the officers can search the central information systems with biometric data such as fingerprints or biographic data such as first name, family name, gender, date of birth, place of birth, and nationality to find targeted persons or search object alerts such as A Vehicle; a Firearm; a Blank Document; an Issued Document; a Banknote; an Industrial Equipment; an Aircraft; a Boat; a Boat Engine; a Container; a License Plate; a Security; a Vehicle Registration Document. The EURODAC (Council Regulation (EC) 2013/603) system stores biometric information such as facial images, fingerprints, and identity-related biographic information of asylum seekers and illegal border crossers. The officers can search the central system by any element of the stored information. The VIS or Visa Information System stores the information such as facial images, fingerprints, name, gender, date of birth, place of birth, nationality, and address of the TCNs travelling with a short-stay visa, and the authorities have up to fifteen working days for vetting the travelers and checking for security clearance.

Important to mention that the central EU systems have gaps in covering all the persons of interest living or travelling to the Member States of the European Union. The gap could be summarised in three types of persons of interest: the short stay visa-exempted third country travellers, permanent foreign residents, and EU citizens. The eu-LISA will implement the ETIAS system and units for solving the gap for the visa-exempted TCNs. However, none of the existing or newly established central European information systems will solve the gap for permanent TCN residents and EU citizens. Each Member state is responsible for solving that gap by creating national systems and achieving interoperability between the national and central information systems as per the EU regulations for interoperability. Clause 22 of regulations (EU) 2019/817 and 2019/818 states that [Member States dispose of efficient ways to identify their citizens or registered permanent residents in their territory], so each Member State is responsible for solving the gap and issue related to its citizens and permanent residents to avoid security vulnerabilities and to reveal their identities if they became suspects or victims of terrorism or serious crime.

The security authorities have enough time to apply security check and clearance on the travellers on a standard short-stay visa with their information stored at the VIS. At the same time, only 48 hours are available to perform security clearance of the visa-exempted third country visitors as mentioned in the regulation (EU) 2018/1240 on establishing a European Travel Information and Authorisation System ETIAS (Council Regulation (EC) 1240/2018). The ETIAS will solve the existing security gap of the visa-exempted TCNs. However, the central and national ETIAS unit officers should be well-trained to solve the multiple-identity issues. The visa-exempted visitor will apply for a travel authorisation before arrival to the EU Member State. The visitor will submit identity-related information such as a facial image and biographical data, which will be stored and processed by the ETIAS. The identity-related information will be searched against all the central EU information systems to check the former existence of the

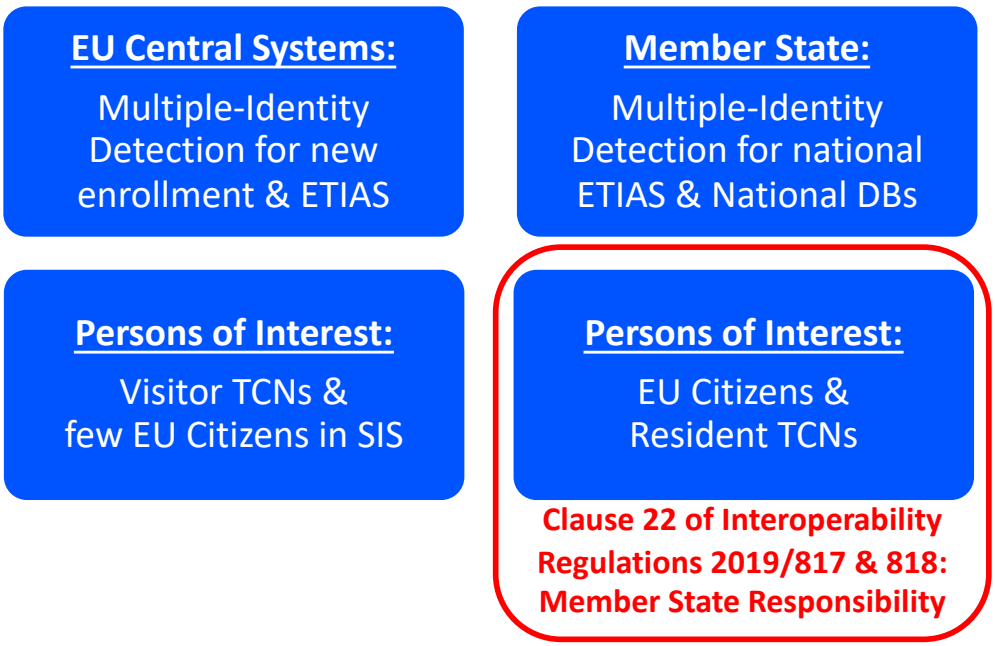
visa-exempted applicant in other EU systems than the ETIAS, and the central MID, Multiple-Identity Detector, will automatically flag the identities with similarities based on biometric matches or biographic matches. The ETIAS unit officers have to manually investigate all the elements of the multiple-identities and confirm or reject the link between identities.

Similarly, the multiple-identity issue may occur when using methods of OSINT to gather more information on suspects, victims, or travellers. The multiple-identity issue gets more complex if the identity-related results of OSINT search is in a language other than the language of the information stored in the EU or Member States information systems. For example, the name, gender, date of birth, place of birth, and nationality are stored using a Latin-based script in the EU information systems. It is very challenging for the officers, detectives, and investigators to decide on the similarities or differences of a multiple-identity with biographic information received from OSINT results and written in Arabic, Cyrillic, Chinese, Greek, Japanese, or Korean scripts. Especially if the officers didn't read or understand the foreign script. The first case of the cases section will clarify an example of multiple-identity.

Identity Fraud

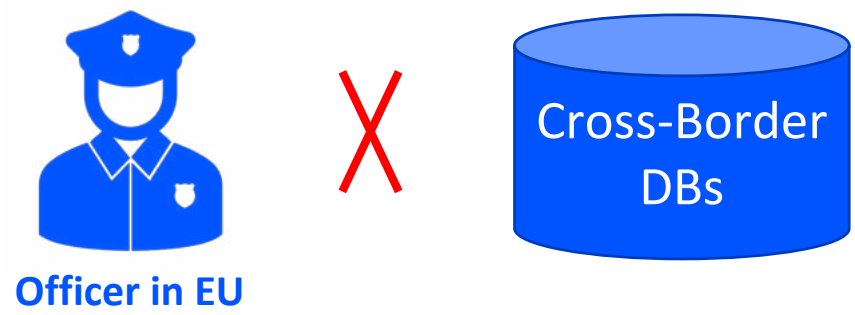
The fraudulent actions and wrong matches are other issues created due to the lack of interoperability and low accuracy of some biometric modalities. For example, the fingerprints of a third-country national could be enrolled in the VIS system with specific identity information, while the fingerprints of the same third-country national might be enrolled in the EURODAC system using different identity information. A second example is that the different facial images of a third-country national could be enrolled in the VIS and EURODAC systems. When submitting a facial query to both systems, the results could be two lists of candidates, instead of one "hit/no hit" from each system, due to the low quality of facial images and the low accuracy of facial recognition algorithms.

