DATA PITCH

H2020-ICT-2016-1

Project Number: 732506

D3.1 – Legal and Privacy Toolkit v1.0

Coordinator: Dr Sophie Stalla-Bourdillon
With contributions from: Alison Knight
Quality reviewer: ODI

Deliverable nature:	Report (R)
Dissemination level: (Confidentiality)	Public (PU)
Contractual delivery date:	30 June 2017
Actual delivery date:	30 June 2017
Version:	1.0
Total number of pages:	75
Keywords:	Law, regulatory compliance, intellectual property rights, data protection, personal data, anonymisation, pseudonymisation, re-identification risk, impact assessments, ethics

Abstract

The legal and privacy toolkit is a crucial component of the Data Pitch project (the Project). The aim for this first version of the toolkit is to address the question - "What are the critical things to consider when sharing and reusing data for defined innovative purposes under the Project?" The focus of the toolkit is on ensuring that legal rights in relation to shared data are respected, and legal obligations in respect of such data are followed in the Project. In particular, the toolkit provides information on how personal data will be anonymised before it is shared for analysis by SMEs taking part in the Project, and will describe techniques to mitigate the risk of harm befalling data subjects as a consequence of data relating to them being processed in new ways (for secondary purposes).

As a project deliverable, this toolkit also provides an overview of the legal and regulatory framework that applies to data sharing and data reuse, including setting out key considerations governing the data sharing arrangements between the parties involved in the Project. The toolkit provides a 'map' of the relevant legal issues arising in the context of data sharing, and an outline of a methodology for handling these in a suitably risk-averse manner. This framework aims to treats data ethically and responsibly, with comprehensive, yet pragmatic guidance on data disclosure and its handling.

The need for a toolkit under the Project is against a backdrop of emerging legal complexities in handling and repurposing data in a 'big data' environment in ways that are fair. This is especially important where personal data are to be subject to secondary processing. The vast power of data analytics means that absolute anonymisation of personal data may not be possible, so organisations have to focus on mitigating the risks associated with data sharing to the point where they are remote. Finding the correct balance between innovation and data protection without running the risks of over-reaction can be extremely difficult. This toolkit sets out a structured strategic approach to managing those potential issues that arise upfront, our goal is to make it easier to innovate safely and fairly. In this way, the toolkit hopes to present solutions to achieve legally-compliant data use in technically-enhanced and risk-aware ways, which also allows organisations to obtain high levels of utility from analysing shared datasets not currently in the public domain and hence to bring better services, products and research to the public.

Disclaimer: The content of this toolkit does not constitute legal advice. If in doubt, you should always contact a lawyer.

Executive Summary

The legal and privacy toolkit is a crucial component of the Data Pitch project (the Project).

The Project is specifically designed around the sharing and reuse of closed data, although there may be some use of open data as part of the Project.

The legal and privacy toolkit aims to provide an overview of the legal and regulatory framework that applies to data sharing and data reuse essentially for closed data, especially in respect of privacy and data protection law obligations. It also sets out key considerations governing data sharing arrangements, such as in respect of licensing intellectual property rights.

Key areas of law relevant to data sharing and reuse are described. These include:

- **Data protection laws** – these set the rules for processing personal data.
- **Electronic privacy laws** – these govern the protection of privacy in the electronic communications sector.
- **Intellectual property laws** – these encompass a number of different rights that may be asserted of a more proprietary type, including in association with the use of data and its expression.
- **Competition laws** – these aim to prevent anti-competitive harm that might result from commercial activities, including from the sharing of information.
- **Laws of confidentiality** – these protect confidential information.
- **Contract laws** – these govern the ways in which private parties can agree to work together, including in respect of data sharing agreements that include certain rights and obligations regarding data usage and access. Ultimately, if terms in agreed contracts are broken, contracting parties could try to enforce such terms in a court of law.

In particular, guidance is given on how personal data should be anonymised/pseudonymised before being secondarily processed for innovation acceleration purposes, and on ways to mitigate the risk of harm to data subjects.

To achieve these outputs, the toolkit describes certain techniques and measures (legal, organisational, and technical), which comprise part of an overall, dynamic risk-management approach adopted under the Project to reduce the likeliness and severity of any harm potentially caused by the secondary processing of data. These legal issues apply at two key stages: before data is shared; and around access to data. In advance of sharing data between two organisations, for example, each must decide how they can comply with relevant laws once data is transferred. Security safeguards and any privacy measures that should be put in place also need to be considered carefully. On occasion, it may not be possible to guarantee the exclusion of the application of data protection law when data relating to persons is shared, while simultaneously preserving data utility when data relating to persons are to be processed for secondary research purposes under the Project. In such cases, further processing will only be permitted after certain measures have been put in place so that legally-compliant access and secondary processing activities are assured.

Four possible options for storing and providing access to data sets shared by the Data Providers by the Participating SMEs have been considered:

1) The University of Southampton hosts the data. Multiple levels of security (data hosting facilities using secure infrastructure managed by the Consortium at the University of Southampton) can be offered depending on what is deemed appropriate to the level of re-identification assessed to be present. These include: connection to the public internet through Janet, the UK university network; or, defining a secure zone Data Pitch via a standard proxy - an academic infrastructure hosting all the academic services and laboratories (this can be supplemented by the addition of a further proxy to create an isolated and secure zone for Data Pitch processing which is managed independently).

2) The Data Provider hosts the data.

3) The commercial cloud hosts the data (under the direction of the Data Provider).

4) The participating SME (chosen to process the relevant data) hosts it.

Ultimately, the option of choice for the secure provision of data relating to individual to be shared by the Data Providers will be led by their preferences and their current legal compliance measures taken as data controllers. Notably, options (1) or (2) are the Consortium's preferred option for hosting data relating to persons under the Project.

The toolkit achieves the overview with reference to existing research and guidance that reflect best practices in this field (including from guidelines, checklists, opinions, and recommendations) based on EU standards. To this end, the toolkit takes into consideration relevant features from different legal systems across the EU that restrict the sharing and reuse of datasets, together with associated rights that are raised. It also set out the outline of a methodology to guide participants in the Data Pitch project from different countries through this area in a way that implements a uniform mechanism.

The toolkit concludes with a list of crucial points to consider for the sharing and re-using of data, which are of interest to both data providers and data recipients:

- Make sure you check the legal agreement you have signed. In the case of Participating SMEs, check the nature of the limitations included which could reduce your planned re-usage of the data being shared with you and in particular the allocation of intellectual property rights between the data provider and the data recipient. One key distinction to bear in mind in order to adequately allocate intellectual property rights is the distinction between the algorithm produced to process the shared data and the output of the processing activity through the means of the algorithm, which could be described as enriched or output data.
- Make sure that the data does not allow individuals to be identified if combined with other information available to you. To note, it is not always necessary to have biographical details in data - such as a requirement for an individual to be named – for it to be deemed personal data. Identification (and re-identification) may be effected in other ways, notably through singling out individuals from others in a dataset via indirect identifiers.
- The risk of re-identification through data linkage is unpredictable because it can never be assessed with certainty what data are already available or what data may be released in the future. On the one hand, while it may not be possible to determine with absolute certainty that no individual will ever be identified as a result of the disclosure of anonymised data, a certain amount of pragmatism needs to be adopted. It involves more than making an educated guess that information is about someone.
- The likelihood and severity of re-identification risk occurring can also change over time (e.g. with the introduction of new technologies that can link data) and, therefore, re-assessments should be carried out periodically and any new risks managed. This should include trying to determine what additional information - personal data or not – could become available that could be linked to the data to result in re-identification.
- General and data-specific safeguards are set out in the legal agreements governing the reuse of data shared. Above and beyond fulfilling their contractual obligations, there is no 'one size fits all' solution to data security – each organisation should adopt a risk-based approach to its assessment and management in conjunction with implementing advice given by the Consortium. Different measures (and combinations of measures - legal, technological, and organisational) may be appropriate depending on the processing activity and other data environment contextual factors.
- Any risk of harm to individuals who are the subjects of anonymised/pseudonymised data should be avoided. This includes consideration of harm that might result from accidental loss or unauthorised

processing by others of data shared under the Project, bearing in mind the nature of the data being processed and any possible sensitivity.

