Feasibility Study of a Common Identity Repository (CIR)

Management Summary

December 2017

Written by PwC

EUROPEAN COMMISSION

Directorate-General for Migration and Home Affairs

Directorate B— Migration and Mobility

Unit B.3 — Information Systems for Borders and Security

Contact: Philippe VAN TRIEL and Richard RINKENS

E-mail: HOME-SMART-BORDERS@ec.europa.eu

European Commission
B-1049 Brussels

Europe Direct is a service to help you find answers to your questions about the European Union.

Freephone number (*)**:**

00 800 6 7 8 9 10 11

(*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

Luxembourg: Publications Office of the European Union, 2017

ISBN: 978-92-79-77739-4

doi: 10.2837/330396

catalogue number: DR-04-18-014-EN-N

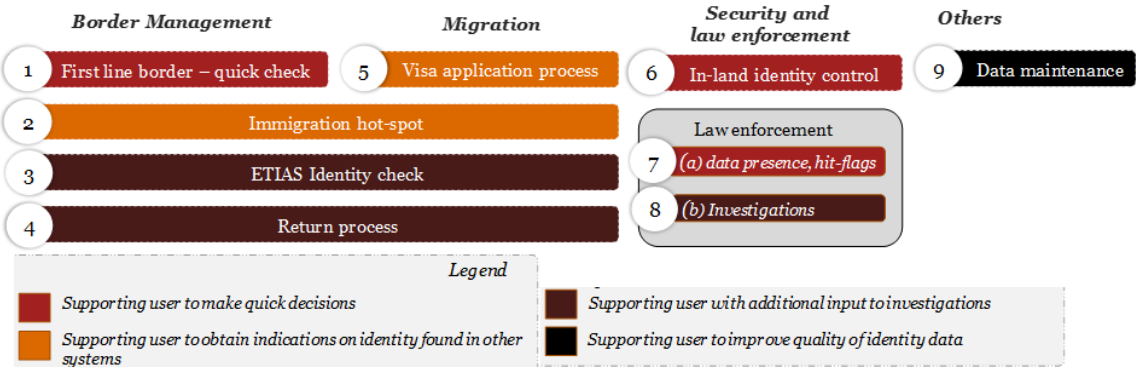
Management Summary

Accurate information is an essential requirement for ensuring the smooth identification of *bona fide* travellers to the European Union (EU) and to detect false identities, which are often a gateway for criminal activities and for irregular migration[1]. The current identity management of third country nationals (TCNs) in the Schengen Area however faces a number of shortcomings. First, it is difficult to identify TCNs with multiple identities and/or detect identity fraud due to: (i) the inability to match one individual to many European central systems (ECS) in real time; (ii) the lack of fast and seamless access to existing information as the data collected on TCNs is currently dispersed in various central systems; and (iii) the difficulty to differentiate between 'false negatives' and positives, which can lead to inadequate decisions being made (e.g. non justified refusal of entry, needless and repeated controls). Secondly, law enforcement authorities face conditions different according to the non-law enforcement information system to be accessed, which can hamper the effectiveness of their controls. Thirdly, and perhaps most importantly, there is no central identity management approach at European level.

In particular, existing border management and migration systems such as the Schengen Information System (SIS), the Visa Information System (VIS) and EURODAC have not been designed for exchanging information with any other system, and no component exists today to interconnect systems that constitute information silos. However, the next-generation of large scale IT systems such as the Entry/Exit System (EES), and possibly the European Travel Information and Authorisation System (ETIAS), is being designed and developed with interoperability in mind, and therefore may contribute to overcome the identity management shortcomings mentioned above.