Finally, when the border security officers and the law enforcement officers use methods of Open Source Intelligence (OSINT) to investigate terrorism and serious crime, it is very difficult to match the identity-related data and facial images of the suspects stored in the EU systems with the data from open sources. Moreover, most law enforcement and border security officers' basic information technology skills are insufficient for detecting fraudulent identities when using OSINT for investigations. The officers should receive advanced biometric training, especially facial recognition training, and Person-Centric OSINT training to be qualified to detect, investigate, and match identity frauds from open source. The second case of the cases section will clarify an example of identity fraud.



Cross-Border Investigation

Cross-Border information exchange is required when revealing the identity of an involved suspect or victim depending on identity information or criminal information that resides in a foreign country outside the borders of European countries. Furthermore, exchanging of cross-border information is required by immigration authorities for the identification and security clearance of TCN asylum seekers and travellers. Cross-Border investigations are challenging because there is no proper way or technical solution for exchanging cross-border information, and the officers in the EU countries don't have access to the cross-border databases and information systems. The third case of the cases section will simulate the challenge and the solution for a valid hypothetical scenario for cross-border investigation.

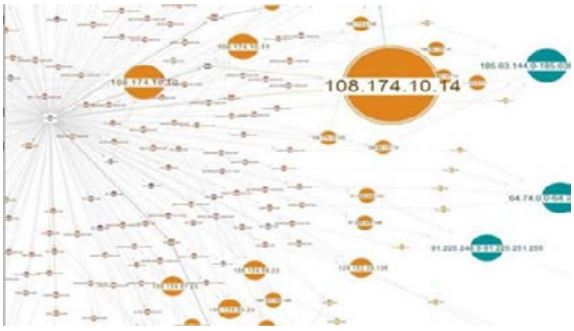


OSINT Complexity

Using tools and methods of OSINT is challenging because it contains various information technology elements such as domains, websites, protocols, headers, codes, scripts, IP addresses, certificates, hashes, usernames …etc. It requires strong IT skills to obtain optimum results in revealing the identities of suspects or victims related to terrorism or serious crime. Moreover, it is difficult to match the suspects' identity-related data and facial images stored across the different databases with the data from open sources. For example, a suspect has a record stored in a national or European database such as SIS or EURODAC. The stored record might be biographic data or a facial image. When the suspect has a different identity on the internet and social media, it is difficult to link the identity stored in the national and EU databases with the fraudulent identity claimed on the internet and social media.

Furthermore, the officers don't get the optimum results from the OSINT tools because they need to understand the tools' mechanism, accuracy, and demographics. Also, they may not differentiate between image recognition and facial recognition in many cases. For example, it is important to understand which type of human images could return good results when searching with tools such as Google, Bing, and Yandex. Those OSINT tools are Artificial Intelligence algorithms for image recognition, not facial recognition. Another example is the facial Recognition AI algorithms used for OSINT have limitations due to their recognition mechanism, the accuracy of algorithms, geographic coverage, and ethnicity bias. Understanding the limitations will lead to optimum results when using such OSINT tools dedicated to facial recognition.

Finally, the different encounters of the same identity are not linked across the different data sources, creating multiple-identity and fraudulent identity challenges due to lack of interoperability and the variations of names and languages.



Artificial Intelligence

The proposal of the European Artificial Intelligence Act defines Artificial Intelligence systems as " 'artificial intelligence system' (AI system) means software that can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with." (Council Regulation proposal (EC) 2021/0106).

Artificial Intelligence technology and interoperability are keys to solving the multiple-identity and fraudulent identity issues. To clarify, the main trigger for the multiple-identity and fraudulent identity issues is that the information is stored in the systems and the databases in silos, and there is no link between the identity-related information stored in the national, European, international, and open sources. There is no SSI "Single Search Interface" till present, and the investigators and officers use separate interfaces for submitting the same queries to national systems, OSINT, and international systems and databases such as Interpol's SLTD "Stolen and Lost Travel Documents", EUROPOL, EURODAC, SIS/SISII, and VIS. When an officer submits a search with one or more identity elements to the different interfaces of data sources, the officer's decision on linking the different encounters of the same identity and discovering frauds depends on factors such as biographic information, name variation, and facial images.

Named Entity Relationship, or NER, is an artificial intelligence method used for automatic extracting, classifying, and categorising the content of a text. NER should be the early step for detectives and investigators when investigating a text written in a language they don't understand. NER will help the investigators understand and target the information useful for investigations, such as names, jobs, and addresses while decreasing the focus on the less useful or less relevant information. The cases section contains three examples for clarifying the practical usage of NER.

Although the stored fingerprints in the EU information systems have good quality, there are challenges to detecting similar or different identities. Natural Language Processing (NLP) AI algorithms can be used for biographic matching across multiple information systems. These algorithms can be trained to link between the different name variations of similar identities. They can detect identity fraud when the same person's fingerprints are enrolled in two systems or more under different identities. The paper demonstrates using an artificial intelligence algorithm for fuzzy name matching, a specific type of Natural Language Processing.

NLP and Named Entity Recognition (NER) AI methods can be combined with domain-specific knowledge to solve the issues related to unstructured data, different data formats, and data mapping. A good example is to search for a license plate number, as each Member State has a different structure and format than the others, and some Member States may have more than one format for license plate numbers. For this example, the AI algorithms will be trained to recognise the license plate number and country of origin. Using representative training data to support a Google-like search and get the best results is essential. However, the paper only presents the pre-trained Artificial Intelligence algorithms.

Using AI for fuzzy name matching for linking the different encounters of similar identities is essential for deciding on similarities and differences between identities triggered by biometric hits such as a fingerprint match or a facial recognition match. Within the paper, AI algorithms are recommended for extracting identity-related information, searching, and matching, while no algorithms will be introduced for anomaly detection or predictive analysis.

Moreover, AI algorithms for image recognition and facial recognition are important for verifying previously known identities and searching for unknown identities. For example, an investigator could search two sources using biographic elements of the identity that resulted in retrieving a facial image from each source. The investigator can use an AI algorithm for facial recognition to verify the facial images. Another example is that the authorities may not have any information about a suspect except a photograph. The authorities can submit the photo to AI algorithms for image recognition and facial recognition to gather more information about the unknown suspect.

The presented concept is to train officers on obtaining the best results from pre-trained commercially available AI algorithms, with any possibility of re-training the AI algorithms.