- Any breach of data protection law must be reported immediately to the Consortium and the relevant Data Provider must act accordingly as soon as possible to mitigate any risk of harm to data subjects.

Given the nature of this document and the fact that the legal landscape is constantly evolving, this toolkit (and its final version due for publication in 2019) will be updated periodically at http://datapitch.eu/policytoolkit. This interactive format reflects the anticipated emerging demands for guidance in specific areas in association with the key features of this toolkit.

Disclaimer: The content of this toolkit does not constitute legal advice. If in doubt, you should always contact a lawyer.

Document Information

IST Project Number	732506	**Acronym**	DATA PITCH
Full Title	Data Pitch		
Project URL	http://datapitch.eu		
Document URL			
EU Project Officer	Francesco Barbato		

Deliverable	**Number**	D3.1	**Title**	Legal and privacy toolkit v1
Work Package	**Number**	WP3	**Title**	Data and data providers' liaison

Date of Delivery	**Contractual**	M7	**Actual**	M7
Status	version 1.0		final √	
Nature	prototype □ report √ dissemination □			
Dissemination level	public √ consortium □			

Authors (Partner)	University of Southampton			
Responsible Author	**Name**	Dr. Sophie Stalla-Bourdillon	**E-mail**	S.Stalla-Bourdillon@soton.ac.uk
	Partner	University of Southampton	**Phone**	023 8059 34 14

Abstract (for dissemination)	The legal and privacy toolkit is a crucial component of the Data Pitch project (the Project). The aim for this first version of the toolkit is to address the question - "What are the critical things to consider when sharing and reusing data for defined innovative purposes under the Project?" The focus of the toolkit is on ensuring that legal rights in relation to shared data are respected, and legal obligations in respect of such data are followed in the Project. In particular, the toolkit provides information on how personal data will be anonymised before it is shared for analysis by SMEs taking part in the Project, and will describe techniques to mitigate the risk of harm befalling data subjects as a consequence of data relating to them being processed in new ways (for secondary purposes).
Keywords	Law, regulatory compliance, intellectual property rights, data protection, personal data, anonymisation, pseudonymisation, re-identification risk, impact assessments, ethics

Version Log			
Issue Date	**Rev. No.**	**Author**	**Change**

Table of Contents

Abbreviations

Art.29 WP = Article 29 (EU Data Protection) Working Party

DPA = UK Data Protection Act 1998

Directive = EU Data Protection Directive 95/95/46/EC

GDPR = EU General Data Protection Regulation 2016/679

ICO = UK Information Commissioner's Office

Definitions

Anonymous information - information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable

Big data – extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions

Data controller - an organisation which alone or jointly with others determines the purposes and means of the processing of personal data

Data processor – an organisation which processes personal data on behalf of the controller

Data Providers – those organisations that agree to supply their data under the Project

Data subject - a living individual who is the subject of personal data

Direct identifier - data that directly identifies a single individual

Indirect identifier - data that indirectly identifies a single individual

Mosaic effect - when it is possible to determine an individual data subject's identity without having access to obvious (direct) identifiers by correlating data pertaining to the individual across numerous datasets or intra-dataset; whereas individual identifiers in these datasets would not otherwise allow a data subject to be re-identified.

Participating SMEs – those organisations that are selected to process the Data Providers' data under the acceleration stage of the Project

Personal data – three alternative legislative definitions are discussed:

- **Directive (to apply until 24 May 2018):** "*any information relating to an identified or identifiable natural person*" (Article 2(a))
- **DPA (to apply under UK law until 24 May 2018)**: *"data which relate to a living individual who can be identified— (a) from those data, or (b) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller, and includes any expression of opinion about the individual and any indication of the intentions of the data controller or any other person in respect of the individual"* (section 1(1))
- **GDPR (to apply from 25 May 2018):** *"any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person"* (Article 4(1))

Profiling – *"any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements"* (GDPR, Article 4(4))

Pseudonymisation – *"the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person"*(GDPR, Article 4(5))

Secondary processing - any processing of data after its initial collection

1 Introduction

1.1 Aims of the Data Pitch Project

The overall aim of the 3-year EU Horizon 2020 (H2020)-funded Data Pitch programme is to increase the value of data and to enable small European companies to reap the full benefits of secure data sharing and reuse. This toolkit is aimed at giving those organisation that take part in the Project confidence to innovate with data in legally-compliant and ethical ways.

In brief, under the Project, Participating SMEs will be able to securely access and process data shared by Data Providers after they have qualified to take part in the main acceleration (experimental) stage of the Project. Such data has been collected externally by the Data Providers for purposes other than this Project. In particular, data is to be shared under the Project focuses on closed datasets. This will be processed in a secure environment by Participating SMEs for multiple defined 'challenge' purposes, and the results from the analysed data provided back to Data Providers. This exercise creates new opportunities, but also legal and ethical challenges.

Our duty as the Data Pitch Consortium is to ensure that Data Providers and SMEs participating in the Project meet the same high standards and effective level of legal compliance respected by the organisations that make up the Consortium (the University of Southampton, Beta-I, Dawex, and the UK Open Data Institute (ODI)). To this extent, the Consortium acts as a facilitator so that Data Providers and Participating SMEs are able to access and process data shared under the Project lawfully and ethically.

At the same time, the Consortium recognises that a balance needs to be struck between the use of new data and techniques, and associated risks (such as re-identification risk arising from the processing of data relating to persons shared under the Project). Proposed solutions to the problem of these risks need to be pragmatic, in particular to mitigate the likelihood of any harm befalling stakeholders to a safe level. We have, therefore, accepted that – while legal risk cannot be excluded totally – a Consortium-wide strategy for dealing with EU and EU Member State law is appropriate. The UK is used as an exemplar for this strategy as it is the residing place of two members of the Consortium and the authors of this toolkit. This strategy is designed to cover key areas of concern that might arise around the sharing and reuse of data.

1.2 Purpose of this toolkit

Data can be a complex subject matter in legal and ethical terms. In particular, while data cannot be owned in the same way that, say, a house or car is owned, extensive rights and obligations can arise in relation to data. For example, there are important legal principles about how data can be shared by those with rights in it with third parties who may in turn repurpose that data for different uses.

At the same time, some areas of law which map such rights and obligations are developing rapidly, and are likely to develop even more quickly as big data analytical techniques become more prevalent. These legislative and regulatory developments also point towards greater legal analysis and risk management being required by organisations handling data in the future - including in respect of analysing what rights subsist in relation to data in any given case and managing associated risks.

Although this report applies generally to the legal and regulatory issues raised by data sharing and reuse, it focuses on the legal issues raised under the Project. It emphasises some of the particularities of the challenges associated with the Project and considers the legal issues with these in mind. The aim for this first version of the toolkit is to address the question - "What are the critical things to consider when sharing and reusing data for defined innovative purposes under the Project?" The focus of the toolkit is on ensuring that legal rights in relation to shared data are respected, and legal obligations in respect of such data followed, in the Project. Information is provided on how personal data will be anonymised before it is shared for analysis by Participating SMEs, and ways to mitigate the risk of harm befalling data subjects as a consequence of data relating to them being processed in new ways (for secondary purposes).