The European Commission's Communication[2] *Stronger and Smarter Information Systems for Borders and Security* (COM(2016) 205) presented ideas on how information systems could be developed in the future to ensure that border guards, migration authorities, police officers and judicial authorities can have the necessary information at their disposal more quickly and easily. One solution proposed and further examined in this report is a Common Identity Repository (CIR) that could act as a single component centralising the search of identity data for third country nationals (TCN) and storing the connections (links) between all the identities for TCNs that appear in more than one of the EU central systems. In the areas of border management (e.g. first-line border checks), migration (e.g. visa application, immigration hot-spots, return process), and security and law enforcement (e.g. in-land identity control, law enforcement investigations), a CIR would also help in detecting and "correcting" multiple (potentially fraudulent) identities on the basis of biometric matches to be provided by the new shared Biometric Matching Service (sBMS). It would thus help users to make quick decisions concerning potential further checks of a given person, to obtain indications on identities stored in other central systems that they would not necessarily directly consult, or to improve the quality of identity data stored in the systems.

The main foreseen uses of the CIR are illustrated in the diagram below:



[1] Source: Frontex Annual Risk Analysis 2017

[2] http://www.eulisa.europa.eu/Newsroom/News/Documents/SB-EES/communication_on_stronger_and_smart_borders_20160406_en.pdf

Key characteristics of the CIR

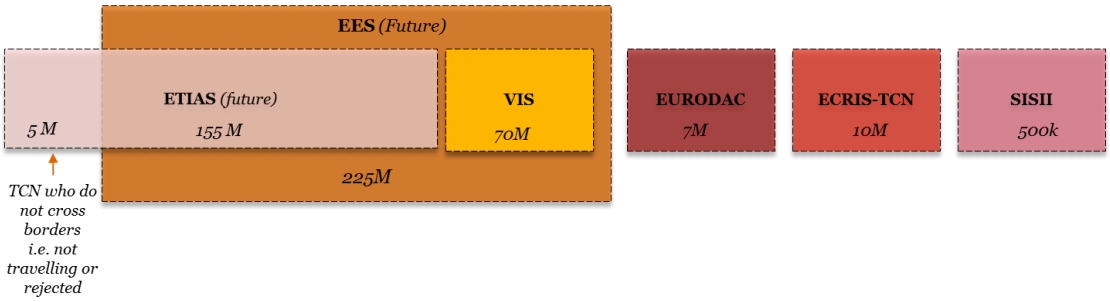
In terms of data management, a person's identity is made up of biometric and biographical information. Following the data model of the Universal Message Format (UMF) standard, the dataset considered for determining an identity and forming the common denominator between the various European central systems is: first name, family name, date of birth, place of birth[3], country of birth, and gender. Biometric data considered in this study for identifying a person includes fingerprints and facial images. Identity document number[4] is also often used as part of the identity information set or as an additional way of verifying the identity, depending on the source system and the conditions related to the document.

It is important to note that the envisaged CIR would not collect any additional data than what is or would already be available in existing or upcoming EU central information systems in the areas of border management, migration, security and law enforcement. It would also not modify the nature of user rights for accessing, updating or creating new records.

However, the CIR would create and store linkage information. It would establish a logical link between identity information recorded in two or more central systems. Such link would only be established under precise conditions, and a justification would be recorded.

The CIR would have to be equipped with a highly performing search engine to facilitate extensive searches across its repository of biographical data and to support biometric queries by making use of the new sBMS. The CIR search engine would potentially replace the need to use the respective search engines of European central systems as it would provide the same results as a series of parallel queries to each central system, with possibly some additional features such as the support for different spellings, transliteration etc. The CIR would not support the creation, update or deletion of identity data in any of the underlying central systems, as such operations would still need to be performed directly in the relevant system(s) and would then be replicated to the CIR in an automated manner.

The potential size of the CIR is estimated to be around 242 million identity records. This is based on adding up the relevant identity records of all central systems, with consideration of the overlap between the EES, the VIS and the ETIAS identity data. However, the final size would depend on the choice of architectural model.