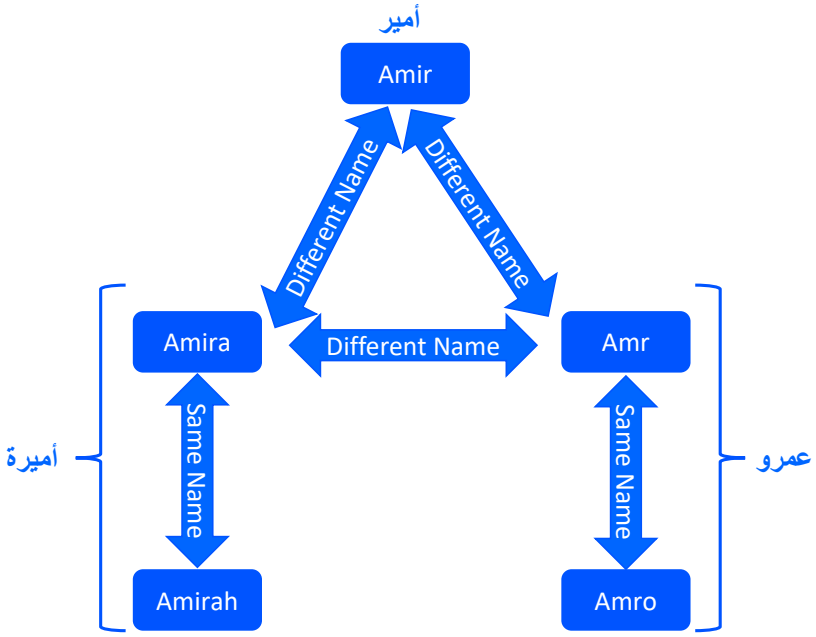
AI for Fuzzy Name Matching

Name matching is essential for linking or unlinking identities. Yet, understanding names is challenging

because the same name is written and pronounced differently across different languages, and the name may have variations due to regional and cultural effects. Furthermore, there are no clear rules for defining nicknames, and a nickname may sound very far from the original name, such as Sasha, a nickname for Alexander. Understanding name variations across different languages and using AI algorithms for fuzzy biographic matching will improve investigation results and solve the problems of multiple-identity and frauds.

In the Arabic language, for example, it is easy for Arabic speakers to identify persons of interest with Arabic names. Still, it is challenging for non-Arabic speakers because the Arabic names have a lot of variations when translated to other languages. Another challenge is that many Arabic letters don't have any phonetical equivalent in Latin-based languages (Sawalha et al, 2014). The below diagram depicts the complexity of name variations of Arabic names.

Figure 1. Complexity of name variations of Arabic names



The above diagram shows an example for three different names, ***Amr"*** عمرو**"**, ***Amir"*** أمير **"**, and ***Amira"*** أميرة**"**, and a good AI algorithm should be able to discover all the variations of one name such as ***Amr*** and ***Amro*** are the same name. ***Amira*** and ***Amirah*** are the same names. The three names are written and pronounced in one way only in Arabic, and Arabic speakers easily distinguish them. However, considering the spoken languages of the European officers, It is difficult for non-Arabic speakers to discover the variations and differentiate between the three names because the names contain letters that don't have phonetical equivalents in Latin-based languages. Each name could be written and pronounced in several ways when translated to other languages. For example, the first sound and letter of the name ***Amr"*** عمرو**"** doesn't exist in English, French, Spanish, Russian, German, Dutch, Italian, or Greek languages, and the letter ***"A"*** is an inaccurate compensation for the letter ***"ع"***, and it is not literal and not correct phonetically. The below table depicts the special phonetics of Arabic letters.

Figure 2. Source: QURANIC WORDS STEMMING (Yusof et al, 2010).

Unicode	*Arabic letter*	*Name*	*SATTS*
0621	ء	hamza	E
0622	آ	alef with madda above	(missing)
0623	أ	alef with hamza above	(missing)
0624	ؤ	waw with hamza above	(missing)
0625	إ	alef with hamza below	(missing)
0626	ئ	yeh with hamza above	(missing)
0627	ا	alef	A
0628	ب	beh	B
0629	ة	teh marbuta	?
062A	ت	teh	T
062B	ث	theh	C
062C	ج	jeem	J
062D	ح	hah	H
062E	خ	khah	O
062F	د	dal	D
0630	ذ	thal	Z
0631	ر	reh	R
0632	ز	zain	;
0633	س	seen	S
0634	ش	sheen	:
0635	ص	sad	X
0636	ض	dad	V
0637	ط	tah	U
0638	ظ	zah	Y
0639	ع	ain	"
063A	غ	ghain	G
0640	ـ	tatwheel	(missing)
0641	ف	feh	F
0642	ق	qaf	Q
0643	ك	kaf	K
0644	ل	lam	L
0645	م	meem	M
0646	ن	noon	N
0647	ه	heh	?
0648	و	waw	W
0649	ى	alef maksura	(missing)
064A	ي	yeh	I

Advanced Artificial Intelligence could help non-Arabic speakers identify and verify name variations for deciding on multiple-identities and frauds. The below table depicts the matching results obtained from a commercial fuzzy name matching AI algorithm for detecting the variations of the Arabic names.

Table 1. Fuzzy name matching AI algorithm

Name 1	Name2	Same Name	Gender	AI Score
Amr	عمرو	Yes	Same	99.0%
Amr	أمير	No	Same	72.7%
Amr	أميرة	No	Different	28.4%
Amr	Amira	No	Different	37.4%
Amr	Amir	No	Same	85.5%
Amir	عمرو	No	Same	72.7%
Amir	أمير	Yes	Same	99.0%
Amir	أميرة	No	Different	80.9%
Amir	Amira	No	Different	51.2%
Amira	عمرو	No	Different	60.9%
Amira	أمير	No	Different	80.3%
Amira	أميرة	Yes	Same	98.2%

Wrong Match

Wrong Match

The above results obtained by the AI algorithms for the three names help determine the similarities and differences between the variations of Arabic names. However, ***Artificial Intelligence is not an absolute source of truth***, and the existing AI algorithms for fuzzy name matching have issues matching Arabic names with their Latin variations. They are not fully matured and not well-trained, and they might wrongly create a high confidence score when matching two different Arabic names. For example, the names ***Amr*** in Latin letters and ***Amir"*** أمير ***"*** in Arabic letters in the second row are two different male names, but the similarity score is higher than 70% which is not correct. The same applies to the names ***Amir*** and ***Amr"*** عمرو***"*** in the sixth row. In the fifth row, the names ***Amr*** and ***Amir*** are wrongly matched with 85.5% because the number of letters is small, and the phonetical difference is minor when pronounced with a Latin-Based language; nevertheless, the two names are very different phonetically when written and pronounced in Arabic. Finally, the names ***Amir"*** أمير ***"*** and ***Amira"*** أميرة***"*** in the eighth and eleventh rows are matched with a score over 80%, although the genders are different because Amir is a male and Amira is a female. The AI algorithm should consider the genders while matching names, especially since the genders already exist in its knowledge base.

AI for Image Recognition vs Facial Recognition

Both Image Recognition and Facial Recognition technologies are based on analysing images. Still, the major difference is that image recognition analyses the whole image for detecting any type of object, such as bags, cars, glasses, clothes, humans, etc. In contrast, facial recognition technology focuses on detecting and analysing human faces. Facial recognition is the most understandable concept in biometric matching because people use it naturally to identify each other daily and without the need for computers. Moreover, facial recognition technology doesn't require special sensors. A facial image could be captured from simple types of sensors such as a webcam rather than fingerprint and iris recognition technologies that require specific and dedicated sensors such as fingerprint and iris scanners. Furthermore, it is easy to obtain facial images from various national, regional, and international data sources available for law enforcement agencies. The availability of facial images from the internet and open source increased after the massive use of social media without good protection of the privacy of personal information.