2 How to use the toolkit?

2.1 Who is the toolkit for?

This toolkit is designed to provide practical advice for all the non-Consortium organisations involved in the Project (Data Providers, as well as Participating SMEs), as well as potentially being of interest to external persons and organisations looking for guidance on how to share and reuse (such shared) data legally. In particular, it covers legal aspects which are crucial for scenarios in which non-open datasets are to be used in combination in innovative ways, and what has to be considered for this process and to further push forward similar data sharing schemes in Europe.

In particular, privacy and data protection guidance is a key focus in this toolkit to assist Data Providers and Participating SMEs. It will help Data Providers by setting out the legal framework and reflect best practice guidance for anonymising or pseudonymising personal data before it is shared. It will help Participating SMEs in ensuring that they subsequently reuse such data appropriately and lawfully once passed to them under the Project.

Notwithstanding the focus on compliance for the Project, the purpose of this toolkit is also a generic one to provide guidelines on data sharing that raise similar sorts of legal challenges in practice. However, it is not intended to be a substitute for obtaining formal legal advice. Case-by-case assessments of the general principles of law and regulation set out in this toolkit will need to be assessed by the Data Providers and Participating SMEs with the Data Pitch Consortium providing support.

2.2 How to use the toolkit?

This guidance will help such organisations think through a methodology of best practice data handling, and help organisations ask appropriate questions at each stage of the Project as relevant to their role. In particular, this version of the toolkit gives a first outline of what to consider when data is to be shared under the Project, as well as setting out what has already been researched in the legal/ethical domain in the first six months of the Project.

At a practical level it is noted that the legal issues emerging under the Project do not arise for organisations in a vacuum. Large organisations in particular, but also small organisations including start-ups, will typically already have in place risk assessment and management strategies for their data activities. These may range from data protection and privacy governance frameworks to more detailed governance and management structures focused on information architecture, data accuracy, security, and regulatory compliance. It is intended that this toolkit will build on such existing good data governance, rather than replace them.

2.3 How you can feedback into updating this toolkit

The production and evolution of this toolkit will be carried out iteratively during the lifetime of the Project. In particular, it is designed to be iterated as it is used.

Thus, this version 1.0 is shared publicly in the expectation that it will encourage feedback and further improvement. The guidance will then be developed with feedback to provide a plan for an extended plan for the final version of the legal and privacy toolkit for publication in 2018-2019.

The following deadlines are relevant:

- By June 2018: Publication of an evolved data situation model to assess anonymisation practices.
- By December 2019: Publication of an extended and final version of the toolkit.

If organisations or individuals have questions and/or suggestions, please contact us at policytoolkit@datapitch.eu.

2.4 Other parts of the strategy

The toolkit complements other parts of the Data Pitch Consortium's strategy in dealing with data processing issues under the Project. These issues are addressed by:

- Contracts agreed with the organisations formally taking part in the Project;
- Oversight provided by the Consortium to such organisations throughout the lifetime of their involvement in the Project; and,
- Training provided by the Consortium.

2.4.1 Contracts with the Data Pitch Consortium

The legal framework in Europe does not often prohibit secondary data processing outright, rather it limits the conditions of usage. For example, EU data protection law is a framework that explains under what conditions you can process personal data. It is not a ban on the processing of personal data *per se*; rather, the rules dictate how you can process personal data only under certain conditions. These conditions are the privacy principles described in data protection law. Legal agreements designed for Data Providers and Participating SMEs under the Project have been prepared with assurances that these conditions will be followed.

The entering into appropriate signed legal agreements with all parties involved in the Project is important from a risk mitigation perspective. The main mitigation of the risk posed by the sharing of personal data is that personal data is not to be shared unless it has been anonymised or pseudonymised. Notwithstanding, the contracts entered into have to address the possibility of residual risk that personal data may yet be deemed shared.

For example, in addition to ensuing that data protection laws (at EU and national levels) will be followed at all times with respect to the processing of datasets containing personal data under the Project, further assurances are included in the legal agreements to guarantee as far as possible that data protection law will be considered in all cases where data relating to persons is to be processed.

The legal agreements require that Participating SMEs must not attempt to re-identify data that has been modified to hide the identities of data subjects to which it relates. This is a means to mitigate the risk that data subjects will be re-identified. There are also provisions in the Data Providers' contracts setting out requirements around data anonymisation. These are required to ensure that data relating to persons are secured adequately on the Data Pitch 'platform'[1] in light of residual re-identification risk.

More generally, terms and conditions are imposed on Data Providers and Participating SMEs in line with a data use licence that they will all sign with the Project consortium which govern the terms of such use and the assignment of certain rights in relation to the data to be shared.

2.4.2 Oversight by the Data Pitch Consortium

The Data Pitch Consortium can provide oversight to help organisations interpret and instil best practices. Measures carried out by virtue of contractual obligations are meant to be carried out by organisations in a spirit of commitment to strong internal checks to prevent shared data from being used in inappropriate ways. These include the imposition of best practice measures as well as the imposition of strong protection against privacy and other legal concerns.

[1] The term 'platform' is used in a notional sense here. The choice of arrangements by which data may be disclosed and processed under the Project are set out in sub-section 6.1.6 below.

Where Data Providers propose to supply data that has been subject to 'pseudonymisation' processes (as discussed below) for processing by the Participating SMEs, oversight into implementing best practice safeguards can also be recommended by the Project on a case-by-case basis.

2.4.3 Training by the Data Pitch Consortium

Training in support of the toolkit will be provided to organisations to cement their legal and ethical data awareness.

3 Structure

This toolkit sets out an overview of the body of common rights and obligations that are binding in all EU countries, as Members, and more generally can be found in domestic laws, in relation to data management and, in particular, on access to and analysis of combined datasets, including exploitation of the results of such analysis.

Section 4: Section 4 is a snapshot of the current EU legal framework which maps the various categories of EU legislation related to data with respect to the restrictions and requirements that may apply in relation to certain types of processing activities as they apply to different types of data.

Section 5: Section 5 provides a self-standing note on anonymisation, pseudonymisation, and re-identification risk, including how interpretation of these terms and the laws relating to them are currently subject to a changing legal landscape in the EU (including at EU Member State levels) which requires the implementation of a dynamic risk-based approach by organisations that process data relating to persons themselves or share such data with others.

Section 6: Section 6 provides more detailed input into the requirements that those involved in the Project must heed when turning the mapped legal 'theory' into practice effectively. Section 6 outlines (for further evolving in 2018/19):

- a data sharing methodology for those providing datasets to the project, related to managing the legal risks associated with data sharing in practice (**section 6.1**); and,
- a data reuse methodology for those participating in the Data Pitch acceleration programme, related to managing the legal risks associated with data reuse in practice (**section 6.2**).

Section 7: Section 7 is a useful checklist to remind organisations in summary form when data sharing/data reuse is likely to be lawful.

Section 8: Section 8 is a short note that considers the issue of data ethics.

Useful background information is contained in the Annexes.

Future versions

The toolkit sets out what has already been researched in the first six months of the project. Plans are already underway to develop this toolkit iteratively into a longer version, along with publication of an evolved data situation model to assess anonymisation practices during 2018-2019.

The structure of the toolkit is therefore likely to change over time. In particular, as the overall aim is to provide guidance that takes into account the realities of the Project, the specific use case scenarios will be developed to add to the toolkit in future iterations. Several common characteristics will be drawn from these use cases, in particular with a focus on the anonymisation of data relating to persons which will provide examples of how real issues that arose within the Project were addressed.

