STUDY ON TECHNICAL REQUIREMENTS FOR DATA SPACES IN LAW ENFORCEMENT

Written by Michael John Flynn
June 2020

EUROPEAN COMMISSION

Directorate-General for Migration and Home Affairs
Directorate F — Financial Audit, Data Management and Risk Assessment
Unit F.2: Situational Awareness, Resilience and Data Management

Contact: Pawel Busiakiewicz

E-mail: HOME-MIH@ec.europa.eu

European Commission
B-1049 Brussels

EUROPEAN COMMISSION

STUDY ON TECHNICAL REQUIREMENTS FOR DATA SPACES IN LAW ENFORCEMENT

Directorate-General for Migration and Home Affairs

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, 2021

PDF	ISBN 978-92-76-29921-9	doi: 10.2837/289402	DR-01-21-054-EN-N

Table of contents

1. ABSTRACT

At the level of the European Commission there has been a realisation that the European Union needs to ensure its place, on the world stage, in the field of Artificial Intelligence (AI), in particular in Big Data analysis, deep learning and machine learning. Essentially, a greatly heightened ability to analyse large amounts of non-homogenous data using analytical tools.

In the sphere of law enforcement, many Member States are pursuing this goal, with a specific focus on the creation of intelligence products to support their tasks. However, although there is a great willingness to share data, good practice and products, the feedback from the Member States is that there needs to be a common (data) framework to pursue this work. Without such a framework, the risks are localised fragmentation and also datasets that could be larger and more representative of the operational data which will ultimately be analysed by the tools created.

In this study, we will explore the state of development, the issues identified, the technical opportunities and the key activities that must be coordinated to make the strategic concept become a reality.

Keywords: law enforcement, data space, data protection, Artificial Intelligence (AI), Big Data, machine learning, deep learning.

2. EXECUTIVE SUMMARY

The Commission's European Strategy for Data[1], expressed in a Communication from the Commission, provides a large part of the context for this study. The goal is for the EU to become a leading role model for a society, empowered by data, to make better decisions – in business and the public sector. This must be seen in the light of the EU's strong legal framework in terms of data protection, fundamental rights, safety and cyber-security, thereby reflecting the best of Europe, that is, open, fair, diverse, democratic, and confident[2]. The Communication recognises the need to act in the areas of connectivity, processing and storage of data, governance structures for handling data and to increase pools of quality data in a cross-border environment.

The ultimate aim is to create a European data spaces for each industrial sector, supporting creation of European data pools enabling Big Data analytics and machine learning, in a manner which is compliant with relevant legislation. This involves Commission investment in European data spaces and federated cloud structures and specifically common data spaces for public administration, including law enforcement needs. The annex to the Communication does not provide much further detail, allowing end-users in Member States and relevant European Agencies to provide creative responses, in line with the principle of proportionality and data protection rules.

In the sphere of law enforcement many Member States are pursuing projects on data spaces, machine learning and deep learning. The focus can vary from very targeted interventions, such as analysis of poor quality audio, video or fingerprint data to more strategic discussion on using advanced technology to streamline everyday bulk processes such as translation or transcription of text or to consolidate a dispersed architecture of law enforcement databases.

Through the use of standard research techniques in this study: literature review, questionnaires and targeted interviews, the following key areas were identified.

2.1. Data standards

One of the goals of Big Data analysis is to be able to receive data, in their original format, and process them without any manipulation which would affect the data quality. This is the theory. The reality is that law enforcement systems, at national level, have evolved to meet national needs, often in isolation from similar systems in neighbouring Member States which do broadly the same thing. The result is that common objects, such as vehicles or firearms, are described in different ways and data on them are not immediately compatible with data from another Member State. This becomes more complex if an advanced analytical tool on object or facial recognition needs to be tested against data from numerous sources where the objects in the data have been annotated differently in the metadata which describe the source data. A data scientist working on a huge non-homogeneous dataset wants to be able to find out what that dataset contains without carrying out an extensive conversion exercise.

Apart from large-scale systems at European level where code tables are used, the work on a Unified Messaging Format and some developments in cyber-crime, it can be seen that there is a need for shared datasets, used for developing analytical models, to have a shared labelling method, known in this context as ontology and taxonomy. This should cover not only the description of the individual records but also how the data are grouped.

[1] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Brussels, 19.02.2020. COM (2020) 66 final

[2] The importance of these values is highlighted in: European Union Agency for Fundamental Rights. Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights. FRA Focus 2019. PDF ISBN 978-92-9474-606-1

2.2. Technical solutions

Development of analytical tools which seek to avoid bias (often caused by a too small or too restricted dataset during development) should be based on datasets which are designed to effectively test what the model sets out to achieve. For example, an analytical tool for identifying people from very poor quality CCTV images should not be tested with video data taken in ideal, fully-controlled conditions. This is a problem with some commercially available datasets. Not only is this a bad research method, the operational application of the tool will be unsuccessful. Therefore, the ideal approach is to be able to assemble appropriately large datasets which contain data which are accurately described to a common format, even to the extent and success of their previous use for testing and training tools, and which can ultimately be demonstrated to be effective for the stated purpose of that tool.

This can be achieved either at the national level, which entails fragmentation and the possible duplication of effort across Europe, often on datasets which are smaller than ideal with all the inherent risk of such a situation; or it can be achieved through European added value at a central level. In this latter case, with the cooperation of the Member States' law enforcement and data protection supervisory authorities, the European Commission, European Agencies and the European Data Protection Supervisor there is the possibility to put in place a technically robust and secure environment to allow all stakeholders to move forward in a transparent fashion.

2.3. Coordination

In order to make progress, there is a critical coordination role. The European Commission is ideally placed to play that role. In short, in order for an initiative for a shared data space in law enforcement to take place there is a need to establish:

- A clear interpretation of legislation on data sharing, data retention and data quality (the question of anonymization and at which point the data cease to be useful for testing) in a scientific, research and development environment.
- On the basis of the point above, common guidelines for all end-users, which are agreed by the supervisory authorities across the European Union, on what is permissible in data sharing and retention.
- If European Agencies are to be involved, a review of their legal capabilities to ensure that their newly-identified tasks can be carried out in the long term.
- If European Agencies are to be involved, a mechanism to coordinate their budget allocations with the identified tasks to be carried out.
- The coordination of stakeholders in identifying the detailed technical solution, its supporting infrastructure and the ideal location and management structure.
- The coordination of stakeholders in drafting and agreeing the rules for use of the system such as user profiles, access rights and mechanisms, required functionalities, granting of permissions and other essential features.
- The coordination of the stakeholders in identifying and managing the projects that will be pursued once the technical solution is in place.
- The coordination of lessons learned, best practice and available datasets and even tools, so that the stakeholder group becomes a genuine learning community across the entire range of data protection and technology.
- The coordination of the evaluation of the initiatives in order to assess the timescale for possible enlargement of the activities undertaken, the technical platform and the user base (an increase in the number of end-users and/or the number of Member States using the service).

2.4. Data protection legislation

There is a widely held perception across the data scientist user community that two things are happening:

- Data protection legislation concerning law enforcement research and development activities has not been homogeneously applied or interpreted across the Member States.
- Guidance is lacking on what is permissible in the exchange and shared use of data between Member States and their retention for the purposes of testing and training analytical models in a scientific (non-operational) environment.

In the rapidly-developing realm of AI it is perhaps not surprising that legal boundaries are being tested. However, the legislation has two intentions, the correct processing of data and the free movement of those data. Accordingly, a perception that the national understanding of sharing or retention of data is not common will naturally militate against the second intention. There is a pressing need for the stakeholders to establish clarity in this area as the EU legislation appears to offer the possibilities required but this is not perceived in reality.

3. RECOMMENDATIONS

3.1. Technical

- The initial approach should be focused upon identifying a mutual problem to be addressed and constructing a specific dataset to test and train tools, built upon the best available data sample size and data type(s).
- The Commission and other stakeholders should draft and agree a list of functionalities to be introduced to the envisaged system based on the observations made by the respondents to this study. A working group should be established for this task, ideally involving the Member States who wish to be the first users of the central technical solution.
- The Commission and other stakeholders should establish a working group to ensure that any new dataset required in the central system conforms to a common ontology/taxonomy in order to facilitate searching and ensuring data standards and quality. This should include a standardised format and set of rules for the completion of metadata.
- The Commission and Member States should carry out a limited study and literature review on the extent to which anonymization may be used in constructing a dataset without compromising the usefulness of the dataset for testing analytical tools.
- The notion of low, medium and large amounts of ambition and/or investment should be viewed in a linear fashion. That is, start small in order to address the establishment of rules and procedures, followed by a number of clearly defined projects. These should be followed by evaluation of the rules, procedures, projects and outcomes before making decisions on expanding the technical provision, user base and revised rules and procedures.
- The technical solution should be a GPU[3] cluster-based private cloud created with open source technology. There would need to be a data platform and a number of simple, generic micro-services (to be decided in detail by the stakeholders). The data platform would contain separate compartments for storing the different datasets.

[3] A graphics processing unit (GPU) is an electronic circuit designed to rapidly manipulate memory to accelerate the creation of images in a frame buffer intended for output to a screen. GPUs are efficient at image processing.

- From a security point of view, there should be redundancy of the system and data fragmentation, distribution and encryption.
- The initial technical solution recommendation, on the basis of possibly six initial Member States' involvement, is for a 64-machine cluster[4]. However, by the time of pursuing this project, new technology may have been marketed and should be investigated.
- A turn-key solution, from an experienced supplier, would reduce the risks in implementation.
- The lessons learned during EUROPOL's initiative on an operational data lake should be shared with other stakeholders, especially during discussions on developing a shared data space for testing and training analytical tools.

3.2. Operational/organisational

- When a common set of rules is established on the sharing of test data it should include a format for describing the respective responsibilities of the joint data controllers.
- On the basis of existing and planned facilities, budgets, expertise and legal possibilities the Commission should propose the most suitable organisation to host the envisaged system.
- The stakeholders should decide on a suitable, secure forum and the necessary sub-sections within it for the sharing of lessons learned, national experience and shared experience.
- The discussion on ethics should be managed at the level of the European Union, so that a common view is achieved. Without this, fragmentation and suspicion are likely outcomes.
- The draft timescale, from a decision to pursue this project should be:
    - Level 1: 12 – 24 months after entry into operation.
    - Level 2: 24 months – 5 years after entry into operations.
    - Level 3: 5 years after entry into operations and onward.

  (Note. The levels above refer to the three scenarios of ambition/investment requested by the Commission in the tender document. It is obvious that not all Member States could use the central facilities provided from day one. From an organisational point of view this is not even desirable. Therefore, as highlighted in another recommendation, the three levels of ambition/investment should be viewed in a linear or time-based fashion. There should be an initial investment for a limited user-group. Following evaluation, there would be the possibility to increase the user-group – more users and/or more Member States - with a resultant need to consider upscaling the processing power of the central system. The third level, still based on evaluation, would be a central system available to all Member States, again with a re-assessment of the processing power required. Therefore, the budgetary provision should be based on increasing provision for five years and then a maintenance/update facility).