Understanding the mechanisms, accuracy, and demographics of image and facial recognition is important for recognising their differences. It is also important to

provide high-quality training for law enforcement officers to qualify them for using those AI algorithms to reveal the identities of suspects and victims. The below table shows a comparison between image recognition and facial recognition.

Table 2. Comparison between Image Recognition and Facial Recognition

Comparison	Image Recognition	Facial Recognition
Mechanism	Analyze full image	Analyze Faces
Limitations	Image-Related	Facial-Related
Accuracy	Low	High
Image Popularity	Important	Not Important
Background & Colors	Important	Not Important
Ethnicity Bias	No	Yes

The image recognition algorithms analyse the full image to classify the type of the image or object and search for similar images, while the first step of a facial recognition algorithm is to detect the existence of a human face inside the image and search for similar faces. The limitations of image recognition tools are related to the whole image of the submitted photo or the photo in the reference database. Nevertheless, the limitations of the facial recognition tools are related to the detected faces only. The accuracy of the image recognition algorithms is lower than the facial recognition algorithms when searching for human faces. The popularity of the equivalent images on the web is important when using image recognition to search for similar images. In contrast, the popularity of the equivalent images is not important when using a facial recognition tool because it searches for similar facial images, even if they are in different photos. Similarly, the backgrounds and colours are important to find equivalent images when using image recognition, while backgrounds and colours don't affect the facial recognition results. Finally, the AI algorithms for image recognition are not affected by ethnicity bias. In contrast, the AI algorithms for facial recognition are prone to ethnicity bias, especially if they were trained with a non-representative dataset.

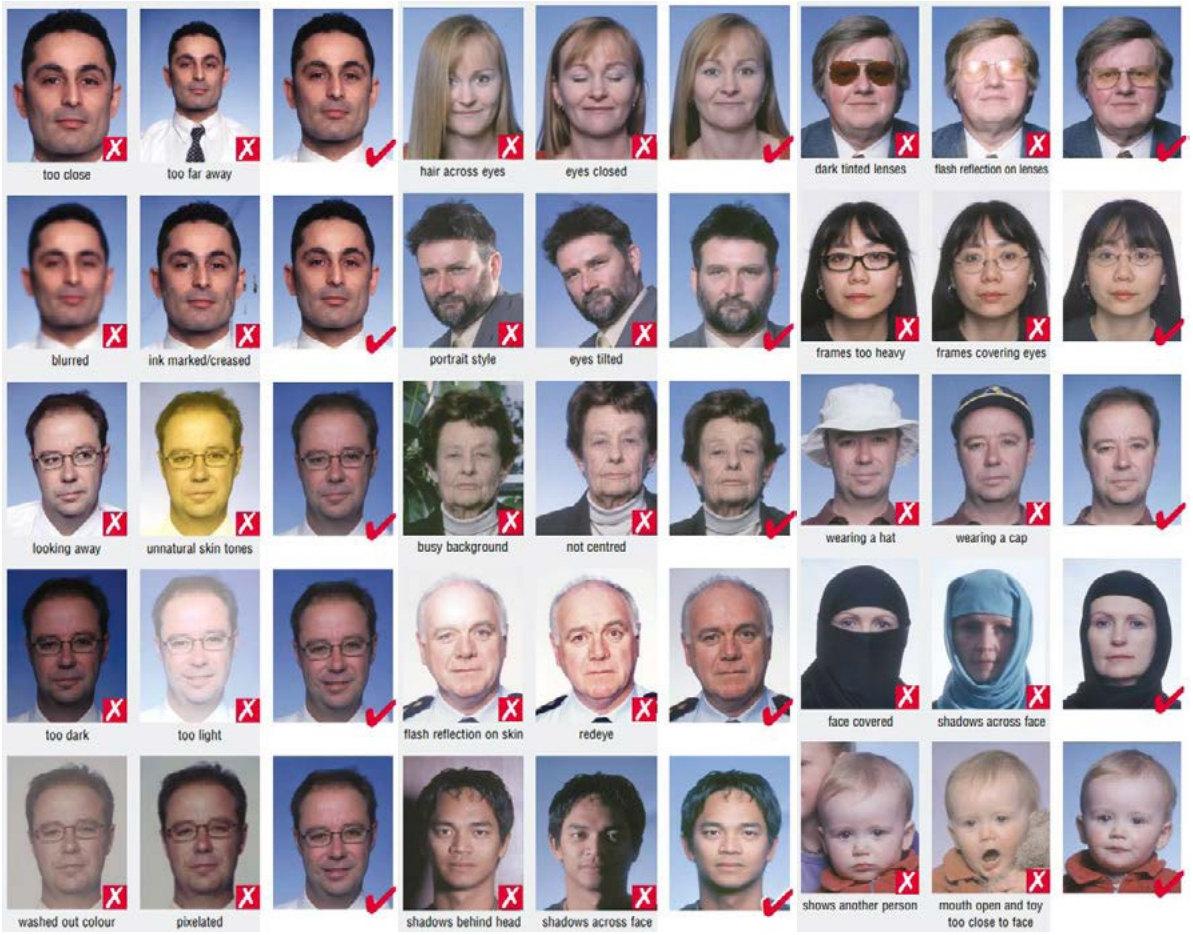
Image Recognition

Artificial Intelligence algorithms for image recognition are used to search for generic and different types of objects. Many image recognition AI algorithms and tools are available publicly and for free, such as Google, Bing, and Yandex. Users can submit an image to search for exactly similar images on the public internet. The detectives and investigators can use such AI algorithms to search and find persons of interest. However, the detectives and investigators should understand the mechanisms, limitations, and factors mentioned in the above table. They should receive high-quality training programs to achieve good results for deciding on multiple and fraudulent identities.

Facial Recognition

Over the last five years, AI has caused a leap in facial recognition technology and is the technology's reason for increasingly accurate results. Nevertheless, new challenges have occurred due to using non-representative data to train the AI algorithms, leading to wrong matches or mismatches. AI algorithms for facial recognition should be trained with representative data that is agnostic of nationality, skin tone, and ethnicity to achieve the target of linking similar identities across the different lists of candidates. The followed approach within the paper is to consider the pre-trained AI algorithms so that the results could be biased. Finally, the AI technology for facial recognition still has technology limitations related to the quality of the submitted images, the stored reference images, and the matching mechanisms. The users should understand the ICAO guidelines for high-quality passport photos (Poon, 2008) and the limitations and effects on matching results related to the distance between eyes, resolution, pose angles, facial expressions, natural skin tone, light exposure, brightness, contrast, and backgrounds. The below images clarify the ICAO guidelines.

Figure 3. ICAO guidelines for high-quality passport photos



The users of the facial recognition AI algorithms should be aware of the facial recognition techniques and the technology limitations and should be trained to achieve the best results from the pre-trained AI algorithms. Furthermore, the users should understand the accuracy levels of the algorithms, the bias of training data, AI mechanisms and demographics, and decide on the correct algorithms that fit the submitted images. The below table shows a comparison between the results of evaluating four commercial AI algorithms for Facial OSINT.