4 Mapping the legal framework

4.1 Why legal compliance is important

4.1.1 Potential legal risks involved in data sharing and reuse

Data sharing and its reuse for new purposes is crucial for the economy and society to work smarter. At the same time, appropriate safeguards must be put in place to protect privacy. This is fundamental to any data sharing and data reuse regime. Organisations involved in the processing of data must comply with the rules set out in any legislation that might apply to such processing, which in turn requires that they be aware of what rules are relevant and what exactly they require.

Unfortunately, the legal rules surrounding data sharing and reuse can be complex. They can arise from different areas of law, and with respect to different regulatory policies, applying to various categories of data/information. This has led to a cautious approach to data sharing which may go beyond what is required by the law in any one situation. Indeed, confusion around the legalities and risks of data sharing can lead to data simply not being made available to third parties, even though in reality there may be lawful solutions to data sharing. These solutions need to be explored and then an appropriate strategy can be identified.

When considering whether to share data for a specific analytical purpose or purposes, organisations should first consider the rights that arise in relation to such data, as well as the obligations upon them in relation to such data, including any constraints that flow from these obligations. For example, some data (such as highly confidential data) may only be legitimate to share in very narrowly-drawn scenarios. Alternatively, other powers may allow for greater flexibility as long as broad principles are complied with. However, that flexibility will rarely be unlimited where rights in relation to data exist. As a general rule, the assessment of whether an individual act of data-sharing is permissible should be considered on a case-by-case basis.

4.1.2 Consequences of non-compliance with relevant laws

Compliance with relevant laws is very important. A breach of data protection laws can result in enforcement action against the organisations involved by national data protection authorities which in the case of the UK, would be the Information Commissioner's Office (ICO).

The consequences of non-compliance can also be severe, over and above the reputational damage that might arise from involvement in, for example, a data security breach. They can give rise to the imposition of fines and other sanctions, as well as ‘damages’ (financial settlement) being awarded by a court against the parties involved following a court case.

The current developments in EU data protection law from next year sees the introduction of a new regime of significant fines that can be imposed by data protection authorities where data protection laws are broken. These fines, in the cases of the most severe data protection law infringements can be as high as €20 million, or 4% of global annual turnover for the preceding financial year, whichever is greater.

4.1.3 The importance of demonstrating legal compliance

Regulatory authorities will give due weight to in deciding whether there has been a breach of the law to any recorded compliance by organisations with authoritative guidance within the organisation. This can also have an impact on what level of sanctions to impose as a result of a breach of data protection laws. Organisations therefore need accurate record-keeping to prove that legal compliance was at least addressed notwithstanding that a breach still occurred. This is also particularly relevant as many data protection laws require risk assessments to be carried out and any identified risks managed as part of legal/regulatory compliance duties that each organisation bears when it processes data (the so-called ‘accountability’ principle).

It is essential that organisations involved in the Project – including the Consortium members – retain an audit trail of steps taken to demonstrate that different areas of the law relevant to the Project and the data sharing activities have been considered and properly addressed throughout the lifetime of the Project. This is also important as there is a need to review legal analyses over time to ensure that they are still accurate and for any analysis to be revisited if there has been a material change in facts.

4.2 Overview of challenging features that apply to data sharing and reuse

As described in the next section, each type of legal right in relation to data has its own rules, and applying those rules leads to a complex, multi-layered analysis with a level of legal uncertainty. Before considering such rights in turn, it is useful first to examine some terms that are commonplace, in describing different categories of data status typified by different attached rights that can constrain their sharing and reuse by a third party.

4.2.1 Types of data: open data and closed data

Open data is data that anyone can access, use and share. A more specific definition of open data is described by the UK Open Data Institute (ODI) as meeting the following criteria: *"[f]or data to be considered 'open', it must be: 1. Accessible, which usually means published on the web 2. Available in a machine-readable format 3. Have a licence that permits anyone to access, use and share it – commercially and non commercially*".[2]

Closed data, by contrast, is – by definition – data that is not open. The ODI has defined closed data as data that "c*an only be accessed by its subject, owner or holder*".[3]

The Project is specifically designed around the sharing and reuse of closed data, although there may be some use of open data as part of the Project.

4.2.2 Access to data: open access versus restricted access

Data sharing implies some level of data access. Access can include being allowed to see and/or use data, including deriving knowledge from such data on its own or in combination with other data. Under laws that protect rights in relation to data, provision of access to data takes two basic modes (although often a combination of the two is employed):

- Open access to data tends to be provided via posting on publicly-accessible websites, and tends to encompass non-personal data (otherwise, normally, consent to public release normally must be granted). On the other hand, data relating to persons may be released on an open access basis after it has been subject to effective de-identification (anonymisation) techniques.
- Restrictive access to data of one sort or another is used for data that is closed data either partially or wholly. For example, such access may be mediated by contractual terms agreed between the requesters and the data holders (who may be either the original data collectors, or an intermediary).
- The Project is specifically designed around access to data of this second type.

4.2.3 Data spectrum

The above distinctions may more properly be considered – rather than absolutes - as part of a multi-faceted data spectrum. Access to and use of data can be allowed or restricted in different ways between two opposite ends of that spectrum – that is, open data and open access, versus closed data and restricted data with no access whatsoever. In other words, all data sits on a spectrum of closed, shared or open, and can move between each of these positions as decisions are made and the surrounding data environments change (deliberately or accidentally).

[2] For the meaning of this term, and related discussion, see https://theodi.org/blog/closed-shared-open-data-whats-in-a-name.

[3] Ibid.

For example, closed data might be shared with a specific class of people, but they might not be able to use it, only access it. There might also be closed data inside an organisation that is for viewing only, as well as closed data that is used in a range of internal products and services.

4.2.4 Openness versus privacy

While there are significant economic and societal benefits associated with the release of data openly, where the information relates to people, open data can pose risks to their privacy. It could be argued that openness and privacy should be viewed as opposing forces. In fact, they should be seen as complementary and equally important, forming a single inextricably linked, albeit complex, issue. Privacy protection can be compatible with degrees of openness in the majority of cases as long as it is done selectively and subject to adequate controls.

Privacy and data protection laws can thus be seen as laying down a governing framework of rules to control access to, collection and usage of information enabling knowledge and control of data about people summarised as 'granting access but with limitations'. They are underpinned by a policy objective to encourage the free flow of data and its use. In the EU, in particular, there is a policy objective to create a single EU Digital Market with free cross-border trade, in relation to which there is a high level of trust that individuals' fundamental rights related to data are protected through the data protection laws that limit and control the privacy risks.

4.2.5 Privacy versus utility

The application of anonymisation techniques to personal data is intended to strip out or mute (so-called data obfuscation, or perturbation) personally-identifying features from it. However personal features may be crucial to the analysis and use of the data. A tension in this case clearly arises between utility and use of the data and the privacy of the data subject's personal data. The challenge is to take appropriate measures to protect privacy with minimum loss of accuracy. In other words, in an ideal world, the recipients of data that has been subject to an anonymisation (also known as 'de-identification')[4] process should be able to run their analyses on that data without losing accuracy compared with the results of those analyses when run on the original data.

The reason for the anticipated decrease in utility post-anonymisation of data is the fact that – in general - linkability is key to obtain information from the fusion of data within a dataset, or across multiple datasets. Anonymisation techniques may limit data fusion capabilities (and thereby linkability, or at least the accuracy of linking inferences), thereby restricting the range of analyses that can be performed on the data and, consequently, the knowledge that can be generated from it. In particular, data analytics rely upon information being gathered from several independent sources from which unknown patterns can be found, and new inferences made. Hence, the ability to link records that belong to the same (or, similar types of) individual is central to analytics involving data relating to people.