[4] Essentially, a grouping of sixty-four computers which acts as one computer for large-scale processing

3.3. Legal

- Legal clarification should be sought to ensure that the least biased dataset (probably a maximal dataset[5]) can be developed for testing and training algorithms. When this can be achieved, the later operational use of the algorithms should be better-placed to reduce unfavourable/undesirable outcomes.
- The Commission should approach the European Data Protection Supervisor and the Board in order to ascertain whether there are divergences in national law and interpretation of the Directive which might cause a difference at national level in the ability to share or retain data.
- Where data are to be used for non-operational research and development purposes, as permitted by the Directive, in addition to their original operational use, this research and development use should be reflected in the stated purpose for the use of the data so that the issues of purpose limitation are managed transparently.
- From a research and development point of view the first question should be, "What is the optimal test dataset?" This should be followed by an exploration based on, "Legally, how do we achieve that or as near as we can to that state of affairs?" If, routinely, there is a conflict between these two positions the matter should be raised, in an appropriate forum, for discussion and resolution between the data researchers and their data protection supervisory authority partners. Any lessons learned should be shared across the stakeholder group learning community.
- The Commission should engage with the Member States and the central and national data protection authorities in order to provide a clear set of rules on dataset retention in the field of development, testing and training tools in the non-operational field of research and development.
- Datasets for training, test and benchmarking purposes should be clearly separated from operational systems. A framework of mandatory disclosure to the supervisory authority of this use could enable data sharing.
- There should be a review of the data processing elements of EUROPOL's legal base in order to ascertain whether there is a need for re-interpretation or revision with regard to EUROPOL playing a full role in the development, testing and training of analytical tools using optimal datasets.

4. A PROPOSED SERIES OF NECESSARY ACTIONS

The following list seeks to place actions in chronological order. However, it is obvious that some actions can be carried out in parallel.

- D-G Home Affairs should submit its vision for a shared data space for law enforcement analytical tool development, testing and training to the wider Commission in order to secure funding for the initiative. The timescale should be one of increasing use and processing capacity in three stages over a five-year period, following entry into operations of the first level of the system. Following that five-year period, sufficient budget should be requested on an ongoing basis for maintenance and routine update.
- D-G Home Affairs should carry out an assessment of legal instruments and existing and planned budget/infrastructure/staff/services in order to identify the most effective/efficient/appropriate hosting body for such a system in the long

[5] For analytical tool training purposes, it might be necessary to interpret the principle of "data minimisation" in a different way. The data scientist is not always in a position, at the start of the process, to predict all the variables concerning the amount or variety of data required. This is especially the case when seeking to avoid bias caused by minimised datasets. Often, data maximisation is required to prevent bias or skewed models.

term. If this process identifies a shortcoming in the legal base of an otherwise ideal candidate, D-G Home Affairs should propose specific amendments.

- DG Home Affairs should establish an Expert Group of Member States (recommendation: in the first instance, approximately six Member States at an advanced stage of development in law enforcement AI matters) and European Agencies that are able to take part in the specification, testing and use of a central system for law enforcement analytical tool development, testing and training. The European Commission should provide the secretariat to this Expert Group as it is ideally placed to coordinate and initiate institutional problem-solving.
- The Expert Group should carry out the following tasks:
    - Confirm the technical architecture so that development will not be delayed and agree a timescale for delivery. The hosting body may need Member State support in activities such as testing.
    - Discuss and confirm the functionalities required in the central system, especially concerning detail such as uploading, downloading, searching, permissions and access control.
    - Discuss and confirm the requirements as regards centralised management, maintenance and security.
    - Discuss and confirm a suitable secured platform for establishing a forum for lessons learned, project updates, available datasets, available tools and other related subjects to be decided.
    - Discuss and confirm the common projects (to be kept to a small number in the first instance) to be pursued, based on a shared understanding of operational problems faced. (Note. Not all members of the Expert Group are obliged to take part in every project but should actively take part in at least one).
    - Discuss and confirm the optimal dataset for each project to be pursued.
    - Discuss and confirm the micro-services required in order to carry out those common projects.
    - Discuss and confirm the common ontology/taxonomy to be used in describing the data which are shared.
    - Discuss and confirm the common specifications for metadata which will required, to a common format, for each type of data to be shared and processed.
    - Discuss and confirm the data that will be made available for sharing in the case of each project.
    - Establish an internal project, with the support of the European Commission, to investigate the concept of data anonymization and the retention of sufficient detail for the data to remain useful. This project could be contracted out but the Expert Group must retain oversight.
    - Discuss and confirm the timescales for projects and the evaluation and reporting method to be used, not only for informing colleagues working in the same field but also for triggering decisions on expanding the capacity/capability of the central system.
- The European Commission should approach the European Data Protection Supervisor and Board to explore the most effective way of forming an ongoing partnership in order to guarantee transparency to this initiative. This could involve the selection of a nominee by the Board to sit on the Expert Group. Such a nominee could greatly assist in developing a format for setting out the limits of responsibility for joint data controllers.
- The European Commission should approach the European Data Protection Supervisor and Board to request a collaborative effort to provide a common set of guidance to the Member States, and specifically the Expert Group, on the

implementation of data protection legislation in this sphere, especially taking into account the recommendations of this study and its detailed content.

- The European Commission should approach the European Data Protection Supervisor and Board in order to ascertain whether there are divergences in national law and interpretation of the Directive which might cause a difference at national level in the ability to share or retain data. This activity should also include an assessment of EUROPOL's legal base. The European Commission should request recommendations to be made in order to overcome obstacles to the stated intention of free movement of data.
- The European Commission should approach the European Data Protection Supervisor and Board in order to jointly manage the debate on the ethics of AI-related technology in law enforcement at the EU level. This should involve the development of a public information strategy in order for this initiative to move forward with transparency and goals on informing and trust-building.

5. STUDY METHODOLOGY

5.1. Description of Tasks

This section provides a reminder of the overall list of tasks, required under the project tender documents, which must be included in the final report for this study.

5.1.1. Identification of technical requirements

This section will identify the technical requirements for the design, implementation and maintenance of a system of data spaces for law enforcement in secure environments at the European level.

The report will identify the requirements so that future setting-up of data spaces complies with the legal framework.

Member States' experience in this field will be taken into consideration. The examples of Belgium and the Netherlands were provided in the tender documents.

5.1.2. Identification of possible technical solutions

The report will identify technical challenges for the design, development, running, management, and maintenance of a system of common data lakes in secure environments at the European level. Aspects such as data protection and ownership, the use of legacy systems, the need for a common ontology at national and international levels, the different roles of the authorities accessing the data, and the maximisation of data usage, shall be analysed. The practical aspects of the design of a cloud storage solution will be addressed, for example, the potential for a cloud solution with separated sub-areas with different user-access levels in order to accommodate developing, testing, benchmarking and operations. The issues of data management over the lifespan of an Artificial Intelligence tool will be addressed as there are clear data protection restrictions on the keeping of datasets. Possible solutions to the challenges shall be proposed.

5.1.3. Recommendations

The report will be made available solely to the European Commission and will cover the technical requirements and challenges described above.

Three scenarios for the implementation of common data lakes will be proposed. These will be based on different levels of ambition and/or resources. These will range from level 1 – low level, through level 2 – medium level, to level 3 – a high level of ambition and/or resources.

Each scenario will include an analysis of the human and financial resources needed in the different phases of system life (design, development, running and management and maintenance).

The intention of this catalogue of scenarios is to provide the Commission with a variety of options for the implementation of the common data platform, depending on the political, social and economic circumstances of the moment. The recommendations should provide evidence-based and practical suggestions to the Commission on how to implement a system of common data lakes in the most efficient manner, and envisage a timeframe of maximum two years for its implementation.

It is important to emphasise that this study will concentrate on the goal of development and testing of Artificial Intelligence solutions based on two complementary scenarios – the possibilities of firstly, pooling operational data for such a goal and secondly, the creation and pooling of non-operational datasets (anonymized or "synthetic") for the development and testing of such solutions[6].

5.1.4. Context and inception report preliminary literature review

With regard to the Commission's European Strategy for Data, the most obvious outcomes in the context of law enforcement would be the development of a common European data space for security to allow research, development, testing and validation of algorithms for Artificial Intelligence-based systems to support law enforcement activities. This would involve:

- The development of the common architecture, the data standards and the criteria for certification and product quality.
- The development, collection and storage of managed anonymized, "scrambled" or realistically-created test data on which to test, train and validate algorithms. A key aspect here is the ongoing management of the data as the use of data for testing law enforcement IT systems has historically been an area of great sensitivity. This is even more important in the face of reports which highlight the risk of Big Data analysis in law enforcement simply reinforcing an existing prejudice[7] and perpetuating discrimination[8]. Equally, the risk of unlawful profiling has also been highlighted[9]. However, such views, in a rapidly-evolving field, should be seen in the light of the capabilities of AI to identify bias in data which also provides an opportunity to advance the desired goals of ethical AI.
- To achieve the bullet-point above, there is a need to improve and develop the mechanisms for creating very large amounts of realistic structured, semi-structure or unstructured test data which are either anonymized or manipulated for test purposes so as to render impossible identification of the data subject. There are issues with such an approach, not least the assessment of the point at which the removal of detail from data renders those data less useful or less realistic for testing analytical models which will ultimately be applied to operational data.

[6] Note from author: There is always the possibility to have a hybrid solution which is based on a fusion of the two scenarios stated. Indeed, across the entire field under study, this might be the preferred scenario.
[7] European Union Agency for Fundamental Rights. Big Data, algorithms and discrimination. FRA 2018. PDF ISBN 978-92-9474-241-4
[8] European Union Agency for Fundamental Rights. #BigData: Discrimination in data-supported decision making. FRA Focus 2018. PDF ISBN 978-92-9474-069-4
[9] European Union Agency for Fundamental Rights. Preventing unlawful profiling today and in the future: a guide. FRA 2018. PDF ISBN 978-92-9474-374-9

The Communication was partnered by a Commission White Paper on Artificial Intelligence[10], continuing the theme of equipping law enforcement authorities with appropriate tools to ensure the security of citizens, with proper safeguards to respect their rights and freedoms, individually and globally.

In the area of operational use of Big Data analysis in law enforcement, the available literature was historically sparse. This has been seen to pick up since 2014 with a large number of studies on ethical and legal matters, a lower number on empirical studies and conceptual matters[11].

Technical papers should be read in conjunction with the papers listed above and the Council of Europe Declaration on the manipulative capabilities of algorithmic processes[12].