Table 3. Comparing Four Commercial AI Algorithms for Facial OSINT

Facial OSINT	American	Chinese	Polish	Eastern Country
Geographic Area	Americas	China	Europe	Eastern Europe
Identify Sunglasses	Yes	No	No	No
Identify Children	Yes	No	No	No
Ethnicity Bias	White, African, Hispanic	Asian	European	European
Websites Coverage	Criminal Records	Asia	Wide	No
Social Media	Facebook, Instagram, YouTube	No	No	VK, Tik Tok, Clubhouse

Facial recognition was limited to a closed set of internal databases for a few law enforcement agencies, but advanced AI tools were recently developed for Facial OSINT. Facial OSINT means submitting a facial image to search the public internet and revealing the identity of the target person through the data available from open sources such as web pages, blogs, and social media profiles. The table above compares four AI algorithms for facial OSINT from the USA, China, Poland, and an Eastern country. To obtain high-quality results, the detectives and investigators should understand each algorithm's demographics and geographic area. For example, each algorithm has better coverage of the area where it was developed, so the Chinese algorithm will not return any results if the investigator used the Chinese algorithm for querying a facial image of a person living in the US and vice versa for the American algorithm. Also, the investigator needs to understand which algorithm returns the best results if the person in the facial image is wearing dark sunglasses. Only the American algorithm returns good results for people wearing sunglasses, while the other three will either return irrelevant results or no results. For the sensitive cases of child abuse and trafficking in children, it is highly important to find an algorithm that can identify children across the web with high accuracy, and only the American algorithm can do that. Ethnicity bias is an important factor for getting good results for combatting terrorism and serious crime, and, unfortunately, all four algorithms have ethnicity bias. For example, the Chinese algorithm will provide inaccurate results if the ethnicity of the submitted facial image is White, African, or Hispanic. The Polish algorithm has the widest website coverage. In contrast, the American covers websites with criminal records only, the Chinese match results from websites hosted in Asia, and the Eastern algorithm doesn't cover any website except specific social media. Finally, and regarding social media coverage, the American covers Facebook, Instagram, YouTube, and Couchsurfing, the Chinese and Polish don't cover any social media, and the Eastern covers VK, Tik Tok, and Clubhouse.

European UMF Standard (P-O-L-I-C-E)

Clause 51 of regulations (EU) 2019/817 and 2019/818 for interoperability clearly states that: [The implementation of the UMF standard may be considered in VIS, SIS and in any other existing or new cross-border information exchange models and information systems in the area of Justice and Home Affairs developed by Member States.]. The UMF standard is well structured and was developed for exchanging information between law enforcement agencies. Complying with that standard format will help solve the challenges of multiple-identities, fraudulent identities, and cross-border investigations. The below image depicts the P-O-L-I-C-E "Person-Organisation-Location-Item-Connection-Event" format of the UMF structure.

UMF and POLE Pyramid

UMF (Council Regulation proposal (EC) 2018) is key to achieving EU Interoperability and solving the multiple-identity and fraud issues, especially for information concerning crimes and persons of interest; twelve countries have already introduced UMF for use by law enforcement authorities in Europe and beyond. At the same time, Police forces have long used Persons, Objects, Locations, and Events (POLE) to classify crimes. Similarly, UMF uses Person, Item (Object), Location, Event (Offence), and a fifth attribute: Biometric Data.

For example, consider a murder incident where an unknown person was the victim of a shooting. Witnesses later described the suspect as a middle-aged white male with blue eyes and red hair, wearing glasses, a red shirt, and blue trousers. Using POLE, the description is *Person*: victim; murderer (40-50, male, caucasian, blue eyes, red hair, glasses). *Object* (Item): gun; red shirt; blue trousers. *Location*: stadium. *Event* (Offence): murder.

With an eye to the future, UMF can represent the data obtained from surveillance systems. So in the above example, face recognition systems will find facial metadata such as age, gender, glasses, and other physical characteristics. Likewise, video analytics can add more metadata, and it can automatically identify items such as a shirt, trousers, and colours. The POLE data model makes it possible to search and correlate this metadata.

The below figures depict the structure of the UMF standard and the equivalent POLE Pyramid that represents a Person-Centric approach to achieving interoperability.