This issue is considered further in section 5 below in relation to considering how EU data protection law determines when data relating to persons is no longer 'personal data', such that is processing no longer has comply with data protection law principles.

4.2.6 The repurposing of personal data

Another distinctive feature of big data analytics - beyond the tendency to collect as much data as possible (to create as much linkability as possible, including from a multitude of different sources) - is the repurposing of

[4] In terms of any potential difference between the two terms, the term 'anonymisation' is commonly used in association with the application of techniques to irreversibly severe the link between a piece of information and identity of the data subject. By contrast, the term 'de-identification' also refers to a severing of data from the identity of the data subject, albeit implied that the modified data may still retain identifying features. In reality, however, we use the two terms synonymously with the latter meaning.

data. This is the use of data for a purpose different from that for which it was originally collected, driven by the fact that existing datasets can be mined to find hidden value.

This is also a key privacy concern. For example, how can individuals know if their data, collected for a specific purpose by a data controller, is being reused and for what purpose? To this end, it is important to have mechanisms to gain data subjects' trust in relation to the secondary processing of data relating to them. Conversely, any undesired reuse of such data would be detrimental to trust levels.

One such mechanism is embedding privacy/data protection 'by design and by default', to mitigate upfront the risks of harm befalling individuals from such secondary use. Transparency on the reuse of personal data is another key factor to ensure fairness, in effect being at least partially 'open by design'.

4.2.7 The need to develop data sharing and reuse protocols

It is important, therefore, to have protocols in place to help bodies understand how to conduct a data sharing arrangement in compliance with the main legal provisions that might apply to them under EU law and their corresponding EU Member State domestic law. In this exercise, aspects such as the extent and timescale of the processing, together with likely impact on individuals should be considered.

For example, the infringement of an individual's privacy from data processing can be significant - potentially including humiliation, financial, or employment status impact, depending on the type of data released and the extent of any identification of individuals. This can happen either as a result of unintentional infringement made without any special effort due to rare characteristics, or as a deliberate attempt to combine various characteristics and datasets.

This toolkit sets out an example protocol with respect to Data Providers, and Participating SMEs, respectively, albeit many other protocols may be suitable. As detailed below, it is focused on a dynamic, risk-based approach in terms of risk assessment and identified risk management. It is also focused on reducing the likelihood of harm befalling individuals flowing from the repurposing of data, as well as the protection of proprietary rights in data.

4.3 Which laws apply?

The laws applicable to data are numerous. They are also scattered across a very large number of legislative instruments, and may be set out expressly or implied. In addition, there is so-called 'common law' developed by judges in courts in countries such as England through case law, contrasting with a system of 'civil law' commonly found in other European countries derived from the interpretation of codified statutes. EU laws, by comparison, come in many different forms. For example, they can be divided into 'primary' and 'secondary' legislation. The treaties (primary legislation) are the basis or ground rules for all EU action. Secondary legislation – which includes regulations, directives and decisions – are derived from the principles and objectives set out in the treaties.

In the following sub-sections, key areas of law relevant to data sharing and reuse are described. These include:[5]

- **Data protection laws** – these set the rules for processing personal data.
- **Electronic privacy laws** – these govern the protection of privacy in the electronic communications sector.
- **Intellectual property laws –** these encompass a number of different rights that may be asserted of a more proprietary type, including in association with the use of data and its expression.
- **Competition laws –** these aim to prevent anti-competitive harm that might result from commercial activities, including from the sharing of information.
- **Laws of confidentiality –** these protect confidential information.
- **Contract laws** – these govern the ways in which private parties can agree to work together, including in respect of data sharing agreements that include certain rights and obligations regarding data usage and access. Ultimately, if terms in agreed contracts are broken, contracting parties could try to enforce such terms in a court of law.

A more detailed version of this section will follow in version 2.0.

4.3.1 EU/national data protection and privacy laws

Driven by societal concerns, regulators and legislators around the world are increasingly concerned about privacy protection, in particular as it relates to issues that affect consumers using services in the online world. Any organisation considering entering into a data sharing arrangement where the data shared relates to persons must consider the relevance of privacy and data protection rules, and ensure that they comply with the law in this respect.

In the EU, the right to personal data protection is a fundamental right, protected by EU law derived from the European Convention on Human Rights (ECHR) and now also enshrined in Article 8 of the Charter of Fundamental Rights of the European Union (the Charter). EU legislation also exists specifically in relation to privacy concerns independently from data protection. Thus, there is a right to respect for one's private life - which must be respected by governmental organisations in EU Member States - which is also enshrined in the ECHR and Article 7 of the Charter.

The protection of such rights in law does not require a prohibition on processing data relating to persons. Instead, conditions – by way of legal principles - are attached to its usage. It is, therefore, important to

[5] Another 'category' of laws not reviewed in this version, but potentially covered in the version 2.0 of the toolkit, are those that mandate the publication of certain types of data which may (or may not) include personal data. See, for example, discussion at: https://theodi.org/blog/what-would-legislation-for-data-infrastructure-look-like.

understand the main conditions, particularly as their application may depend upon a contextual analysis of the facts under consideration. In particular, the adoption of a risk-based approach in ensuring compliance with privacy and data protection laws means organisations must understand what safeguards need to be implemented based on the extent and severity of the risks that intended processing operations would raise to individuals in the circumstances.

4.3.1.1 Data protection laws

Data protection rules applying in the EU are currently derived from the EU Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data (Data Protection Directive [1]). Directives in EU law are binding as to the objective to be achieved but leave implementation to each EU Member State – such as in the UK via the adoption of the Data Protection Act 1998 - leading to the potential for significant differences in national approaches.

However, a new set of EU data protection rules will take effect from 25 May 2018, which provide stronger regulation in this area specifically developed to take into account the vast technology changes since 1995. These new rules are set out in the General Data Protection 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 (GDPR [2]). As the project timeline goes beyond May 2018, the description below includes reference to the higher set of standards to be introduced under the GDPR.

Under both regimes, which apply to the processing of personal data, conditions are set out under which personal data may be processed. These conditions are the principles described in data protection law. To note, while many of the principles in the GDPR are much the same as those in the current Data Protection Directive, nevertheless, there are some important new concepts and obligations introduced by the GDPR, and different actions may be required for compliance with the GDPR. If an organisation is already familiar with the current data protection regime in their EU country in implementation of the Data Protection Directive, then this will provide strong starting point of understanding to build upon for future compliance with the GDPR.[6]

EU Data Protection Definitions

EU data protection law imposes certain standards on the entities (**data controllers**) who alone or jointly with others determines the purposes and means of the processing of personal data. **Data processors**, by comparison, refers to an entity that processes personal data on behalf of the controller.

Personal data is defined as information relating to an identified or identifiable natural (living) person (a **data subject**), that is, information about a person whose identity is either manifestly clear or can at least be established by obtaining additional information. In particular, a person is deemed identifiable if he: "*can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity*." Interpreted broadly, personal data can include: personal details; family and lifestyle details; education and training; employment details; financial details; and, contractual details (e.g. goods and services provided to or by a data subject).[7]

To note, special rules also apply to the processing of special categories of data (also known as '**sensitive personal data**') that reveals: racial or ethnic origin; political opinions; religious and philosophical beliefs; trade union membership; health information; or, details about sexual orientation.

[6] Since the UK voted in June 2016 to leave the EU, it remains unclear how data protection law in the UK will be reshaped in the future. However, as the UK will not have exited the EU by that date, the GDPR will become UK law replacing the Data Protection Act 1998.