Concerning biometric data, horizontal biometric matching across the central systems would not be handled by the CIR but by the future sBMS. The CIR would act as a client of the sBMS, and the links it would create and store would depend on the horizontal biometric matching (probabilistic match[5]) delivered by the sBMS. The sBMS would also provide the CIR with biometric references (e.g. source ID of a person's fingerprints in VIS), which could then be stored in the CIR together with the person's biographical data. Populating the CIR would be based on a one-to-many ("1:N") matching of the sBMS. A change operation (update/create/delete) at central system level would trigger a matching at sBMS level. This operation would be repeated as many times as the number of central systems. As a result, the sBMS would send its matches

[3] Note that, in respect of usability, the Machine Readable Zone (MRZ) of an identity document, which is often used as a source when making queries, does not contain the place of birth.

[4] The identity document number here is assumed to be unique and to combine the issuing country and the number displayed on the document itself.

[5] In probabilistic matching, several field values are compared between two records and each field is assigned a weight that indicates how closely the two field values match. The sum of the individual fields weights indicates the likelihood of a match between two records.

to the CIR for storing. The technical characteristics of sBMS would therefore constitute a major determining factor of the CIR's matching capabilities. This study analyses how the CIR could collaborate with the sBMS, either as a single component or as separated ones.

In terms of functionalities, the CIR would need to support search, 'create-update-read-delete' (CRUD operations (for storing data in and retrieving data from the CIR) and link detection operations. Its main users would be national-level end-users at entities mandated to access the system on behalf of the Member States, and support functions at national or European level (administrators, operators and auditors mandated to ensure the system's functioning.

Architectural options for the CIR

In order to define and assess architectural options for the CIR, different features were considered, including: (i) access to the CIR services; (ii) architectural position of the CIR in the ECS landscape (front- vs. back-end); (iii) data model; (iv) co-existence with the future European Search Portal (ESP), (v) co-existence with sBMS (separated vs. single component); and (vi) possible central systems coverage of CIR. Different potential combinations of these various features were identified, which define the various possible operating models for the CIR. These options were combined into four Target Operating Models (TOMs) considered as having the highest potential in terms of feasibility. These four TOMs were designed in accordance with the principle of data minimisation (i.e. privacy by design principle).

TOM 1, 2 and 3 represent the various ways in which the CIR could service the central systems in the scope of the study, either being at the front- or the back-end of the overall systems landscape, with full data set or links only and being either separated or closely coupled with the sBMS.

TOM 4 would split the repository functionality from the linkage functionality. For this purpose, besides the CIR itself, an additional and separate component would be introduced, called Common Identity Linker (CIL). In this TOM, the CIR would store the full set of biographical identity data from EURODAC, VIS, EES and ETIAS (and possibly ECRIS TCN). This would introduce the possibility (but not the necessity) to service the central systems with identity data from one single location, allowing for more efficient and effective ways to guarantee consistency of core identity information. The CIL, on the other hand, would act as an interface to the SIS, linking to its core identity data and providing safeguards for accessing its sensitive data. The CIL would complement the CIR by detecting and storing the relevant links between identities stored in the CIR and TCN identities stored in the SIS. This would help address some potential concerns regarding mixing alerts with identity data (different nature, source and different use cases). The CIL would leverage on the sBMS to detect the potential links that would need to be confirmed by a competent authority. The authorised users could search the CIL via the ESP or directly.

	TOM 1:	TOM 2:	TOM 3:	TOM4:
Central systems serviced by the CIR	✓ VIS ✓ SIS II ✓ EURODAC ✓ EES ✓ ETIAS ✓ ECRIS-TCN (potentially)			*CIR:* ✓ VIS ✗ SIS II ✓ EURODAC ✓ EES ✓ ETIAS ✓ ECRIS-TCN (potentially) *CIL:* CIR and SIS II
Access to the CIR services	Member States must connect directly to the CIR to access its services.	Member States can benefit from the CIR services through the central systems or through direct connection.	Member States can benefit from the CIR services through the ECS only, no direct connection is possible.	Member States can access both the CIR and the CIL directly or via the ESP.
	Architectural characteristics			
Position of the CIR	CIR as a front-end component.	CIR as both a backend component for some operations and a front end component for searches at first line	CIR as back-end component	CIR and CIL as front-end components
Data model	CIR contains: • links and decisions • core identity data[1]	CIR contains: • links and decisions • core identity data	CIR contains: • links and decisions	CIR contains only the full set of identity data from the ECS (excl. SIS II) CIL contains: • links and decisions • extract of core identity data of SIS II
Co-existence	The ESP acts as the first single search entry point to the CIR. However,	No connection between the ESP and the CIR	No connection between the ESP and the CIR	The ESP acts as the first single search entry point to both the CIR and the CIL.