Academics often concentrate on the need for effective governance of Big Data mining. This was argued by Brinkhoff[13] and must be seen as relevant in the sensitive context of cross-border sharing of data for the development of algorithms in the field of security.

5.1.5. Legal Framework

Relevant EU legislation can be found in the General Data Protection Regulation[14] and the specific data protection instrument pertaining to prosecution of criminal offences[15] and the movement of personal data[16]; the specific European legislation on legacy IT systems operating in the domain of security and migration and on the running and management of those systems. Several of these instruments are under active review.

In order to avoid excessive broadening of the scope of this report, the information held in the existing large-scale EU IT systems (Schengen Information System, Visa Information System, Eurodac) is viewed as out of scope. Where it is necessary to investigate the roles of EUROPOL and eu-LISA, this will be done in specific relation to the relevant sections of their governing legislation.

Equally, this is not an operational study. The first step is to concentrate on how to access sufficient shared, securely managed, good quality data in order to be able to have ongoing confidence in the algorithms developed and tested.

5.1.6. Research Methods

The research was focused on the area highlighted in the Commission's European Strategy for Data. With regard to the development of data spaces and the testing of algorithms, the study intends to discover:

- The current state of play, including data sources and data used.
- Aspirations at the national level.
- Limitations: legal, ethical and technical.

---

[10] European Commission White Paper on Artificial Intelligence – A European approach to excellence and trust. Brussels, 19.02.2020. COM(2020) 65 final

[11] Anneleen Rummens, Wim Hardyns, Lieven Pauwels. A scoping review of predictive analysis techniques for predicting criminal events. Institute of International Research on Criminal Policy (IRCP), Ghent University. A paper in Data Protection and Privacy under Pressure – Transatlantic tensions, EU surveillance, and Big Data (Gert Vermeulen and Eval Lievens (Eds). Maklu 2017. ISBN 978-90-466-0910-1

[12] Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes. Adopted by the Committee of Ministers on 13 February 2019. Decl(13/02/2019)1

[13] Big Data mining by the Dutch Police: Criteria for a future method of investigation. S. Brinkhoff. 10 February 2017. Springerlink.com. Eur J Secur Res (2017) 2:57–69 DOI 10.1007/s41125-017-0012-x

[14] Regulation (EU) 2016/679

[15] Directive (EU) 2016/680

[16] Regulation (EU) 2016/679

- The conditions under which Member States might cooperate on sharing operational and test data.
- The scenarios, from modest to significant, where there is potential for European added-value through targeted interventions at the Union level.

The stages for the study have been:

- Inception Report.
- Questionnaire to Member States.
- Targeted interviews with Member States (ideally NL, DE, FR) identified as having taken a lead in Big Data analysis and operational use[17].
- Targeted interviews with eu-LISA on the security, development, technical and management aspects of European data space for law enforcement purposes.
- Targeted interviews with EUROPOL on the legal, ethical and operational aspects of the use of such a space.
- Draft final report for European Commission comments.
- Final report for the European Commission.

5.1.7. Risks

The major risks in this study were:

- Maintaining focus – the risk was that the responses to the questionnaire and interviews would be so wide-ranging as to be incapable of meaningful analysis and therefore introducing difficulty in arguing for targeted interventions. All tasks were firmly set in the context of the European Strategy for Data, as other topics, however interesting, fall outside the scope of foreseen interventions.
- Timescale – the available time for this study was limited to 21 working days (total working days allocated, not consecutive days), a bare minimum for achieving an acceptable overview of the situation and proposing solutions as per the tender documents. This was mitigated by flexibility on the part of the contractor and ensuring that documents submitted for review are turned around quickly so as to be sent to respondents as early as possible to allow the maximum time for drafting replies.
- Access to experts – this was mitigated through the good offices of the Commission, making early contact with known leaders in the field and forwarding their details as quickly as possible. The level of cooperation across the Member States, European Commission and European Agencies proved to be exemplary.

[17] Belgium had been highlighted as worthy of further study and this is indeed the case. At the time of the study Belgium was in the middle of a complex procurement exercise and the COVID19 crisis, limiting the availability of staff for interview.

6. IDENTIFICATION OF TECHNICAL REQUIREMENTS

6.1. Possible technical solutions and challenges - technical recommendations – three scenarios or a linear approach?

6.1.1. Introduction

There are two ways of looking at the brief for this report, with regard to the low/medium/high levels of ambition and resources.

The first would be firmly limited on what could be done today, resources permitting; that is, a one-off investment that would be low, medium or large.

However, it is clear from the submissions of several respondents to this study that a more linear approach should be taken. It is preferable to start small, in order to prove the concept and better define any limitations and then to expand capacity, based on what has been learned. This is more structured and more likely to succeed. Accordingly, any approach to an outcome of this study should focus on what is needed in the short-term with a very clear view to identifying and resolving issues in the expectation of much greater investment in the medium to long-terms as more Member States use the facilities available, more sophistication is introduced and more storage and processing power are needed. Therefore, it is not recommended to think of the three scenarios in the tender document as options to what could be done right now; it is more manageable and sustainable to view the three scenarios as happening over a period of time.

The information below should be read in this light. This will allow those Member States which are ready and willing to take advantage of a limited central facility to do so in an ambitious timeframe whilst accepting that many more Member States (as can be seen from the levels of actual and planned work in this domain) will wish to join in within, perhaps, one to two years. Budget should be requested accordingly in order for the goals of the Commission's European Strategy for Data to come to fruition.

In summary, the levels 1, 2 and 3 described in the tender should be viewed as timescale stages; in all probability: up to 12-24 months after entering operations; the period 24 months – 5 years after entering operations and 5 years and onward.

Recommendation. The notion of scenarios of low, medium and large amounts of ambition and/or investment should be viewed in a linear fashion. That is, start small in order to address the establishment of rules and procedures, followed by a number of clearly defined projects. These should be followed by evaluation of the rules, procedures, projects and outcomes before making decisions on expanding the technical provision, user-base and revised rules and procedures.

6.1.2. Prompt questions

In order to give some structure to the discussions on technical solutions the following prompts were used.

1) In the first instance what functionalities would you wish to be made available at the central level? (This can include an idea of the number of end-users that should be considered in the first instance and also some sort of searchable forum on lessons learned).

2) What technical platform should these functionalities have? (This to be explained in terms of function and technical requirement).

3) Any other technical intervention that would be required in order for the technical solution to enter production (for testing, training etc. NOT operations)?

4) An estimate of the cost of procuring the required hardware/software. (The cost of ongoing running costs/management will be addressed separately).

5) For how long should this first stage be run before the stakeholders would be in a position to consider expansion?

6) Please also consider the same points at 1) - 4) above in this "more ambitious" investment stage as described at point 5.

6.1.3. High level architecture and available micro-services

From a technical point of view the solution should be a private cloud and should be open-source based. This is important, as in such a sensitive field the solution should be transparent with no proprietary "black box" components which cannot be explained. In Europe we do not have anything like Amazon and so, open source is the only way.

There would need to be a data platform and a number of simple, generic micro-services. Examples would be object recognition on images or language detection in text. As the platform develops more of these micro-services could be added. It could also be possible to add a level of pre-processing, such as a tool to anonymize data. This could assist data sharing.

Within the services available it would also be possible to incorporate safety check features. In this way, a user uploading a dataset containing anonymized facial images could run a check to ensure that there is not a clear, un-anonymized facial image left in the dataset by accident. Tools could also check the age/gender range of images against the stated metadata to check that the balance claimed is actually "true".

There would need to be a number of application programming interfaces (API's). Although some of the data would obviously be operational data the solution must be entirely separate from any operational systems.

Within the data platform there should be separate compartments in which to store the different datasets. This set-up should be based on privacy and security by design as data leaks must be viewed as not permissible. There should be redundancy of the system and data fragmentation, distribution and encryption so that, if for some reason, there is compromise of the system, the loss will be limited.

From a component point of view, the configuration can be built on standard equipment.

As the intention is to facilitate machine learning and deep learning the architecture of the solution would be based on a graphics processing unit (GPU) cluster or a GPU supercomputer. Essentially, the system would be a private cloud environment but based on a GPU cluster, naturally with central processing units (CPU).

Recommendation. The technical solution should be a GPU cluster-based private cloud created with open source technology. There would need to be a data platform and a number of simple, generic micro-services (to be decided in detail by the stakeholders). The data platform would contain separate compartments for storing the different datasets.

Recommendation. From a security point of view, there should be redundancy of the system and data fragmentation, distribution and encryption.

6.1.4. Sizing and scalability

If, for example, an initial core set of six Member States wished to make use of this solution, sufficient processing power could be provided by a 64-machine cluster. The advantage of this approach is that it is fully scalable to cover the later stages of a

medium and a high level of demand, ambition and resources. It is important that performance speed remains good to avoid frustration amongst users. The capacity of storage should be considered, as a minimum, to be in the order of terabytes with a view to it eventually being larger.

During the writing of this report, NVIDIA, a market leader, announced a new architecture, DGX A100[18]. This could be something of a game-changer as the physical size of the hardware and the power consumption are significantly reduced. The potential is to procure an initially expensive solution but to save on hardware, floor-space, power and air-conditioning whilst providing an easily scalable solution with enormous processing capacity. As the specialist media reported on this development in May 2020 and the first customers are expected to be large government services it is difficult to assess the cost; but it is already clear that the price/performance of this new architecture will set some new standards in relatively cheap supercomputer power for AI. This significant leap in technology should always be considered when drawing up the required solution.

Recommendation. The initial technical solution recommendation, on the basis of possibly six initial Member State involvement, is for a 64-machine cluster. However, by the time of pursuing this project, new technology may have been marketed and should be investigated.

6.1.5. Network

Regarding the network, as most processing would take place on the platform itself, in the first instance the network capacity is less important. This however, would need to be reviewed over time, and especially if more users come on-board and/or large datasets are uploaded containing images, video and audio. End-users will soon lose interest if such simple tasks as uploading are problematic. Network security is a prime concern.

6.1.6. Turn-key solution

Risks in setting up the platform can be reduced if an experienced external company is contracted. At the moment, NVIDIA has an effective relationship with DELL. By using such a commercial arrangement, it is possible to ask for a turn-key solution.

Existing software frameworks, such as TensorFlow or PyTorch, could be used in the environment to build models for testing. A good workload scheduler would be required; this could be proprietary, such as the NVIDIA Bright cluster manager, or an open source scheduler such as Slurm.

Recommendation. A turn-key solution, from an experienced supplier, would reduce the risks in implementation.