Figure 4. UMF Structure

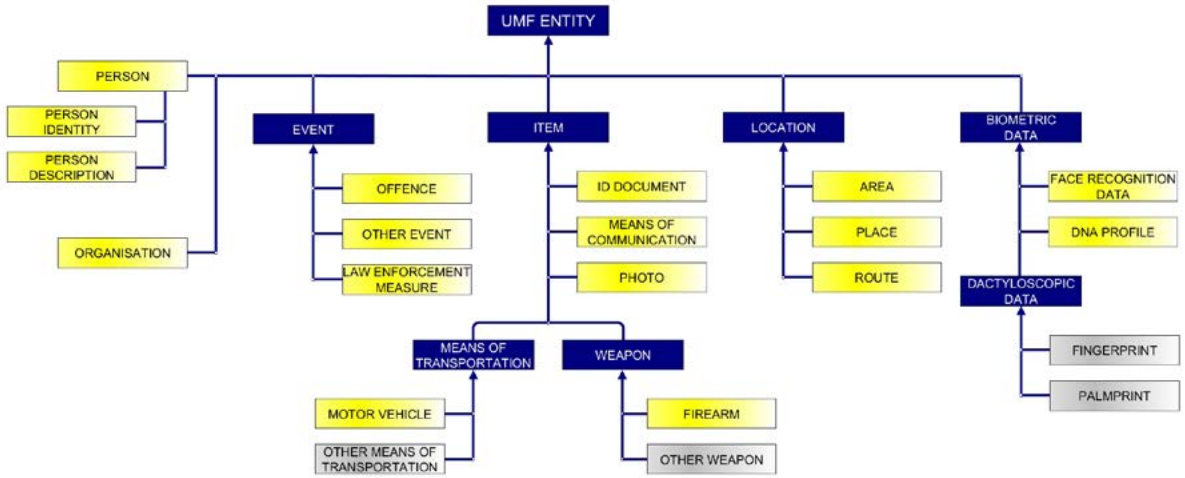
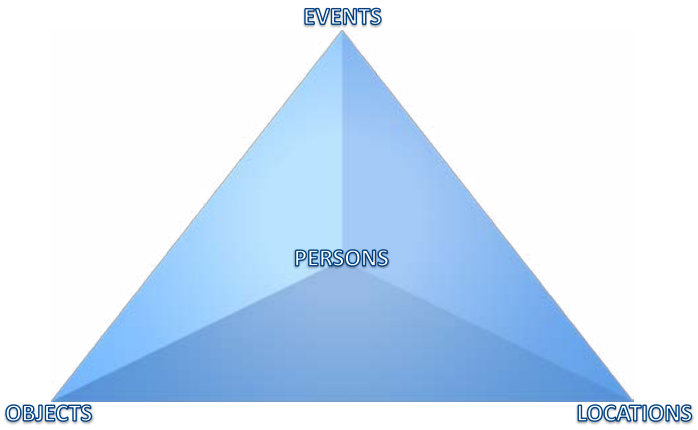
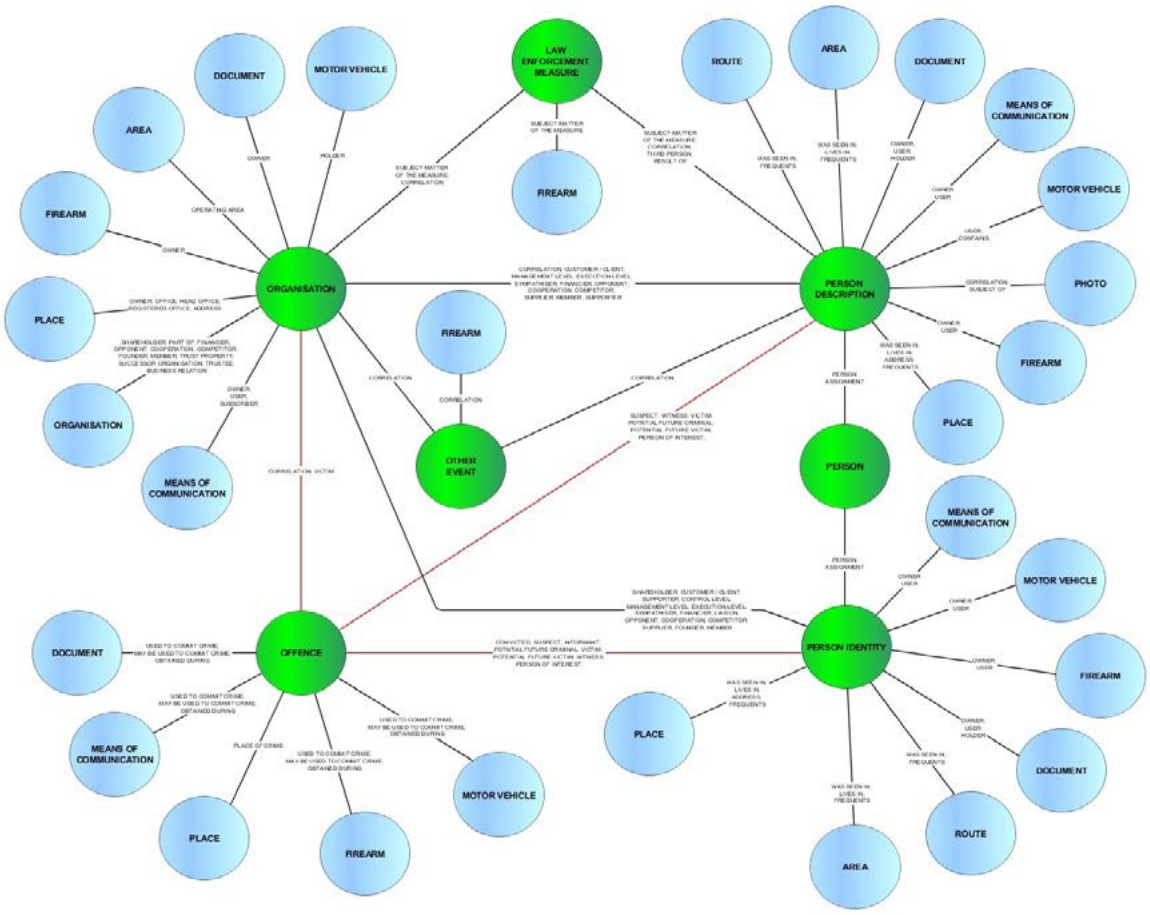


Figure 5. POLE Pyramid



The UMF standard is a Person-Centric representation for criminal investigations. Person-Centric means that all the elements, classes, subclasses, properties, objects, and instances are centred on a person, and revealing a person's identity, such as a suspect, target, or victim, is the optimum target of investigators. The Person-Centric structure is well-understood by law enforcement and border security officers. It can be used for searching or querying databases or exchanging criminal-related information among competent authorities. That structure helps the operational and field officers focus on the functional aspects they understand by heart while avoiding being involved and overwhelmed with learning about the complexity of the technical aspects. For example, biometric-based structures, such as the NIST format for fingerprints, facial images, and iris, focus on data representation and modelling technicalities. They don't highlight the full characteristics of a person. Furthermore, fingerprint and facial NIST formats are only interoperable on the biometric level but not on the higher identity levels. The below image depicts the Person-Centric approach of the UMF standard, where all the items and elements,

Figure 6. UMF Items



such as "Person Description", "Person Identity", Offence, Event, Organisation, Document, Motor Vehicle, Means of Communication, Firearm, Route, Area, and Place, are connected to a "Person".

HORUS Method - UMF Bidirectional Data Mapping

Data mapping and interoperability between the newly established and legacy systems will be required to correctly identify the different encounters of the same passenger across the different watchlists and information systems. The UMF standard can be used for bidirectional API and PNR data mapping with the EES, ETIAS, SIS, and VIS. The UMF "Person Identity" contains the "Person Core Name" to map the passenger's given name, family name, and other names. The value (Yes or No) of the "Primary ID" determines whether the

identity data belongs to the main passenger or the emergency contact. The address structure of the UMF contains "Location" and "Place" to map the passenger's contact address, billing address, mailing address, home address, and intended address. The UMF will map the email address and telephone details to the "Means of Communication" (MoC) item and specify the MoC type and identifier. The "ID Document" item of the UMF will map the travel document information. The UMF "Means of Transportation" (MoT) contains the License Plate Number, VIN, Make, Model, Vehicle Type, and Color to map the vehicle information. Finally, the UMF will map the fingerprints to the "Dactyloscopic Data" and the facial image to the "Face Recognition Data" of the "Biometric Data" item.

The table below depicts using the UMF to map the biographic and biometric data of the passengers with the central EU information systems.

Table 4. UMF Mapping of Air and Sea Passenger Information

Data Group	Data Element	UMF Mapping
Passenger Name Details	Passenger Name	Person Identity → Person Core Name
	Family Name	Person Identity → Person Core Name
	Given Name/Initial	Person Identity → Person Core Name
	Title	
	Other Names	Person Identity → Other Name
Address Details	Emergency Contact	Person Identity → Person Core Name Primary ID = No
	Contact Address	Location → Place → Address
	Billing Address	Location → Place → Address
	Mailing Address	Location → Place → Address
	Home Address	Location → Place → Address
	Intended Address	Location → Place → Address
	Email Address	Item → MoC → MoC Type Item → MoC → MoC Identifier
Contact Telephone Number	Telephone Details	Item → MoC → MoC Type Item → MoC → MoC Identifier
Travel Document Data	Name on Passport	Item → ID Document → MRZ Content
	Date of Birth	Item → ID Document → MRZ Content
	Sex	Item → ID Document → MRZ Content
	Nationality	Item → ID Document → MRZ Content
	Passport Number	Item → ID Document → Document Number
Vehicles Registration	License Plate Number	MoT → Motor Vehicle → VIN
	Car Brand	MoT → Motor Vehicle → Make
	Car Model	MoT → Motor Vehicle → Model
	Body Style	MoT → Motor Vehicle → Vehicle Type
	Color	MoT → Motor Vehicle → Colour
Biometric Data	Fingerprint	Biometric → Dactyloscopic → Fingerprint

Person-Centric OSINT

AI and UMF for enhanced interoperability will be the bridge between Cybersecurity and Biometric Technology. To clarify, linking similar identities from OSINT and the EU information systems can be achieved using a hybrid solution of Knowledge-Based Domain-Specific AI for UMF, NLP and NER for advanced matching identities and AI for facial recognition. All can be done within the legal framework and by considering the regulations for protecting personal data like the GDPR.