[1] The option of extended dataset is discarded as there are not sufficient benefits to warrant additional duplication of data

This study assesses the four TOMs and compares them on the basis of seven criteria: interoperability, operational aspects, implementation complexity, performance, security, impact on legal basis, and financial impact/cost. Considering that there are specific security requirements regarding access to the SIS data, TOM 4 appears to be the preferable option overall. It however constitutes a slightly more complex operating model than the other ones, and hence a potentially more costly too.

Deployment and roll-out options

Multiple options exist with regards to the deployment and the rollout of the CIR. First of all, there are different possible ways to initialise the CIR:

- Launch of a CIR fully populated with relevant ECS data ahead of its entry into operations;

- Launch of a non-populated CIR, to be then progressively populated as relationships between identities are established within the existing workflows (e.g. during the issuance of visa, during investigations, etc.);
- Launch of a CIR only populated with data originating from a single central system, and subsequent progressive addition of new systems.

Regardless of the choice of initial population option, there would be a significant need for manual data processing as potentially millions of records in the central systems would be analysed and de-duplicated using biographical information and/or biometrics. Whatever TOM is selected, a dedicated specialised unit or entity would therefore need to be established to manually verify the links and associated decisions.

The CIR could in principle be deployed as an identity analysis solution or as identity repository tool. As an identity analysis solution it would perform analysis of the identities stored in underlying central systems. This could facilitate linking multiple identities together based on biographic and biometric information, by linking identifiers that reference to the actual data in the underlying systems. In this case the CIR would make it possible to link records from the various ECS together, without interfering with or having impact on the ECS themselves. The required data would only be replicated to the CIR and links would be established and managed in the CIR only.

Alternatively the CIR could be set up as an identity repository tool, extracting identity-related data from the underlying systems and managing it in a centralised way. This approach could eventually lead to removing all identity data from the central systems. This data would then be stored in the CIR and a reference to its new location in the central repository would be added in the ECS of origin. Identities appearing in multiple central systems would be automatically merged together, inherently creating links between identities in several central systems.

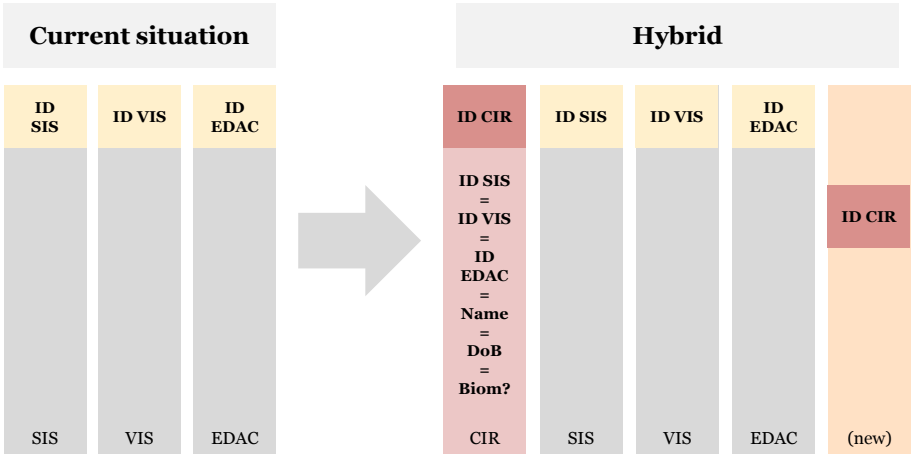
Compared to an identity analysis tool, an identity repository would allow for centralised identity management as all data related to an identity would be stored in a same system. Such central management would be much more efficient and less error prone, as centralised identity information would always be up-to-date in each system and as management processes (including governance, security, etc.) would only need to be defined once. This would however require significant changes in the already existing central systems, as all identity-related data would have to be extracted to the repository. Furthermore, the business applications would need to be re-factored to be able to both connect to the repository for obtaining their data and to work with identity information that is managed externally and can thus be updated by other central systems.