[7] More guidance on interpreting the data protection law concept of 'personal data' (as well as other data protection law terms and obligations) is discussed in the next section. Also available are guidelines issued by the EU Article 29 Working Party (Art.29 WP), an influential independent supervisory authority with responsibility for advising on policies and legislation that affect privacy and cooperating with similar authorities to ensure consistent data protection. For example, in 2007, the Art.29 WP adopted an opinion on the concept of personal data (WP136, Opinion 4/2007) that summarises: the "*common understanding of the concept of personal data*" in the EU Member States; the situations in which national data protection legislation should be applied; and, how it should be applied. The opinion analyses the main elements which make up the concept of personal data and adopts a wide interpretation, particularly on the question of when the information' relates to' (is about, even if the information does not focus on him/her) an identified or identifiable individual.

Data protection principles

Data protection law sets out a number of principles with which those organisations that are processing personal data must comply. These principles form the core of the obligations of the data controller and nearly always form the basis of any claim that data protection law has been breached.

The principles oblige data controllers of personal data:

- To process the data fairly and lawfully.
- To collect data only for specified, explicit and legitimate purposes, and not to further process it in any manner incompatible with those purposes (the 'purpose limitation' principle).
- To collect and store data only to the extent that is adequate, relevant and not excessive in relation to the purposes for which it is collected and further processed.
- To ensure that all data held is accurate and, where necessary, kept up to date. Every reasonable step must be taken to ensure that data which is inaccurate or incomplete, having regard to the purposes for which it was collected or for which it is further processed, is erased or rectified.
- Not to keep data in a form which permits identification of data subjects for longer than is necessary for the purposes for which the data was collected or for which it is further processed.

Justifications for personal data processing

Within the EU, a data controller is required to justify the processing of personal data before it will be considered lawful under data protection law. To that end, there are a number of conditions with which data controllers must comply.

The processing of ordinary personal data (as opposed to sensitive personal data) is only lawful if it satisfies one or more of the following:

- The data subject has unambiguously given his or her consent to the processing activity being carried out.[8]
- It is necessary for entering or performing a contract with the data subject.
- It is necessary for compliance with a legal obligation to which the data controller is subject.
- It is necessary to protect the vital interests of the data subject.
- It is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller or in a third party to whom the data are disclosed.
- It is necessary for the purposes of legitimate interests pursued by the data controller, except where these interests are overridden by the interests for the fundamental rights and freedoms of the data subject (the so-called 'legitimate interest' legal basis).

[8] Article 4(11) GDPR defines consent as meaning a "*freely given, specific, informed and unambiguous indication of the data subject's wishes" signifying agreement to the processing of personal data relating to him or her"*. Article 7 GDPR sets out the conditions for valid consent and places the burden of proof on the controller. These conditions include, for example, the fact that the consent must be clearly distinguishable from any other matters in a written document and it must be "*in an intelligible and easily accessible form, using clear and plain language*." In addition, where the processing has multiple purposes, consent should be given for all of the purposes (Recital 32, GDPR). Consent is presumed not to be freely given if it does not allow separate consent to be given to different personal data processing operations (Recital 43, GDPR).

For sensitive personal data, the data controller must also comply with one of the additional requirements, for example, where:

- The data subject has given ‘explicit’ consent.
- In processing the data, the data controller exercises a legal right or performs a legal obligation under employment law.
- The processing is necessary to protect the vital interests of the data subject or of another person where the data subject is physically or legally incapable of giving his consent.
- The processing is carried out by non-profit making organisations (for example, charities, churches and trade unions) for certain specified purposes.
- The processing relates to data which are "manifestly made public" by the data subject or is necessary for the establishment, exercise or defence of legal claims.

Secondary processing and compatibility of purpose (the purpose limitation principle and data minimisation)

As previously described, this relates to the repurposing of data – that is, the intended processing of data for a different purpose than that for which it was originally obtained. Where personal data are intended for re-use, the application of data protection in these scenarios is particularly important.

Some of the key decisions an organisation must make under EU data protection law include:

- whether any secondary use of data complies with the principle of ‘purpose limitation’ – in particular, whether data initially used in one context can be considered adequate, relevant, and proportionate to be reused in another context, and,
- whether, in the absence of obtaining consent from the individuals to reuse personal data relating to them for secondary purposes, an organisation can rely on any legal basis to justifying their processing of the data (including, in particular, the necessity-based ‘legitimate interest’ legal basis).

Regarding the first point, the purpose of secondary data use must be clearly and specifically identified. In particular, it must be detailed enough to allow the data subject to determine what kind of processing is, and is not, included within the specified purpose. Consequently, data controllers should ensure that the purposes they notify to data subjects on the collection of their personal data are not too vague or general as by doing so they may fail to meet the criteria of being specific.

As an example for the second point, an organisation that has collected personal data for one purpose legitimately, must - before it starts analysing such data for a different purpose – get consent from that data subject for this secondary processing usage. Alternatively, the organisation must find another legal basis to justify that secondary processing, unless the secondary processing purpose is compatible with the original justification (the initial legal basis) for processing the data upon its initial collection.

Under data protection law, to determine whether a new purpose is compatible with the original purpose, guidance has been provided by the Article 29 Working Party (Art.29 WP) in 2013 in its Opinion on Purpose Limitation (‘WP203’ [3]) setting out a detailed approach to help data controllers carry out a compatibility assessment. In particular, the Art.29 WP opines that a number of key factors should be considered to assess the purpose limitation as follows:

- **The relationship between the purposes for which the data was originally collected and the purpose of secondary processing.** The focus of this assessment should be the substance of the relationship. Questions to be asked include whether the purpose of the secondary processing was already more or less implied in the original purpose, or is there only a partial or non-existent link between the two purposes. The greater the ‘gap’ between the two purposes, the more likely it is that the further processing is incompatible with the original purpose.
- **The context in which the data has been collected and the reasonable expectations of the data**

subject about its further use. The data controller should consider what a reasonable person in the data subject's situation would expect his data to be used for, based on the context of the collection. This requires consideration of any legal statements made, as well as what would be customary and generally accepted practice in the given context. The more unexpected or surprising the further use is, the more likely it is that the further processing is incompatible with the original purpose. In addition, an assessment of the nature of the specific relationship between the data controller and the data subject in context should include an investigation of the balance of power between them. In cases where the data subject was obliged legally to provide their data, or was in a weak commercial bargaining position, this could give rise to reasonable expectations of stricter confidentiality and stricter limitations on future use.

- **The nature of the data.** The nature of the data will determine the level of protection that the data controller must afford to the data subject. Further processing that involves sensitive personal data is likely to require stricter limitations on further use.
- **The impact of the further processing on the data subject**. The data controller should take into account both negative and positive likely consequences of secondary processing on data subjects. The former could include emotional impacts such as negative feelings related to a loss of control over personal data because the data is shared widely.
- **The safeguards applied by the data controller to ensure fair processing and to prevent any undue impact on the data subject**. Appropriate additional safeguards taken by data controllers may occasionally compensate for a change of purpose, or the fact that the purpose has not been as clearly specified as it should have been. Those measures may include technical or organisational measures to ensure functional separation of different processing activities, as well as additional steps taken for the benefit of the data subject like increased transparency.

By way of illustration, WP203 gives the example of repurposing data for big data analytics. Art.29 WP identifies two types of further processing in this respect:

- where it is done to find out about individuals and make decisions affecting them; and,
- where it is done to detect trends or correlations.

In the first case, the Art.29 WP says that "*free, specific, informed and unambiguous 'opt-in' consent would almost always be required, otherwise further use cannot be considered compatible*". An example is where information that people have put on social media is going to be used in unexpected ways, such as to assess their health risks or their credit worthiness. Unless they are informed of this and asked to give their specific consent to this secondary use, it is unlikely to be compatible (or fair) personal data processing.