6.1.7. Costs

Costing is not an exact science as each agency/body that could host such a platform has its own relationship with suppliers. In terms of simple procurement of a build for the platform based on a 64- machine GPU cluster the price would be approximately 1,5m euros (but see previous update on the DGX A100 solution, which is relatively cheap, starting at 200k euros for the hardware). It is necessary to factor in other costs, such as power supply, air-conditioning, security, technical support, network requirements and to realise that once investment has been made the stream of new equipment arriving on the market every year does indicate the need for a budget just to keep pace and additional budget in the medium to long-terms as more Member States use the facilities available, more sophistication is introduced and more storage and processing power are

[18] https://venturebeat.com/2020/05/14/nvidia-unveils-monstrous-a100-ai-chip-with-54-billion-transistors-and-5-petaflops-of-performance/

needed. It is for these reasons that the architecture chosen should be scalable and the hosting agency be able to provide the infrastructure. GPU's are very power consuming. Additionally, with the massive processing activities taking place the machines in the cluster should be closely located (within a few metres of each other) to prevent disruption to calculation caused by the "slowness" of the connections between the machines (they run at the speed of light). For this reason, an integrated solution such as NVIDIA's DGX A100 is ground breaking in terms of computer power, throughput and power consumption.

6.1.8. High-level functionality

The solution would essentially be a storage and processing facility with features such as log-in, search, management of access rights, upload data, request data access, download data once permission to do so is granted and completion of metadata.

Regarding functionality, the solution must handle different levels of authorisation and access from day one of operation. Transparent, good access management is critical. The data owner should assign the access rights to the data. Then, a user with searching rights could search the metadata only in order to find what he/she is looking for. Once the correct dataset has been identified by the metadata search a "request access" function, with appropriate safeguards such as a short series of questions to describe the requester and the intended use, could cause a message to be sent to the data owner to authorise access to the dataset.

The platform is intended to assist the development, testing and (pre)training of models or analytical tools, not run in production, that is, an operational policing system. Once a model has been trained it can be used at the national level in an operational environment. This is why the platform should host datasets that focus on real operational issues. An example would be the analysis of video images that are poor quality, such as surveillance footage taken in conditions that are less than ideal (unusual angles, poor light). Other examples might be speech recognition based on a shared dialect within a criminal group or speech that has been distorted, such as during poor quality telephone communication.

6.1.9. Metadata

Metadata should describe the dataset but also describe what the data have been used for. In order to do this, the metadata must be largely structured, using for example, a JSON structure based on boxes to be completed. There is a need for version control of the datasets. If the dataset is updated or otherwise modified this should be captured and visible to the searcher of the metadata. The data owner can include information such as the age or gender balance within the dataset and information on what the dataset has been used for in analysis. Metrics on how well the dataset performed should also be available. As there might be a need to explain the use of the dataset or its performance there should also be a free-text facility in the metadata. This helps to develop trust in the dataset and use of the platform itself.

6.1.10. Challenges, including a common ontology and diverse data sources

There are various challenges to face. One is how to use the various data types, including synthetic data and also establishing the impact of anonymization on the usefulness of the dataset for its intended purpose.

Another significant challenge is the lack of a common ontology/taxonomy, on how to describe objects, for example, and how to group them. This could be addressed with a strict approach to metadata, that is the data that are attached to each data object and which describe the object. The metadata should also describe the data type; that is, are the data anonymized, pseudonymized etc.

It has been seen that the lack of common ontology between systems has even caused problems at the national level thus compounding the problems of data sharing at the European level. Solutions have been identified in the field of cyber-crime but this notion should be extended to other fields and the opportunity should be seized early on to ensure that it is a founding principle of a European law enforcement data space.

It is ironic that data spaces and Big Data analysis intend to store and analyse data in their original format in order not to lose detail but, if across several Member States it is not possible to describe a car or a pistol in the same way the programming of analytical tools becomes more complex than is necessary. There is a need to reduce complexity from the data sources, many of which are legacy systems designed only to meet national needs.

There are solutions to this issue. Firstly, new developments to systems at national level should look to the lessons learned during the Unified Message Format (UMF) projects at European level and, at least, attempt convergence with UMF data models. Secondly, right from the start, this shared data space initiative should mandate an ontology sub-group to specify the fields and rules for the descriptions held in metadata in order to facilitate searching and use of datasets.

6.1.11. A scientific, learning community

The users of the technical platform should consider themselves a scientific, learning community and therefore, it is healthy to publish the findings of work undertaken and the standards required, such as the ontology, so that other users and potential candidate users can see what is available and what is expected of them. A structured forum should be available to allow such publication, exchanges and should contain a search function. The forum could be hosted on an existing platform such as those provided by EUROPOL. Building on this, there will be a need for a common level of understanding of services and their use. This falls within EUROPOL's mandate on training and capacity building, as it would be advantageous for the users of a shared system to have standardised training on artificial intelligence and the use of the available micro-services in order to establish a European level competence.

The learning community should start small, for example, six Member States, and should carry out highly focused work, demonstrate its value and publish the findings. There should be regular evaluation of the technical solution and what it has achieved in order to attract new Member State users, demonstrate transparency and trustworthiness and thereby retain political support for further development. Interested Member States should be canvassed on how many end-users they might provide at the first stage. For a large Member State this might initially be up to fifty users.

6.1.12. Timescales

The first stage of use should be between 12 and 24 months, including thorough evaluation of use, outcomes and issues addressed. The roll-out schedule for the long term should envisage a timescale of five years or slightly longer. Accordingly, the three scenarios envisaged in the tender could be expressed as:

- Level 1: 12 – 24 months after entry into operation.
- Level 2: 24 months – 5 years after entry into operations.
- Level 3: 5 years after entry into operations and onward.

It is possible to stop additional development investment at any stage of this timescale but this simply implies limiting the number of users and services at the level already achieved. As long as the platform is in existence and use there will be the usual costs of running, maintenance and routine upgrade.

Recommendation. The draft timescale, from a decision to pursue this project should be:

- **Level 1: 12 – 24 months after entry into operation.**
- **Level 2: 24 months – 5 years after entry into operations.**
- **Level 3: 5 years after entry into operations and onward.**

6.2. Management and infrastructure costs

In assessing the way forward, it is not enough to simply consider the cost of purchasing a central system. The European Commission and other stakeholders will have to consider such elements as a suitable, secure network; connection to that network; power supply for the central system and its related air conditioning; 24/7 security of the hosting site and all the services related to maintenance, upgrade and intervention for incidents. The most experienced European- level body in this field is eu-LISA. Accordingly, eu-LISA specialists where invited to study the outline technical solution and the timescale in the above recommendation and forward their assessments and observations to complete the technical aspect of this report.

The response from eu-LISA can be found at Appendix 2, entitled, "Preliminary assessment of eu-LISA on the European Security Data Space (technical considerations, synergies and proposed way forward)".

7. WHERE TO HOST A SYSTEM AND AT WHICH STAGES OF ITS LIFECYCLE

7.1. eu-LISA

In the sphere of large-scale IT systems in the area of justice, freedom and security, the EU has an existing Agency, short name eu-LISA, for the operational management of such systems. The legal basis for the activities of the Agency is set out in Regulation (EU) 2018/1726[19].

One of the major reasons behind the establishment of the Agency was that it is not an end-user of any of the operational data processed in the systems which it manages, thereby effectively introducing a significant element in the security of sensitive personal data.

Under Article 1.7.7, inter alia, the Agency is responsible for carrying out research activities, carrying out pilot projects, proofs of concept and testing activities and providing support to Member States and the Commission.

Article 15 describes the conditions for setting up and running pilot projects, proofs of concept and testing activities. These include notifying the European Parliament and the Council and a positive decision of the Management Board. It is important to note that the pilots are for the development or the operational management of large-scale IT systems and so the goal must always be an operational system. The duration of pilot project budgets can be no more than two consecutive years.

[19] REGULATION (EU) 2018/1726 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 14 November 2018 on the European Union Agency for the Operational Management of Large-Scale IT Systems in the Area of Freedom, Security and Justice (eu-LISA), and amending Regulation (EC) No 1987/2006 and Council Decision 2007/533/JHA and repealing Regulation (EU) No 1077/2011

It is clear that the system under discussion in this report will be required for as long as the technology is in use, that is for the foreseeable future.

Additionally, the Agency may plan and implement testing activities on the development, establishment, operation and use of the systems.

At the request of a group of at least five Member States the Agency can be entrusted with the task of developing, managing or hosting a common IT component to assist them in implementing technical aspects of obligations deriving from Union law on decentralised systems in the area of freedom, security and justice. This can occur only after prior approval by the Commission and a positive decision of the Management Board.

7.2. The Joint Research Centre (JRC)

The European Commission has a scientific research body, with the status of Directorate-General, called the JRC, located in Ispra, Italy. The JRC has already demonstrated its capacity to provide authoritative reports and research on law enforcement related topics such as biometrics. The JRC also has a long-standing relationship with EUROPOL in which JRC seeks to provide technical solutions to operational issues, based on such technology as machine learning. The JRC possesses the data processing facilities, especially in terms of infrastructure such as power supplies and air conditioning, for large servers to be housed.

As is the case with eu-LISA, the JRC has no access to the data processed in these problem-solving projects and has a comprehensive data protection regime in place.

As can be seen from the legal base of eu-LISA, the focus is on operational systems. Accordingly, there is a limit on the length of time for which pilot or developmental projects can be funded and managed. The JRC draws together users and researchers to act as a scientific research body or innovation hub which can work in partnership with other Directorates-General of the Commission, other European Agencies such as eu-LISA, EUROPOL and Frontex and Member States to ensure that maximum advantage can be gained from technical facilities and making maximum use of legal opportunities.

The goal must be, essentially, the right system, under the right data protection regime, in the right place at the right time in its life-cycle. For example, it is not necessary for the JRC to store all the data for analytical purposes. The analytical tools can be hosted and the Member States and Agencies, as data controllers, can use tools hosted at the JRC. The same ultimately applies for operational systems hosted in eu-LISA.

An important element of such work is to demonstrate the purpose limitation of the dataset being used. The Directive provides the opportunity for the definition of the purpose to be quite broad. This should be used to advantage in a tightly-secured environment.

Recommendation. On the basis of existing and planned facilities, budgets, expertise and legal possibilities the Commission should propose the most suitable organisation to host the envisaged system.

8. EXPERIENCE OF THE MEMBER STATES AND EUROPEAN AGENCIES

8.1. The summary of the questionnaire responses by the Member States

The master overview of the questionnaire can be found at Annex 1, based on the original questionnaire format. This contains the complete text of all responses from Member States. Fourteen Member States responded.

This section will address each question in turn and draw out the key themes.

> *1. Do you have or plan a strategy to develop, train, test, benchmark or validate technological solutions (e.g. tools or services for digital investigations, Big Data analysis, image recognition etc.) under the control/supervision of Law Enforcement Authorities (LEA's)?*

Fourteen Member States answered, "yes" to this question. One Member State answered, "no".