Favoured approach

This study's technical analysis, and the developed target operating models (except TOM 4), are primarily based on the assumption of a CIR conceived as an identity analysis tool. However, it clearly shows that the identity repository solution would provide significantly more benefits, such as a way to organise the management of identity-related data in a centralised way. Yet, the important changes that would be required to the already existing central systems may hamper its feasibility and influence the overall cost-benefit analysis significantly.

Therefore, a third, 'hybrid' solution could be envisaged, aiming at combining both options to obtain a centralised identity management environment in which existing systems would only require limited changes. In this hybrid combination, identity data would not be extracted from the currently existing systems but copied to the repository. A synchronisation mechanism would ensure that all changes of identity data belonging to this central system, done in either the central system itself or in the repository, would be immediately synchronised so that both instances of the data are always in sync.

This combination of identity analysis and identity repository approaches would allow central management of identity-related data (in the repository), but would not require severe changes to the central systems and their business applications.



CIR as 'hybrid' identity analysis and repository solution

In order to facilitate the roll-out of a 'hybrid' solution, the CIR could be implemented in a staged approach leveraging the ongoing development of new central systems, i.e. the CIR would only be rolled-out together with the deployment of one of the to-be-launched future systems (such as EES, ETIAS or new EURODAC). This would make it possible to already incorporate the design principles of a 'hybrid' identity management solution in the design of these new systems, resulting in a much more efficient connection process during the deployment.

A 'basic' CIR could for example be deployed together with the EES. The EES would then be equipped with a CIR identifier linking its identity-related data to the identity related data stored and managed in the CIR. No changes would be required for currently existing systems, as the CIR would in a first stage only act as an identity management system for the EES.

A second stage of the CIR roll-out could then consist in the extension of the CIR to also manage the ETIAS identities, leveraging the close connection between the EES and ETIAS. In this way, the identities of visa-exempt TCNs registered in ETIAS would be stored in the CIR and could be easily re-used by the EES when they would cross the borders. By doing this, the future EES and ETIAS systems could already reap the benefits of a central identity repository and more easily comply with what is currently envisioned in their legal proposals.

Afterwards, the CIR could be gradually extended to include the identities that are already present in the current systems as well. For example, the strong connection between EES-ETIAS and VIS could be leveraged to copy, or better migrate, the identities from VIS to the CIR as well. Depending on the feasibility, these identities could be either moved to the CIR (which would require changes to the VIS system) or simply copied to a 'hybrid' CIR (with a synchronisation mechanism in place).