In the second case, the law is more lenient. In other words, the organisation concerned will not need to seek specific consent for intended further processing of personal data where the activity would not involve making decisions about the individual data subjects. However, if data controllers intend to rely on a legal basis other than consent, they should still assess whether the new purpose is compatible with the original reason for processing the data in accordance with the compatibility assessment above. In other words, technical and organisational safeguards should be applied by the data controller to prevent any due impact on the data subject. Furthermore, the Art.29 WP advocates a clear functional separation between analytics operations that detect trends and those that involve making decisions, as well as the carrying out of pseudonymisation processes (for more on pseudonymisation, see below) on the personal data.

Where attempting to rely on the 'legitimate interest' legal basis to justify secondary processing, similar considerations are required under the GDPR as under the Art.29 WP's recommended compatibility assessment. In reminder, assessing whether this legal basis can be relied on requires analysis of whether the processing is necessary for the purposes of the legitimate interests pursued by the controller (or a third party), and are not

overridden by the interests or fundamental rights and freedoms of the data subject which require personal data protection. The purpose of this balancing exercise is to prevent *disproportionate* impact on individuals. In practice, carrying out this exercise will also require a full assessment of the facts and context of each case are required. Consideration of the reasonable expectations of the data subject individual based on their underlying relationship with the controller is also required. As mentioned, in cases where the data subject was obliged legally to provide their data, or was in a weak commercial bargaining position, this could give rise to reasonable expectations of stricter confidentiality and stricter limitations on future use.

Moreover, in both cases (compatibility assessments and legitimate interest assessments) organisations need to carefully consider how they document their analyses.

Exemptions from data protection principles

Scientific research exemptions

Recital 33 of the GDPR acknowledges that it is often not possible to fully identify the specific purpose for which personal data may be further processed – at least at the time of initial collection - when it is to be carried out in the public interest, for scientific or historical research purposes, or for statistical purposes (hereafter, collectively 'scientific research purposes').[9]

As such, when they are being carried out for scientific research purposes, data subject consent is not required for secondary data processing activities. Moreover, such activities would not have to be deemed compatible with the initial purposes where the secondary processing legal basis to be relied upon is not consent;[10] the justification for this is partially because such secondary processing would be presumed to be carried out in keeping with ethical standards for scientific research in any event. However, to ensure this takes place, the GDPR requires that appropriate safeguards ("*for the rights and freedoms of the data subject*") must be implemented. Moreover, research should be carried out on data that has been subject to anonymisation techniques, unless the scientific research purposes cannot be otherwise fulfilled. These safeguards include:

- ensuring that technical and organisational measures are in place (e.g. such as keeping data enabling the attribution of information to an identified or identifiable data subject separate from other information); and,
- ensuring that the principle of 'data minimisation' (only the minimum amount of personal data should be processed to achieve the processing purpose) is upheld.

Which measures are appropriate to adopt will depend on the context, the nature of the data, and the assessed impact of further processing on the data subjects.

Where personal data are processed for research purposes, derogation from data subject rights (as described under the next sub-heading) are possible, but only in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes.

Processing which does not require identification

Data controllers may modify personal data in such ways that prevent a data subject from being re-identified from it. Article 11 GDPR acknowledges this situation and provides an exemption from data subject rights to access, rectification, erasure, and data portability (described below). The exemption applies only if "the

[9] To note, **t**he GDPR adopts a broad definition of research, encompassing the activities of public and private entities alike (at Recital 159 -*"technological development and demonstration, fundamental research, applied research, privately funded research"*).

[10] Recital 50 of the GDPR states that *"further processing for [scientific research purposes] should be considered to be compatible lawful processing operations"*. This is an important provision where the limited scope of limited consent forms would not make such secondary research possible.

controller is able to demonstrate that it is not in a position to identify the data subject". To note, the GDPR does not require a controller to hold additional information "*for the sole purpose of complying with this Regulation*."

Data subject rights

The upholding of data subject rights (rights that can be exercised by the individuals vis-à-vis data controllers who process data relating to them), including some new types of rights, is a key focus of the GDPR. Underpinning these rights are the concepts of fairness and transparency towards data subjects and these are critical in ensuring compliance with data protection laws.

Data subjects are granted the following specific rights:

- **The right to be provided with certain information from the data controller** – this should occur – either where personal data are collected directly from the data subject, or indirectly from a third party - at the time when the data is first processed or as soon as possible afterwards. This should include details about: the identity of the controller; the intended purposes of the processing; and, any further information necessary to render the processing fair, such as recipients to whom data are disclosed, and the existence of right to access to, and rectification of, the data concerning the data subject.
- **The right of access to personal data held by the data controller relating to the data subject** - a data subject has a right, on making a request to the data controller, to be informed whether personal data of which he/she is the data subject is being processed by or on behalf of that data controller. If so, the data subject also has a right to: a description of the personal data undergoing processing and the purposes for which it is being processed and the recipients or classes of recipients to whom the data may be disclosed; and, subject certain exceptions such as in relation to confidentiality, any information available to the data controller as to the source of the data.
- **The right to have data corrected, erased or blocked** – where it is not being processed in accordance with data protection principles.
- **The right in certain circumstances to object to processing of data** - such as the processing of data for direct marketing purposes.
- **The right not to be subject to a decision based solely on the automated processing of data** – this applies where such automated processing is intended to evaluate certain personal matters relating to the data subject, such as his/her performance at work, creditworthiness, reliability, and conduct.
- **The right to receive compensation from a data controller** – this applies where damage has been suffered as a result of unlawful processing of personal data.
- **The right to erasure (also known as the 'right to be forgotten')** – under the GDPR, individuals have the right to request that businesses delete their personal data in certain circumstances (for example, where the data subject withdraws their consent, as well as where data are no longer necessary for the purposes for which they were collected). Controllers are also required to take all reasonable steps to inform third parties that any links to the copy or replication of personal data must be deleted, on request from the data subject.[11]

[11] To note, the GDPR provides an exemption from the right of erasure of personal data for scientific research purposes, in so far as the right of erasure is likely to render impossible or simply impair the achievement of objectives of this type of processing. Notwithstanding, Recital 65 of the GDPR states that, where data is processed for scientific research purposes, the further retention of such personal data should be lawful; in practice, therefore, the consequences of the right of erasure is limited in relation to scientific research.

Data security

A key data protection principle relates to data security, in keeping with the adage that '*data privacy cannot exist without data security*". Under data protection rules, both data controllers and data processors must implement "*appropriate technical and organisational measures*" to protect personal data – in effect, by ensuring a level of security appropriate to associated risks of data breach in each case. These include risks of: accidental or unlawful destruction, accidental loss, alteration, unauthorised disclosure or access, and all other unlawful forms of processing.

As well as being responsible for ensuring adequate security of personal data, data controllers (and to a lesser extent, data processors) are also responsible in the case of a data breach. This implies that data controllers need to vet thoroughly those who are formally granted access to personal data to process that personal data on its behalf. Data processors also need to be careful when securing personal data on behalf of others, and should ensure that they follow all instructions issued around data security by data controllers.

The adequacy of measures taken to ensure personal data security should take into account the state-of-the-art technologies available, the costs of implementation, and the nature, scope, context and purposes of processing, as well as the risk of varying likelihood and severity "*for the rights and freedoms of natural persons*". Examples of measures that might be appropriate could include:

- applying pseudonymisation process to personal data (for more on pseudonymisation, see below);
- data encryption;
- the ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and services;
- the ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident; and,
- implementing processes for the regularly testing, assessment, and evaluation of the effectiveness of technical and organisational measures for ensuring the security of the processing.