> *1a. Are the plans or strategies based on mechanisms that enable the use of **operational datasets***?*

Fourteen Member States answered, "yes" to this question.

> *1b. Are the plans or strategies based on **datasets that are specifically designed for that purpose** e.g. synthetic data, anonymised data, data retrieved with the consent of the people involved, etc.?*

Ten Member States answered, "yes" to this question. Four Member States answered, "no".

> *2. If you have plans or strategies (as explained in question 1.), could you specify for each category of data:*
>
> *a) The type of data? (e.g. LEA database, video, audio, unstructured text, also please see guidance provided).*

In order to assist with structured responses to this question the respondents were asked if they used data which fulfilled the following descriptions - (A) operational and identifiable; (B) operational and anonymized; (C) operational and "scrambled" so that personal detail is swapped between several identities; (D) synthetic data (realistic data, similar to operational but based on the use of actors or persons who have given their consent); (E) pure test data, machine-generated for the purpose.

Most respondents stated that they use test data in the following two categories: operational and identifiable; operational and anonymized. Several Member States stated that they were also using synthetic data, especially to test in the areas of unstructured text, video, audio. One Member State reported using scrambled operational data in video or image data training sets.

One Member State reported using machine-generated pure test data for replicating information from LEA databases.

Recommendation. From a scientific point of view the first question should be, *"What is the optimal test dataset?"* This should be followed by an exploration based on, *"Legally, how do we achieve that or as near as we can to that state of affairs?"* If, routinely, there is a conflict between these two positions the matter should be raised, in an appropriate forum, for discussion and resolution between the data researchers and their data protection supervisory authority

partners. Any lessons learned should be shared across the stakeholder group learning community.

b) The source of the data?

Several Member States reported the use of data extracted from data generated by investigations (based on individuals and events), statistically processed LEA data and open source data. This can include confiscated, intercepted and covertly-obtained data. Biometric capture stations and document application processes also provide fingerprint and photographic data. Unstructured text can originate from repetitive requests to LEA support centres.

c) In order to develop plans or strategies what kind of support would you expect from the Commission?

The responses fell under the following broad headings:

FINANCIAL/TECHNICAL PROVISION - This includes the provision of centralized storage facilities for test data and the facility to assemble the individual test datasets themselves. Commission-funded or jointly-funded studies on problem-focused solutions would form part of this. The Commission has a central role in influencing the conditions for the infrastructure requirements to host components to perform such tasks.

This heading also includes the provision to develop training datasets, specific for LEA use, that contain audio, image and video and which realistically resemble the data that are processed daily. In this way it is possible to train models much better or have a much better transfer learning effect on pre-trained models.

Anonymization and pseudonymization are not just legal concepts. If there is best practice in constructing test datasets involving these techniques it would help not only data protection but the construction and integrity of the datasets themselves.

LEGAL CLARITY AND GUIDANCE – This would include clarification of data protection legislation regarding limitations on cooperation between LEA's and the private sector and clarification on anonymization and pseudonymization.

Many Member States requested clarity on purpose limitation and storage/retention limitation as the understanding and implementation does not seem to be harmonised across the Member States. Accordingly, Member States report that much time is wasted on legal discussion when the rules should be clear across the Union.

Clarity on rules would include establishing the clear separation of the operational and scientific/research/testing domains as the legislation accommodates this distinction and the transfer of data from the former to the latter.

The Commission would have the pivotal role in providing a legal interpretation, supported by the EDPS for using operational data for developing/testing/benchmarking amongst the Member States, and resulting in the permission for EUROPOL to also do so.

Common guidance could also be drafted on dissemination of results and trained models.

PROMOTION OF COMMON STANDARDS - This includes the notion of common standards/formats and "interoperability" of data in order to be able to carry out analysis across the widest possible range of existing databases and also semi-structured and unstructured data.

The notion of common standards also extends to the unification of the platforms used so as to ensure that data integration and analysis are the least complex achievable.

Regarding common systems, it would be more helpful to see clear rules on mandatory features and functionalities and agreement on data types/formats. Indeed, it is probably only through advanced analysis that such wide varieties of data can be processed in any case.

SHARING OF LESSONS LEARNED AND INFORMATION - This includes regular updates on activities across the Member States and within Agencies to allow a true community to be built where expertise is shared and lessons learned on both success and failure are readily available to the community. This would also allow reporting on solutions developed at Member State level to permit sharing of best practice.

Recommendation. The stakeholders should decide on a suitable, secure forum and the necessary sub-sections within it for the sharing of lessons learned, national experience and shared experience.

3. Are you building or planning to build a "shared" pool of data for the development of technological solutions that could be accessible for multiple projects or multiple police services (as opposed to datasets that are created for a particular purpose/project but are not designed to be reusable)?

Eight Member States responded, "yes" to this question, whilst seven responded, "no".

Recommendation. The approach should be focused upon identifying a problem to be addressed and constructing a specific dataset to test and train tools, built upon the best available data sample size and data type(s).

4. Once you have built a tool for analysis are you permitted to retain the original test dataset so that you can use it again for testing if you further refine the tool/algorithm?

Eleven Member States responded, "yes" to this question, whilst five responded, "no". This included one Member State which answered that it depended on the data source and so responded both positively and negatively. This was reinforced by another Member State which pointed out that there were no difficulties posed by data which are synthetic or anonymized.

Another Member State highlighted that it is possible to retain the dataset within national legislation but there is a limit of five years. This implies that use of historical data is limited. Certain types of police data can be processed for a longer period.

Recommendation. The Commission should engage with the Member States and the central and national data protection authorities in order to provide a clear set of rules on dataset retention in the field of development, testing and training tools in the non-operational field of research and development.

5. What type of stakeholders can access and process the data (e.g. authorized LEA only; academia or private entities/industry partners under specific contractual agreement and supervision of the LEA)?

An over-riding theme in the responses from the Member States is that, in general, only authorized law enforcement agencies can access the data at national level. Where there is an intervention by an academic or system manufacturer partner, special confidentiality or non-disclosure agreements are imposed and access is minimized or limited, for purposes such as maintenance or error management and supervision is strict. Access is also limited by national law.

6. What are the main constraints that currently limit your sharing or pooling of data with other Member States or EU Agencies for development, training, testing, benchmarking, modelling and validation purposes?

- *Legal*

This question elicited a large body of responses.

It is clearly perceived, at Member State level, that the raft of European and national legislation is understood and interpreted differently across the European Union. This very perception undermines the intended concepts of free movement of data and common rules on the movement and processing of data. The responses include statements which set out that unless data sharing/pooling is specifically foreseen in national legislation then it simply will not happen without changes to that national legislation. For most respondents to the questionnaire, it is simply not clear what can be shared and under which common rules. This is the main constraint to data sharing.

This problem is compounded by different interpretations and implementation of purpose limitation. If the purpose limitation is set too narrowly the opportunities afforded by the Directive on understanding criminal phenomena from data and carrying out scientific research in that regard cannot be used to full effect.

Different rules on data retention will invariably cause difficulties for a shared dataset if, at some point, part of that dataset must be deleted during the lifecycle of an analytical tool which has been calibrated on that dataset and would be again, if the data were available.

The responses suggest that an absence of EU legislation or guidance on data retention and access for law enforcement and security purposes means that national law and Courts will seek to interpret Court of Justice rulings on data retention, either widely or narrowly. This again militates against a common set of procedures and understanding.

- *Ethical*

The responding Member States clearly appreciate the ethical issues involved in Big Data analysis and machine learning, including the question of how to assess and prevent the use of personal data in the absence of consent. This could be summarised as how to enshrine ethical constraints within agnostic, self-learning models. This is especially important in the face of potential criticism on the unethical bias in algorithms, the trustworthiness of results or observations generated by a machine before any human intervention on interpretation has taken place and, indeed, the trustworthiness of the law enforcement agencies pursuing such work, given their role in upholding the law. Not only will oversight by the relevant data protection be required this should really take the form of an ongoing partnership. Additionally, the type of information campaign launched with the second generation Schengen Information System (SISII), in partnership with the supervisory authorities, could play an important role in transparency but also showing European added value. This transparency should also extend to explanation about safeguards in place with academia and the private sector. All this complies with the notions of fairness, transparency and "explainability" which should always provide the cornerstone for forward-thinking, sensitive initiatives.

The use of operational data for research and development purposes raises the issue of how to anonymize sensitive content without losing information or detail that is relevant to the development and testing of tools.

The Commission's stated intention of pursuing a transparent "European way" is particularly relevant, as it is felt that there will have to be a high level of openness in order to ensure that there is not a groundswell of public opinion, especially if it is felt that certain sectors of society will be disproportionally and adversely affected by this project.

The discussion on ethics should be managed at the level of the European Union, so that a common view is achieved. Without this, fragmentation and suspicion are likely outcomes.

Recommendation. The discussion on ethics should be managed at the level of the European Union, so that a common view is achieved. Without this, fragmentation and suspicion are likely outcomes.

- *Technical*

One of the clearest observations from Member States is the problem caused by the quality and diversity of data sets. This is an interesting point; Big Data analysis is often vaunted as a solution for data analysis without manipulating data into a new format. This is indeed true but the pragmatic stakeholders point out that if the same object is described in several different ways and even grouped into larger categories that vary across the Member States, we would be starting from a position of complexity in that the programming of the analytical tools would have to be able to identify one type of object even though it had been described in radically different fashions even at the input stage.

One responding Member State highlighted previous technical implementations across the Union where the technical capabilities of the lowest quality national implementation became the de facto data standard.

Another Member State pointed out that ethical considerations have an impact on technical development and data quality. If anonymization is required, how can sufficient meaningful detail be retained for training and evaluating models?

One Member State highlighted that LEA's of the EU Member States have their own IT infrastructure and solutions to manage available information. Current technical advances restrict fast, efficient and up-to-date data transfer without involving a considerable amount of the institution's resources. The planned implementation of a Business Intelligence platform seeks to solve this problem.

Several Member States stated the need for high levels of technical and data security to prevent unauthorised access. For example, encryption should be used where possible as well as extensive logging of the actions performed especially on a shared platform.

Where the developer and the application provider of the national data asset management systems supervised at Ministry level are wholly state owned authorities, technical barriers are much less of a concern than financial and time barriers. However, this does not necessarily overcome the problem of the availability, formatting, quality and the amount of data for testing and training and the need to invest heavily to keep pace with the rapid development of AI technology.

Two Member States highlighted the need to work across data providers to improve the availability of common data identifiers and to develop a common EU-level ontology and taxonomy for data description.

The lack of a shared platform for storage and processing and related secure network for data transfer hinders sharing as there is simply no single way of achieving such an activity.

For a future dedicated space, an appropriate universal data model would have to be created. Every contributing party would have to map their data into this data model. This effort should not be underestimated.