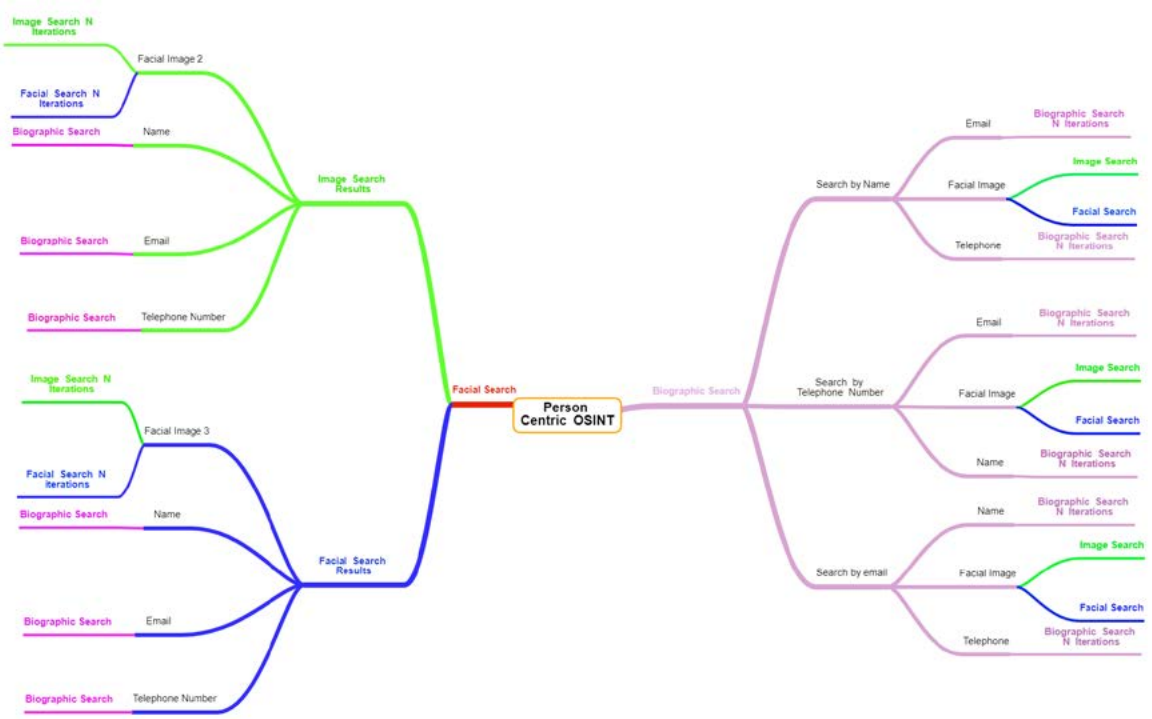
Person-Centric Approach

Person-Centric OSINT constructs the lost bridge between OSINT and biometrics, especially facial recognition. The Person-Centric OSINT approach uses open-source data to investigate cases and assemble their identity footprints to reveal their identities on the internet. The searches will be limited to a biometric search

using a facial image and a biographic search using first name & family name, email address, or telephone number. The cases can be categorised into three groups; the first group of cases is fully identified where the facial image and the identity-related biographic data are known to the investigator, the second group of cases is partially identified where the identity-related biographic data is known. Finally, the facial image is unknown to the investigator, and only the facial image is known for the last group of cases.

Searches will always have a single starting point in the Person-Centric approach, either to start with a facial search or a biographic search. The elements of the results will be submitted for successive iterations of searches until the identity is revealed or more information is gained. For example, the OSINT search could start by submitting a facial image for search using AI tools for image recognition or facial recognition. The result could be a name submitted for the second iteration of the biographic search to reveal an email. The email can be submitted for the third iteration of a biographic search to reveal a telephone number and so on. Another example, the Person-Centric iterations could start with a biographic search using the first name and family name. The result could be a facial image that could be used for the second iteration of a facial search or an email that could be used for the second iteration of a biographic search. The third iteration could fluctuate between a facial or biographic search, based on the obtained results, and so on. The below image depicts the mind map for the recommended iterated Person-Centric OSINT searches.

Figure 7. Mind map for the iterated Person-Centric OSINT searches



Rule-Based Decision Making

The final decision on confirming or rejecting the link between two identities is a human-based decision for that paper. The results are evaluated through a Multi-Attribute Rule-Based decision-making approach (Bohanec, M. and Rajkovic, V., 1999). The investigator can use that approach to identify and decide the similarities and differences between identities. The scale and weight of the rules are subject to change based on continuous studying and evaluating results. The below table depicts a weighted evaluation using a Rule-Based decision-making method for comparing the results of two identities.

Table 5. Multi-Attribute Rule-Based decision-making

Facial Match	Email	Telephone Number	Name	Decision
Different Persons	Different Emails	Different Numbers	Different Name	Different Persons
Different Persons	Different Emails	Different Numbers	Nickname or Variant	Different Persons
Different Persons	Different Emails	Different Numbers	Exact Name	Different Persons
Different Persons	Different Emails	Same Number	Different Name	Investigate more
Different Persons	Different Emails	Same Number	Nickname or Variant	Investigate more
Different Persons	Different Emails	Same Number	Exact Name	Investigate more
Different Persons	Same Email	Different Numbers	Different Name	Investigate more
Different Persons	Same Email	Different Numbers	Nickname or Variant	Investigate more
Different Persons	Same Email	Different Numbers	Exact Name	Investigate more
Different Persons	Same Email	Same Number	Different Name	Same Person
Different Persons	Same Email	Same Number	Nickname or Variant	Same Person
Different Persons	Same Email	Same Number	Exact Name	Same Person
Same Person	Different Emails	Different Numbers	Different Name	Investigate more
Same Person	Different Emails	Different Numbers	Nickname or Variant	Same Person
Same Person	Different Emails	Different Numbers	Exact Name	Same Person
Same Person	Different Emails	Same Number	Different Name	Same Person
Same Person	Different Emails	Same Number	Nickname or Variant	Same Person
Same Person	Different Emails	Same Number	Exact Name	Same Person
Same Person	Same Email	Different Numbers	Different Name	Same Person
Same Person	Same Email	Different Numbers	Nickname or Variant	Same Person
Same Person	Same Email	Different Numbers	Exact Name	Same Person
Same Person	Same Email	Same Number	Different Name	Same Person
Same Person	Same Email	Same Number	Nickname or Variant	Same Person
Same Person	Same Email	Same Number	Exact Name	Same Person

- Red is a low possibility.
- Black is a medium possibility.
- Green is a high possibility.