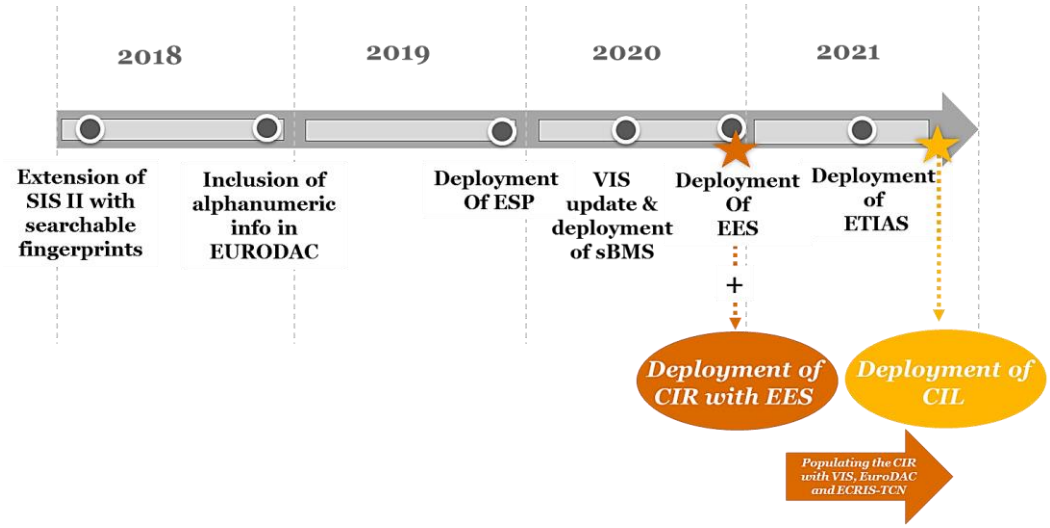
Once EES, ETIAS and VIS make use of the CIR, adding identity data from the other ECS would only add relatively small amounts of identity data, which this would nevertheless enable the CIR to also contain valuable information on identities linked to persons that could be of interest to authorities for various reasons.

All new and currently existing systems would therefore be able to benefit, in consecutive stages, from the central identity management capabilities of the CIR. The deployment of the CIR would thus constitute the first step towards an optimised and person-centric environment in which all aspects of a person's identity are fully managed centrally and separated from the core business of each European central system.

Financial aspects

Financial estimates for the development and implementation of the CIR are provided in a separate Cost Report. This Cost Report only computes expected costs for TOM 4, which the is shown by the Feasibility Study to be the potentially most beneficial operating model, but also the most complex in architectural terms. It is therefore assumed that cost estimates for the other TOMs may be derived from the cost of TOM 4 by subtracting relevant elements.

The cost model is based on the following high-level deployment timeline for the CIR/CIL as per TOM 4 (indicative only):



In TOM 4 the CIR is due to be deployed together with the EES. Hence some CIR development costs are assumed to be shared with the EES or even to be already covered by the EES budget, especially as regards hardware, software and network. Moreover, rationalisation should be aimed at by setting up common programme management for upcoming EES, sBMS, ESP and data warehouse initiatives.

The CIL on the other hand is expected to be developed and deployed as a standalone component after the CIR. Potential cost-sharing opportunities except search engine, between the CIR and CIL in terms of development are thus limited.

The cost model considers development costs both at central level and for the Member States, as well as operation and maintenance costs for a period of three years.

The overall estimated cost (in millions of €, rounded) for TOM 4 are as follows:

	Development costs	**Operations and maintenance costs (3 years)**	**Total** (‘000,000,000 €)
CIR	16,1	6,5	**22,6**
(incl. Member States)	3,5	-	
CIL	41,8	11,8	**53,6**
(incl. Member States)	30,0	-	
Total	**57,8**	**18,3**	**76,2**

HOW TO OBTAIN EU PUBLICATIONS

Free publications:

- one copy:
  via EU Bookshop (http://bookshop.europa.eu);

- more than one copy or posters/maps:
  from the European Union's representations (http://ec.europa.eu/represent_en.htm);
  from the delegations in non-EU countries
  (http://eeas.europa.eu/delegations/index_en.htm);
  by contacting the Europe Direct service (http://europa.eu/europedirect/index_en.htm)
  or calling 00 800 6 7 8 9 10 11 (freephone number from anywhere in the EU) (*).

  (*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

Priced publications:

- via EU Bookshop (http://bookshop.europa.eu).

Priced subscriptions:

- via one of the sales agents of the Publications Office of the European Union
  (http://publications.europa.eu/others/agents/index_en.htm).

Catalogue number: DR-04-18-014-EN-N

doi: 10.2837/330396