- *Organisational*

Two Member States added "organisational" constraints to the questionnaire response.

These States pointed out that building shared datasets is an intensive effort with few short-term benefits. Teams that have the required skills and data-access struggle to get the required capacity available for shared projects. Such projects have to compete for attention with operational day-to-day concerns. Additionally, police services and their related Ministries might not have their own development capacities and self-sufficiency in high technology can be limited in extent.

There is considerable potential here for European added value.

7a. For the purposes of a common European Data Space for Law Enforcement what types of data would you be willing to share with other Member States or with an EU agency such as EUROPOL (or with other Member States via an EU agency)?*

The Member States indicated considerable willing to share data, within the obvious legal constraints.

The range of available data covered all data types, including:

- Open source.
- Operational LEA data.
- Anonymized data, including national anonymized teaching database.
- Synthetic data.
- Machine-generated test data.
- Information on analytical models and the models and products themselves.
- Non-personal data, such as statistics, modus operandi, trends, threat assessments, traffic accidents, offences.
- Image data (operational and identifiable) including photos and video.
- Audio data.
- Data on a case-by-case basis.
- Criminal, infringement and fines: administrative procedure related data.
- GIS data supported by geocoded position data – geospatial data.
- Electronically authenticated, digitally signed documents.

It was highlighted that it would be advantageous to be very focused in data sharing; that is identify the issue to be addressed and then to gather the dataset. A good example of this would be facial recognition in a law enforcement environment, often working with poor quality images in difficult circumstances. The test dataset should reflect this environment instead of being a range of standardised, good quality images.

Such an approach would also assist with standards and data standardisation, such as the way metadata are provided.

7b. Under what conditions would you be prepared to share the data you listed at 7a?

The overwhelming response concerned the absolute clarification of legal issues and the technical and organisational structures required at European level in order to ensure clear procedures, security, overall compliance with the law and no compromise of or adverse effect to the work of the providing authority. This would include nomination of the data controller, authentication, authorisation, auditing, logging and extent of permissible use procedures.

LEA's should be able use the shared datasets for training, test and benchmarking purposes but should be prevented from sharing them further. A framework of mandatory disclosure to the supervisory authority of this use could enable data sharing. This would also prevent misuse of data such as using data to evaluate a tool which has been used by a different organisation to build the tool.

It is also clear that there is a strong desire to separate the research and development from the operational sphere. Data provided for the former cannot find its way into the latter, especially in hit/no hit systems.

Recommendation. Datasets for training, test and benchmarking purposes should be clearly separated from operational systems. A framework of mandatory disclosure to the supervisory authority of this use could enable data sharing.

7c. In your work, have you identified any legal difficulty in achieving such sharing?

The majority of respondents stated, "yes". Reasons included:

- Lack of clear framework defining use and the protection of data provided and received.
- Lack of clarity over licensing issues.
- National constraints on data sharing. European and national rules are not clear enough.
- Information exchange is normally carried out in the framework of an investigation upon judicial request. Broader information sharing is out of the ordinary.
- Achieving a sufficient level of anonymization while retaining the usefulness of the data can be difficult.
- Often the purposes for the data collection are limited, creating many difficulties when talking about data sharing and auditing.
- A need for clarity regarding the interpretation and explanation of norms. As yet, there is not a lot of experience in the sharing of these types of data, there is no jurisprudence and little legal explanation on how to comply with legal and ethical norms.

One Member State pointed out that other information exchange models function well, such as EUCARIS. Others are under development. This would reinforce the need to set out a clear set of common rules for data sharing.

8. From a business point of view, what types of transaction/procedures (e.g. with regards pooling the data at national or EU level, organizing the access to data, or preparing specific datasets) would you wish to carry out in order to advance your work in development, testing, training, validation, benchmarking, modelling?

Member States were very forthcoming about the transactions and range of services they would wish to see. This included faster procedures at the EU level when approval is needed. As the responses are very varied they are reported here without much summarising.

- Pooling data:
  - Unified interfaces (authentication, data transmission, queuing systems)
  - Field-specific standardised data formats (like UMF for LEA)
  - Support for all character sets globally used
- Organizing the access to data:
  - Standardised querying interfaces
  - User rights management based on the "need to know" principle
  - Information about the reliability of the data
  - Logging and statistics
- Preparing specific datasets:
  - Support for transforming national data formats to and from EU or global standards

- Industry standards (open source) should be used wherever possible as proprietary systems may cause a large amount of adaptation.
- Some Member States find differences with systems and data standards within their own borders and therefore find the possibility of European solutions and standards attractive, as a way of tackling the problem. This would necessitate clear rules on data ownership and use.
- Sharing data (download and upload), models and code. Benchmarking and evaluating products. The possibility to comment on code, data or any artefact to foster information and facilitate exchange of data and knowledge.
- Comprehensive stocktaking exercise at national and EU level.
- Possible identification of similar cases among LEAs at European level and cases where victims and/or criminals are the same or they have similar profile.
- State-owned companies perform their tasks on the basis of legal designation; there is limited performance of such tasks. Business considerations such as this are difficult in case of state owned companies. It would be most optimal if the law enforcement agencies and the companies supporting them could carry out such developments and activities within EU cooperation and with EU funding as part of their already existing tasks.
- When checking with other parts of the public sector to see if an individual is on their system a manual process is completed on a case-by-case basis. Advancing this process to a more efficient direct 'hit/no hit' approach, and expanding such an approach to other areas of public sector data would be of benefit. Given that the stated objective of pooling data at EU level is to facilitate the uptake of AI solutions for law enforcement it is necessary to strongly consider the listed mandatory requirements applying to high risk AI applications in the EU White Paper on AI. The quality of training data is a function of the underlying processes and practices. It is necessary to establish how data modelling can support the achievement of better outcomes, i.e. first identify the desired outcome and then develop models on the basis of current data while recognising the limitations within that. Accordingly, the overarching challenge is to ensure that the data driving the development of AI systems is subject to a process of continuous improvement in line with the goal of delivering positive outcomes. In this respect the development of mechanisms by which to judge whether AI is leading to discriminatory practices is considered a prerequisite to the effective development and deployment of high-risk AI applications in the Justice sector. The extent to which current statistical modelling techniques could achieve this, or whether a bespoke formal process is required needs to be determined.
- The best way to answer this question is probably to start a small pilot and use case with a few LEA's to see what we need. It is necessary to establish how to align data handling procedures in an EU context. Additionally, sharing data in the national context between different legal regimes (for example: criminal law regime and health care) is difficult.
- A solution could be the migration of IT&C infrastructure to cloud services.
- Secure and common technical standards (e.g. API), pool of experts within the field of data engineering.
- It would be more helpful to address areas of commonality, such as irregular migrants, sexual child abuse, trafficking of human beings, firearms, cybercrime than areas of national specificity.
- As far as the organization of data at the national level is concerned, in addition to setting up the right algorithms, we want to enable searching in the database from other records as well.

Recommendation. The Commission and other stakeholders should draft and agree a list of functionalities to be introduced to the envisaged system based on the observations made by the respondents. A working group should be

established for this task, ideally involving the Member States who wish to be the first users of the central technical solution.

Recommendation. The Commission and other stakeholders should establish a working group to ensure that any new dataset required in the central system has a common ontology/taxonomy in order to facilitate searching and ensuring data standards and quality. This should include a standardised format and set of rules for the completion of metadata.

9. What additional shared technical facilities at EU level would you wish to have available in order to support the transactions/procedures you listed at question 8?

- One Member State highlighted the desire to have centralised (e.g. at eu-LISA) services for:
  - Normalisation of text thus enabling searches ignoring diacritic characters
  - Transliteration for non-Latin characters
  - Term banks for area-specific terms in all EU languages (like IATA) which can be integrated in queries
  - Phrase sets for standard communication
- Many databases do not offer a standardised way to query them, even if they contain the same type of information. Search portals should be able to convert the search criteria into the query logic of the databases they connect to and present the results in a unified way. Where this is not possible, users should be made aware of this fact to avoid false positive or false negative results.
- It is important that the selected technical solution is up-to-date and chosen for a long term.
- Implementation and development intermediate system for processing requests.
- A hosting platform for data and models for training, testing and evaluating AI models and products.
- A common user interface/platform for all EU LEA Users, i) in order to have a common base of how to handle the results, ii) where the information of each Member State could be evaluated and processed in order to extract valuable results that could support the work of all participating Agencies.
- Big Data analysis tools (the most up-to-date and validated ones).
- Pattern recognition and pattern matching (applying proper algorithms on criminal cases, similarities of cases can be revealed immediately and efficiently).
- Development, application and integration of transparent interoperability data exchange and process control technologies related to European law enforcement systems.
- The recent EU-level discussions on the setting up of an EU Innovation Hub are relevant to the stated objective of pooling data at EU level to facilitate the uptake of AI solutions for law. In this respect the proposed EU 'Hub', collaborating with Member State Innovation Centres, would seem the logical home for the development of an inventory of 'use-case' applications of AI technologies for law enforcement purposes. Such an inventory may benefit from grouping use-cases according to type of law enforcement need, for example, AI tech supporting: administrative efficiencies; investigative & forensic capabilities; and predictive analytics (for crime forecasting, and in managing the risk of recidivism and the risk of being a victim of violent crime). The development of a use-case inventory is also consistent with the need to elaborate the ethical principles for AI of fairness, transparency and "explainability" by way of practical applications of AI technology in the justice field.
- At present, the greatest need is for a platform to exchange experiences. Only later can specific technical facilities be considered.

- One Member State expressed a need for a place where LEAs can download well-described, proven, secure, and maintained algorithms and tools; e.g. it is waste of time that each Member State develops its own entity recognition model for English language.

10. What legal conditions would facilitate your work?

Member States reported that a Common understanding among domestic stakeholders regarding legal limitations should be developed. The following points were raised:

- Establish licensing usage agreements. Clarify product and data ownership. Clarify rights and responsibilities of users and institutions. Establish terms of use and clarify financing issues, should they be necessary. Clarify storage and platform hosting issues to enable sharing.
- Simple and smooth conditions/prerequisites for accessing other information systems and also for data sharing in order to train algorithms and provide better quality results.
- More efficient treatment for proposals which, many times, fall under shared ministries and other public entities' competence.
- The current legislation is not clear and Member States have different implementations of the Directive. There need to be a common understanding and guidance as to the legal basis for access to data and data retention of operational datasets for the specific purposes of development of AI solutions for law enforcement. This includes clarity as to the legal basis for the use of non-operational datasets (inter alia, synthetic, pseudonymized) to develop AI solutions for law enforcement as consistent with the mandatory requirements applying to high risk applications, and with the ethical principles of fairness, transparency and "explainability".
- For development of some police tools personal data would most likely have to be used for tests, validation etc. so further considerations are most likely required from a data protection and confidentiality point of view.
- The focus should be on what is possible, not on what is not possible. Furthermore, for training purposes, often data maximisation is required to prevent bias or skewed models. The current legislation has a flaw; it does not take training data into account but focuses on using (minimal) data in models/algorithms. Data retention periods also differ between different EU states. With respect to the ability to get access to third party data, specific for use by LEA's (e.g. Social Media, automotive, telecom, financial) the European Commission could play a role.
- There should be a clear agreement on data ownership. Additionally, there are legal constraints regarding intellectual property rights, trade secrets and private international law when working with private parties.
- One of the problem is that is not always clear when we could process user data from social networks. This is especially problematic in H2020 project.