Conclusion

The major investigation challenges are summarised as multiple-identity, fraudulent actions, lack of interoperability and absence of an effective technical solution for exchanging Cross-Border information, and complexity of OSINT investigations.

The recent global threats such as the increase of illegal immigration, the high risks of terrorism and serious crime, the COVID-19 pandemic, and the war between Russia and Ukraine created the essential need for exchanging Cross-Border information for preventing, detecting, and investigating terrorism and serious crime across Europe and the neighbouring countries.

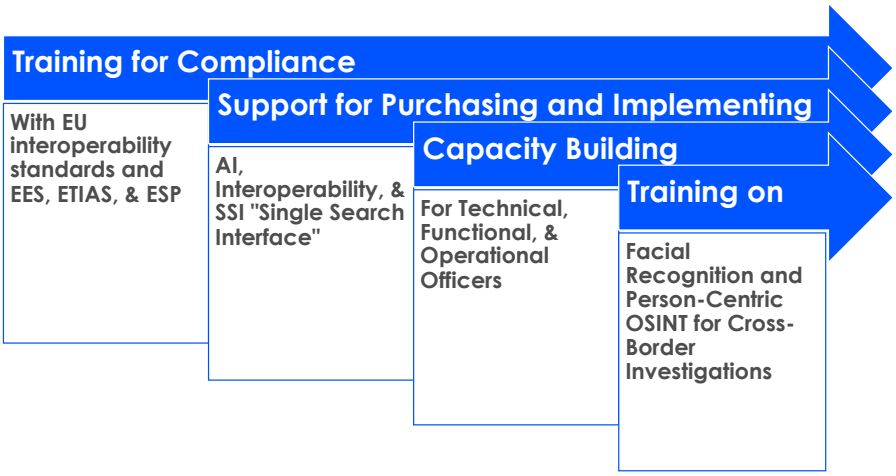
Providing high-quality training for law enforcement officers is an essential step for solving the investigation challenges. Importantly, *the training programs should contain Artificial Intelligence mechanisms, limitations, and demographics, and it is recommended to cover the proposed Person-Centric OSINT approach.*

Moreover, the training programs for each EU and non-EU Member State are recommended to include the following: Training for compliance with the EU interoperability regulations and standards and the new EU systems such as the EES, ETIAS, and ESP, Providing support for purchasing and implementing Artificial Intelligence, interoperability, and SSI "Single Search Interface", Capacity building for the border security and law enforcement agencies' technical, functional, and operational officers, and Training on facial recognition,

facial OSINT, and Person-Centric OSINT for cross-border investigations.

Finally, the training tools should include mock trials and criminal case simulation, and the training syllabuses should cover using modern technologies and digital skills for solving the challenges of multiple-identity, fraud, and cross-border investigation. The below image depicts the recommendations.

Figure 8. Recommendations for Member States



References

- Bohanec, M. & Rajkovic, V. (1999) Multi-attribute decision modeling: Industrial applications of DEX. Informatica (Ljubljana), 23(4), pp.487-491.
  Available at: https://kt.ijs.si/MarkoBohanec/pub/Inform99.pdf [Accessed 13 July 2022]
- Council Regulation (EC) No 2019/817 of 20 May 2019 on establishing a framework for interoperability between EU information systems in the field of borders and visa.
  Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019R0817 [Accessed 7 July 2022]
- Council Regulation (EC) No 2019/818 of 20 May 2019 on establishing a framework for interoperability between EU information systems in the field of police and judicial cooperation, asylum and migration.
  Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019R0818 [Accessed 7 July 2022]
- Council Regulation (EC) No 2019/816 of 20 May 2019 on establishing a centralised system for the identification of Member States holding conviction information on third-country nationals and stateless persons.
  Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019R0816 [Accessed 7 July 2022]
- Council Regulation (EC) No 2017/2226 of 30 November 2017 on establishing an Entry/Exit System (EES) to register entry and exit data and refusal of entry data of third-country nationals crossing the external borders of the Member States.
  Available at: https://eur-lex.europa.eu/legal-/EN/TXT/?uri=CELEX%3A32017R2226 [Accessed 8 July 2022]
- Council Regulation (EC) No 2018/1862 of 28 November 2018 on the establishment, operation and use of the Schengen Information System (SIS) in the field of police cooperation and judicial cooperation in criminal matters.
  Available at: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32018R1862 [Accessed 8 July 2022]
- Council Regulation (EC) No 603/2013 of 26 June 2013 on the establishment of ·Eurodac· for the comparison of fingerprints for the effective application of Regulation (EU) No 604/2013.
  Available at: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32013R0603 [Accessed 8 July 2022]
- Council Regulation (EC) 2018/1240 of 12 September 2018 on establishing a European Travel Information and Authorisation System (ETIAS).
  Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32018R1240 [Accessed 8 July 2022]
- Council Regulation amended proposal (EC) No 2018/480 of 13 June 2018 on establishing a framework for interoperability between EU information systems (police and judicial cooperation, asylum and migration) and amending regulations.
  Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018PC0480&from=EN [Accessed 10 July 2022]

- Council Regulation proposal (EC) No 2021/0106 on 21 April 2021 for laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts of 21 April 2022.
  Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206[Accessed 11 July 2022]
- European Commission, Directorate-General for Migration and Home Affairs (2018) *Feasibility study of a Common Identity Repository (CIR): management summary.* Publications Office.
  Available at: https://data.europa.eu/doi/10.2837/330396
- European Union Agency for the Operational Management of Large-Scale IT Systems in the Area of Freedom, Security and Justice (2018) *Shared Biometric Matching Service (sBMS): feasibility study - final report*. Publications Office.
  Available at: https://data.europa.eu/doi/10.2857/84504
- European Council: Council of the European Union (2019) *Interoperability between EU information systems: Council adopts regulations.*
  Available at: https://www.consilium.europa.eu/en/press/press-releases/2019/05/14/interoperability-between-eu-information-systems-council-adopts-regulations/ [Accessed 10 July 2022]
- Europol (2014) Universal Message Format: faster, cheaper, better. Publications Office.
- Poon, J. (2008) Annex A Photograph Guidelines.
  Available at: https://www.icao.int/Security/mrtd/Downloads/technical%20reports/annex_A-photograph_guidelines.pdf [Accessed 8 July 2022]
- Sawalha, M., Brierley, C. & Atwell, E. (2014) Automatically generated, phonemic Arabic-IPA pronunciation tiers for the Boundary Annotated Qur'an Dataset for Machine Learning (version 2.0). *In Proceedings of LRE-Rel 2: 2nd Workshop on Language Resource and Evaluation for Religious Texts, post-conference workshop (LREC 2014)*, Reykjavik, Iceland, 31 May (pp. 42-47). https://doi.org/10.13140/2.1.2887.2640
- Yusof, R.R., Zainuddin, R., Baba, M.S. & Yusoff, Z.M. (2010). Qur'anic words stemming. *Arabian Journal for Science and Engineering*, *35*(2), pp.37-49.
  Available at: https://www.researchgate.net/publication/298945997_QUR%27ANIC_WORDS_STEMMING [Accessed 12 July 2022]