11a. In relation to research carried out by third parties (e.g. private enterprise, universities):

1. *Do you have available data for research purposes (e.g. data that could be made available to researchers that are not under the direct supervision of LEA's or the Ministry of the Interior?).*

Nine Member States responded, "yes" to this question, five Member States responded, "no".

For several Member States, these arrangements are still under development. Equally, several Member States indicated limitations on such sharing. There would be value in describing the partnerships that are evolving and the legal considerations. This could form part of an information-sharing platform.

2. If yes, what types of data?

Generally, the data available for such sharing would be standardized specifications regarding criminal offences (anonymized) for research and statistical purposes only. Public statistics, anonymized or synthetic data might also be made available.

3. If yes, would you share that data with other MS or EUROPOL for similar purposes?

This seems to be largely dependent on several factors:

- There are few problems with publicly available statistics.
- Non-disclosure agreements might have to be in place.
- The source of the data is key. Open source provides few difficulties but the issue is that 'private data' for LEAs should not be used for development of products, because it then cannot be used to benchmark and evaluate products. Establishing separate data pools for R&D and benchmarking purposes could help overcome this issue.
- Whether a clear legal basis for sharing can be identified. In some cases, there can simply be no sharing.
- Publicly available data might be accessible but is it sufficient for development of law enforcement machine learning models? Often, the dataset should reflect the adverse conditions of collection in order to be a realistic basis for the operational dataset to be analysed (e.g. poor quality surveillance images as opposed to images captured in controlled conditions).

4. Under which conditions could they be shared?

In brief, the conditions required for data sharing are based on clear legal compliance, existing frameworks, non-disclosure agreements and a case-by-case consideration following an appropriate request. As this is a relatively new field, there is a lack of clarity on how to share "regularly", purpose limitation being a very important legal condition.

11b. As an example: a private company has independently developed a new tool for image recognition in cases of child abuse. After a first set of tests on their own dataset (which does not include real images), they wish to test and validate this solution on real data. Would that be possible in your country?

The general view is that such a practice would not be possible. In the isolated cases where it would be possible, there would be restrictive conditions, such as the test environment or indeed the whole infrastructure being set up in law enforcement premises and all testing carried out by the LEA itself with no participation by the private developer. It might be possible to share if the development and testing were carried out by a trusted partner, such as Interpol.

12. Any other information.

Several Member States indicated that they have ongoing projects on subjects such as open source intelligence capabilities and scientific research pools. Results from these activities are not yet available.

It is clear that there is considerable activity and a desire that the goal for the creation of a platform would be to facilitate trusted use and sharing of data and models within LEAs. This would have to be enshrined in a carefully crafted concept of rights and roles, access to and storage on such a platform.

Looking further ahead, regardless of the quality of results of new algorithms-systems and their intelligence, there is a concern of whether/how these final output-results and findings can be used: i) in the context of case investigations, and ii) can be used and recognized as valid in the court of law. These are important factors which, later, need to be taken into consideration.

One Member State is planning to develop and integrate AI and Big Data system solutions that support civil and law enforcement goals with the support of Universities. In line with this, an AI Lab has been established with the support of the Ministry of Interior. These technologies and data can be shared with the Member States on the basis of proper legal authorisation.

Prior to the availability of a shared data pool, there is still the possibility of Member States sharing data to facilitate development within national secure environments. This however, will not fully lead to the development of EU-wide standards on data description.

9. THE LEGAL ENVIRONMENT

9.1. Data protection

Introduction

In order to make this section more "readable", it will take the form of key questions and answers. It should be read in conjunction with the legal texts.

Data protection – what does the law intend?

The two principal instruments on data protection are Regulation (EU) 2016/679[20] and Directive (EU) 2016/680[21]. In short, the Regulation is the General Data Protection Regulation and the Directive specifically covers the processing of personal data (pertaining to natural persons)[22] by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties[23]. In addition, since 2019, there is a specific piece of EU legislation on the free-movement of non-personal data[24], important when datasets contain both personal and non-personal data. Recitals 18 and 19 of this Regulation concentrate on public security and its context.

Very importantly, both Regulation (EU) 2016/679 and Directive (EU) 2016/680 were passed on the same day and it is clear from the cross-references held in the texts that they should be addressed together. There is even explanation of the scope of the Directive and at which point matters pass from the Directive to the Regulation. Member States may entrust competent authorities, within the meaning of the Directive, with tasks which are not necessarily carried out for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties,

[20] 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)

[21] DIRECTIVE (EU) 2016/680 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 by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data

[22] It is important to specify data pertaining to natural persons as geographical data, such as street signs, data on firms and data pertaining to deceased people are all exempt

[23] There is also an important need to study the legal basis of EUROPOL to ensure that unnecessary or unintended impediment to data sharing does not hinder development. There are further points on EUROPOL and data protection later in the report.

[24] Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union

including the safeguarding against and prevention of threats to public security. Where the processing of personal data is for those other purposes, in so far as it is within the scope of Union law, it falls within the scope of the Regulation.

Both instruments provide a big hint, in their titles, as to their intention. Both cover the processing of personal data and the free movement of such data. Therefore, it was always the legislators' intention to facilitate the free movement of personal data whilst setting out the conditions for that movement to take place. The legislation should be read in this light. It is useful to read both instruments as one sets out the "big picture" whilst the other recognises some of the unique challenges of law enforcement.

Although the recitals to the instruments are not enforceable, they do provide necessary background to the legislation and an explanatory text as to what the later Articles intend to achieve.

The Regulation[25] sets out that the exchange of personal data … across the Union has increased and that national authorities in the Member States are being called upon by Union law to cooperate and exchange personal data so as to be able to perform their duties or carry out tasks on behalf of an authority in another Member State.

Further, the Regulation[26] describes the situation whereby the objectives and principles of the previous (repealed) legal instrument (Directive 95/46/EC) remained sound, but it had not prevented fragmentation in the implementation of data protection across the Union, legal uncertainty or a widespread public perception that there are significant risks to the protection of natural persons. These differences in the level of protection of the rights and freedoms of natural persons, in particular the right to the protection of personal data, with regard to the processing of personal data in the Member States may prevent the free flow of personal data throughout the Union. Those differences may therefore impede authorities in the discharge of their responsibilities under Union law. Such a difference in levels of protection was considered to be due to the existence of differences in the implementation and application of Directive 95/46/EC.

Please note that activities that fall outside of the scope of Union law are not covered under this legislation. This includes activities which are considered matters of national security.

What is the environment framing this study?

It is important to recall that the context for this study is the provision of data, in tightly controlled circumstances, for development, testing and proving purposes, not for an operational system. Accordingly, the term "data laboratory" is very apt. The concerns set out at Article 28.1.a. of the Directive (prior consultation of the supervisory authority) are largely absent in that the controlled environment of a data laboratory allows considerable risk mitigation measures. However, the legislation foresaw the use of new technologies, mechanisms and procedures, even though again the risk to rights and freedoms can be strictly controlled. Accordingly, it would seem legally and politically appropriate for transparency to prevail and for the data controllers involved to consult their supervisory authorities. There may be considerable added value in involving the Board[27] in order for it to undertake its role in ensuring that there is no divergence in interpretation which might frustrate the development of appropriate technology when this was not the intention of the legislation.

[25] Regulation (EU) 2016/679. Recital 5
[26] Ibid. Recital 9
[27] See Article 68 Regulation (EU) 2016/679

How can we maximise data usage?

The principal solution to maximising data usage is to ensure that Member States each face the same legislative framework. The Directive states[28] that in order to ensure a consistent and high level of protection and to remove the obstacles to flows of personal data within the Union, the level of protection of the rights and freedoms of natural persons with regard to the processing of such data should be equivalent in all Member States. Consistent and homogenous application of the rules for the protection of the fundamental rights and freedoms of natural persons with regard to the processing of personal data should be ensured throughout the Union. This is mirrored in the Directive[29] where it is stated that ensuring a consistent and high level of protection of the personal data of natural persons and facilitating the exchange of personal data between competent authorities of Members States is crucial in order to ensure effective judicial cooperation in criminal matters and police cooperation. To that end, the level of protection of the rights and freedoms of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security, **should be equivalent in all Member States.**

What about data maximisation?

There is a separate concept, data maximisation. In this, the largest appropriate dataset is assembled in order the seek to reduce bias in the algorithm. The current legislation may introduce a difficulty, in that it seems to not take maximum (and therefore least biased) training data into account but focuses on the importance of using the minimum amount of data in models/algorithms.

Recommendation: Legal clarification should be sought to ensure that the least biased dataset (probably a maximal dataset) can be developed for testing and training algorithms. When this can be achieved, the later operational use of the algorithms should be better-placed to reduce unfavourable/undesirable outcomes.

Is there any margin for legal manoeuvre at the Member State level?

Yes. The Regulation and Directive[30] also set out that there is a margin of manoeuvre for a Member State to specify its rules, including for the processing of sensitive data. To that extent, the legislation does not exclude Member State law that sets out the circumstances for specific processing situations, including determining more precisely the conditions under which the processing of personal data is lawful.

This is welcome, as long as the margin of manoeuvre does not result in a national specification which militates against the twin goals of the legislation: correct processing AND free movement of personal data.

If I share data with another Member State can I impose conditions that I could not impose in my own country?

No. The Directive[31] sets out that Member States shall ensure that the transmitting competent authority does not apply specific conditions to data shared to recipients in other Member States or to agencies, offices and bodies established pursuant to Chapters 4 and 5 of Title V of the TFEU which would differ from those applicable to similar transmissions of data within the Member State of the transmitting competent authority.

[28] Ibid. Recital 10
[29] Directive (EU) 2016/680. Recital 7
[30] Regulation (EU) 2016/679. Recital 10 and Directive (EU) 2016. Recital 15
[31] Directive (EU) 2016/680. Article 9.4

This re-enforces the long-standing principle of availability of information between Member States.

However, as with existing sharing of data between Member States where Union or Member State law provides for specific conditions applicable in specific circumstances to the processing of personal data, such as the use of handling codes, the transmitting competent authority should inform the recipient of such personal data of those conditions and the requirement to respect them. This covers issues such as a prohibition against transmitting the personal data further to others, or using them for purposes other than those for which they were transmitted to the recipient, or informing the data subject in the case of a limitation of the right of information without the prior approval of the transmitting competent authority.

What are the risks if Member States apply diverging rules on protection and free movement of data?

This risk is highlighted in the Directive[32] where we see that in order to ensure the same level of protection for natural persons through legally enforceable rights throughout the Union and to prevent divergences hampering the exchange of personal data between competent authorities, the Directive should provide for harmonised rules for the protection and the free movement of personal data processed for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security. The approximation of Member States' laws should not result in any lessening of the personal data protection they afford but should, on the contrary, seek to ensure a high level of protection within the Union.

The final sentence of the relevant recital states that Member States should not be precluded from providing higher safeguards than those established in this Directive for the protection of the rights and freedoms of the data subject with regard to the processing of personal data by competent authorities. Again, it is important that a national implementation does not remove the balance between protection and free movement as this would lose sight of the legislators' intention, remove the "level playing field" within the Union and effectively subvert the intention of the instrument. The Board, introduced by the Regulation and described in the Directive has a vital role in contributing to the consistent application of the Directive throughout the Union, including advising the Commission and promoting the cooperation of the supervisory authorities throughout the Union. This is important, as it places the Board in a position of facilitating the intent of the legislation and highlighting where correct processing AND free movement of data have been frustrated or addressed incorrectly.

Recommendation: The Commission should approach the European Data Protection Supervisor and the Board in order to carry out a short study on whether there are divergences in national law and interpretation of the Directive which might cause a difference at national level in the ability to share or retain data.

At this stage we are in the realms of research and development. What does the law say on this?

The scope of this study covers development, testing, training and similar activities, carried out by competent authorities, concerning algorithms for analysing very large amounts of data in the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security.

[32] Ibid. Recital 15

Where that activity is carried out by a competent authority within the meaning of the Directive, the Directive applies. Where the activity does not fit the law enforcement definition above or the activity is carried out by a third party which is not a competent authority, the Regulation applies.

Much academic study on the use of Big Data analysis covers the potential for bias and the need for sufficiently large and representative data pool sizes to minimise or remove bias. Therefore, it is important at the development, testing and training stages to ensure that the algorithms developed are sufficiently academically sound for their future use in an operational sphere. Accordingly, we are in the realms of data science. Article 3 of the Directive covers this eventuality, recognising that processing by the same or another controller may include archiving in the public interest, **scientific**, statistical or historical use, for the purposes of the Directive, subject to appropriate safeguards for the rights and freedoms of data subjects. This is reinforced by the explanation[33] that for the prevention, investigation and prosecution of criminal offences, it is necessary for competent authorities to process personal data collected in the context of the prevention, investigation, detection or prosecution of specific criminal offences beyond that context in order to develop an understanding of criminal activities and to make links between different criminal offences detected.

The EDPS has published an opinion on scientific research and data protection[34], but it would not seem to be based in the same context (clinical trials, commercial use of data). The explanation at Recital 27, explained above, is specifically targeted on law enforcement activities and the need for data to be analysed in order to be able to understand criminal activities and make links between criminal offences. On this basis, the Directive directly addresses the issue and stresses that law enforcement authorities have the ability to carry out such analysis. It is vital, therefore, that these authorities reflect this possibility in their statement on the purposes for analysing data gathered. Naturally, any data analysis outside these purposes would fall under a different legal regime.

Are we allowed to process data for a purpose other than that for which they were originally gathered?

Yes, under prescribed circumstances.

In the framework of Big Data analysis for law enforcement purposes, the data gathered must have been obtained lawfully. By this we mean that, for example, the video footage, the social media messages, the scans of vehicle registration plates must have been collected within national law. For further processing to take place, these data must have been collected for specified, explicit and legitimate purposes within the scope of the Directive[35].

Further the data should not be processed for purposes incompatible with the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security.

The explanatory recital continues by explaining that it is possible to further process those data for a purpose (within the Directive) other than that for which they have been collected, as long as such processing is authorised in accordance with applicable legal provisions and is necessary for and proportionate to that other purpose (also see the explanation of Recital 27 of the Directive in the previous question).

[33] Ibid. Recital 27
[34] Preliminary Opinion on data protection and scientific research. EDPS. 06.01.2020.
[35] Ibid. Recital 29

Where personal data were initially collected by a competent authority for one of the purposes of the Directive, if there is a lawful need to process the data outside those purposes Regulation (EU) 2016/679 should apply to the processing of those data.

Is there always only one data controller?

No. Continuing from the above answer, the Directive explains that this further analysis can take place under the auspices of the original data controller or another controller, thereby explicitly accepting the transfer of data for further processing, under given conditions.

Pseudonymization and anonymization

The principles of data protection apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymization, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly.

In the sphere of a data laboratory the data are not being used operationally and in very controlled circumstances, probably under conditions agreed with the supervisory authority as foreseen in Article 28 of the Directive on prior consultation.

Where there is the possibility to anonymize data entirely, preventing the identification of the data subject, the principles of data protection do not apply.

Recommendation. The Commission and Member States should carry out a limited study and literature review on the extent to which anonymization may be used in constructing a dataset without compromising the usefulness of the dataset for testing analytical tools.

How should we handle user access, that is, the different roles of authorities accessing the data?

As highlighted in the introduction to this section, the Directive applies to competent authorities when acting within a specified set of responsibilities which could be described as criminal law enforcement. Another body or entity which processes personal data on behalf of such authorities, within the scope of this Directive, should be bound by a contract or other legal act and by the provisions applicable to processors pursuant to this Directive in order to be able to process such data. Any processing that falls outside such a relationship or not for such law enforcement purposes is covered by the Regulation.

I need to keep a test data set. What are the rules on deletion of data?

Under Article 4.1.e of the Directive Member States shall provide for personal data to be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which they are processed. Therefore, if the purpose is set too narrowly the opportunities for dataset retention are limited. Article 4.3. permits that processing by the same or another controller may include archiving in the public interest, **scientific**, statistical or historical use, subject to appropriate safeguards for the rights and freedoms of data subjects. Article 5 describes the necessity for **time-limits for storage and review in that** appropriate time limits must be established for the erasure of personal data or for a periodic review of the need for the storage of personal data. Procedural measures shall ensure that those time limits are observed.

Recommendation. Where data are to be used for non-operational scientific purposes, as permitted by the Directive, in addition to their original operational

use, this scientific use should be reflected in the stated purpose for the use of the data so that the issues of purpose limitation are managed transparently.

9.2. Data "ownership"

In strict terms, personal data belong to the data subject. However, for the sake of this study we will concentrate on data controllers and processors[36]. It is clear that the legislation intends sharing of data as it sets out provisions for there being more than one data controller. Member States shall, where two or more controllers jointly determine the purposes and means of processing, provide for them to be joint controllers. They shall, in a transparent manner, determine their respective responsibilities for compliance with the Directive[37].

In this way, the legislation allows data to be transferred, processed for certain purposes (which are in fact, quite broad) and all responsibilities set out transparently. As some of the processes might be viewed as novel, due to emerging technical capabilities, the provisions on prior consultation of the supervisory authority would seem to be appropriate, especially as this would also allow discussion on supervision, access, processing and logging in order to maintain a transparent data processing regime.

Recommendation. When a common set of rules is established on the sharing of test data it should include a format for describing the respective responsibilities of the joint data controllers.

10. Conclusions

The key to success is the ability to coordinate a number of activities, some of which are politically-charged.

To illustrate this statement, there is a need to ensure that the intentions of the data protection legislation are capable of being achieved, both at EU and national level. If not, a divergence of legislation, which was repeatedly highlighted in the text itself as something to be avoided, will inevitably cause a blockage to the desired free movement of data. This situation should be verified.

Once the harmonised legal situation can be seen to be a reality, the interpretation of what can be done in both data sharing and data retention, in the research and development realm of data science to support law enforcement, should be subject to common guidelines which are agreed by data science practitioners and their partners, the data protection supervisory authorities. The notion of partner is important. The European project on data spaces wishes to progress in a very European way, that is one based on being open, fair, diverse, democratic, and confidence-inspiring. This would be very difficult to achieve unless there is a long-term, transparent relationship between these partners based on how to achieve what needs to be done to keep citizens safe in a manner that is lawful and instils trust. This means that the supervisory authorities need to have sufficient resources and a mandate to enter this dynamic partnership.

The approach to be taken to the gathering of datasets should be based on a problem-solving approach, that is, a dataset that most accurately reflects the operational problem to be solved and can be scientifically demonstrated to be the most relevant. This will entail discussion on issues such as anonymization, that is, at what point do the data lose their usefulness to the exercise through the removal of personal detail. Additionally, there will need to be discussion on the size of the dataset, which needs to be sufficiently large to be able to show that bias has been minimised. Throughout the development

[36] Directive (EU) 2016/680. Article 3.8 & 3.9
[37] Ibid. Article 21.1

lifecycle, as the tool is enhanced and, essentially, recalibrated there is a need to retain the test dataset in order to recreate earlier test results; this suggests rules on data retention and the need to ensure that purpose limitation is set broadly enough for research and development-use data to be retained once the original operational data have been deleted. Naturally, this should be under the tightest separation and control. There is the potential for data protection officers to play an important role here.

Concerning the hosting of any centralised system, there is a need to address the legal basis of any potential hosting agency to ensure that later legal problems will not arise, such as time limitation on pilot projects. The necessary legal and budgetary provisions will have to be coordinated.

Regarding use, it is clear that EUROPOL has a critical role in not only analysing data but also carrying out common training across the EU to ensure that the level of understanding is shared. In the training sphere, the legal basis is already in place, however, in the processing of data, there is a need to ensure that EUROPOL can play its full role.

The divergences in source systems, data standards and data quality across the European Union make life more complex for those designing analytical tools. Ideally, within the working structures supporting this initiative a data quality and standards group would set rules for the datasets, including how the metadata are described. This would facilitate searching the available datasets but also assist the data owner(s) in managing access to the dataset.

The proposal is to start small and grow incrementally. There will be a need to carry out evaluations, not only of individual problem-focused initiatives but of the capacity of the central system in relation to the user demands placed upon it and also the extension of micro-services to support the users. This should lead to a working group on evaluation and technical development.

In summary, there will be an intensive demand, probably over a five-year period but persisting after that, to align data protection, several strands of legislation, rules, appropriate partnerships, project selection, data standards/quality, allocation of roles, evaluation, technical development and a coordinating mechanism in a transparent, trust-building environment. It is a tall order but the goal is worthwhile.

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

–by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

–at the following standard number: +32 22999696, or

–by email via: https://europa.eu/european-union/contact_en

DR-01-21-054-EN-N

Publications Office
of the European Union

doi: 10.2837/289402
ISBN 978-92-76-29921-9