To note, the GDPR requires businesses to notify the national supervisory authority of all data breaches without undue delay and where feasible within 72 hours unless the data breach is unlikely to result in a risk to the individuals. If this is not possible it will have to justify the delay to the authority by way of a "*reasoned justificatio*n". The requirements for a reasoned justification are not clear prior to the implementation of the GDPR but it can be anticipated that a high level of justification would be required to explain any delay.

Data protection by design and default

As indicated, the GDPR requires businesses to implement technical and organisational measures to facilitate compliance with data protection principles even before they start processing personal data. This requirement is introduced as a new data protection principle under the GDPR: 'data protection by design and default'.

An example of a measures which may satisfy the requirement to implement data protection by design and default is pseudonymisation. Pseudonymisation refers to a process applied to personal data, which is defined as a new legal concept introduced by the GDPR. In particular, it means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific individual without additional information kept separately to prevent inadvertent re-identification.

Pseudonymisation is referred to further in the next section on anonymisation. However, to note, the processing of personal data which has been pseudonymised is actively encouraged under the GDPR as a means of implementing appropriate safeguards to protect personal data. For example, Article 6(4) of the GDPR states that the pseudonymisation of data is a factor controllers should consider when determining whether secondary processing of data has a purpose which is compatible with the original purpose when such data was initially processed.

Data protection impact assessments

A data protection impact assessment (DPIA) is a tool to be used before processing begins to assess how planned processing activities are likely to affect the individuals whose data is being processed and whether processing is fair. Under the GDPR, data controllers are required to perform DPIAs before carrying out any processing that is likely to result in a "high risk" to data subjects (taking into account the nature, scope, context and purposes of the processing). In particular, DPIAs are required for:

- A systematic and extensive evaluation of personal aspects by automated processing, including profiling (see below, including most forms of online tracking and behavioural advertising), and automated processing carried out on which decisions are based that produce legal effects concerning the data subject or significantly affecting the data subject.
- Processing of special categories of personal data or data relating to criminal convictions and offences on a large scale.
- A systematic monitoring of a publicly accessible area on a large scale.

The GDPR does not specify which DPIA process must be followed but instead allows for data controllers to introduce a framework which complements their existing working practices provided it takes account of certain components.

Guidance on these components can be found in the Art.29 WP's 2017 Guidelines on Data Protection Impact Assessment and determining whether processing is "likely to result in a high risk" for the purposes of the GDPR ('WP248' [4], summarised in **Annex A** below).

The results of DPIAs should then be taken into account with an ongoing requirement to keep those measures up-to-date. A clear record of personal data processing activities, the main elements of DPIAs carried out, and compliance measures taken in consequence of results, should also be kept by controllers and processors.

Profiling

Profiling is defined in the GDPR as meaning, "a*ny form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements"* (GDPR, Article 4(4)).

The fact of profiling must be disclosed to the data subject who may object to it. Under Article 14(2) GDPR, the information to be disclosed should include the fact that such activity is to take place, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject. Moreover, a DPIA is required to be carried out before profiling activities take place to ensure that risks of harm to data subjects that could ensue are identified and mitigated appropriately.

Under the GDPR, data subjects also have a right not be subject to profiling (or other forms of automated processing) which produces legal effects on, or significantly affects an individual, unless the profiling is necessary for the performance of a contract, it has been authorised by Member State law, or it is conducted with the explicit consent of the individual.

Decisions based solely on the profiling of sensitive personal data are only permitted in very limited circumstances (that is, with the explicit consent of the data subject, or where the profiling is necessary for reasons of substantial public interest) and where appropriate safeguards are implemented.

4.3.1.2 Electronic Privacy Laws

Privacy laws independent from data protection regulation also exists under EU law, notably the E-Privacy Directive (formally, Council Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector, as amended in 2009 by the Citizens' Rights Directive 2009/136/EC [5]). This Directive's minimum standards have been implemented into the national laws of EU Member States.[12]

The E-Privacy Directive is currently under review as part of the European Commission's agenda to reinforce trust and security in the Digital Single Market. Several policy issues have emerged from this review as potentially needing to be addressed. These include ensuring consistency of the e-privacy rules governing electronic communications services with the GDPR; updating the scope of the E-Privacy Directive in the light of new market and technological reality (such as instant messaging, voice over IP, and web-based email); enhancing security and confidentiality of communications; and, addressing inconsistent enforcement and fragmentation of legal rules and their interpretation at national level.

To these ends, in January 2017 the European Commission published a draft E-Privacy Regulation (COM(2017) 10 final [6]) which is intended to replace the current E–Privacy Directive from 25 May 2018 alongside the introduction of the GDPR.

The draft Regulation would be directly applicable, like the GDPR, meaning that organisations across the EU would only need to comply with one set of e-privacy rules, rather than varying rules in different EU Member States.

Key provisions in the draft Regulation include in respect of proposed Articles 6 and 7 permitting processing and storage or erasure of communications data. Data such as the timing, location and duration of a call ('metadata') and user browsing history will need to be anonymised or deleted if users have not given their consent to its retention, unless the data is required for certain purposes. The new proposed rules expand the possibilities of how telecoms providers may use this type of data.

For example, existing rules only permit telecoms providers to process traffic data and location data for value-added services (such as suggesting products better suited to individuals' needs based on their usage, or offering Wi-Fi hotspots, or for billing purposes for verification purposes).[13] The proposed Regulation provides for other purposes, if users have provided consent, provided that the business complies with certain privacy safeguards.

The impact of this draft Regulation will need to be considered if communications data is used within the Project.

[12] For example, in the UK the Privacy and Electronic Communications (EC Directive) Regulations 2003 (SI 2003/2426) as amended by the Privacy and Electronic Communications (EC Directive) (Amendment) Regulations 2011 (SI 2011/1208) (E-Privacy Regulations). See, for example, the ICO's Guide to Privacy and Electronic Communications Regulation, available at https://ico.org.uk/for-organisations/guide-to-pecr/.

[13] Traffic data includes any data that identifies the person transmitting the communication, the person to whom it is transmitted and the circumstances under which it is transmitted. In the case of an email, for example, related traffic data may include the time the email was sent, the sender and the addressee as well as the size of the file. In the case of a telephone call, it may include the number called and the number from which the call was made as well as the call length. Location data is a sub-form of traffic data and includes the geographical coordinates (cell-site information) used by telecommunications providers to track the location of portable communications device, such as mobile phones.

4.3.2 EU/national intellectual property laws

Data/information can attract intellectual property rights (IPR). However, analysing these rights in data often requires a multi-layered legal assessment as requirements for different types of IPR to exist – and related rules - can vary widely. For example, they may provide copyright protection protecting a particular expression of information (such as a literary work); have the attributes of a database to attract database rights protection (and, also, database copyright); and, in some circumstances, could give rise to trade mark or so-called 'passing off' rights.

To note, as well as different types of rights, another distinction across the spectrum of IPR types relates to legal enforceability of the IPR (including potentially powerful infringement remedies, from temporary and permanent injunctions to damages and account of profits). Some IPR are in theory enforceable against the world; however, others are primarily national rights that operate different in different countries, and which are enforceable in one country (though the courts of that country) but not others. This includes disparities within the EU bloc, as well as outside it.[14]

For these reasons, contractual agreement is often relied upon to clarify and confer strong, enforceable rights at least between the contracting parties, including by assigning rights of use from one party to another via licence. Contracts are considered further below.

What follows is a brief analysis of some of the key different IPR types in relation to data/information.[15]

4.3.2.1 Copyright

Copyright protects data/information in its expressed form, not the underlying idea. For example, simple literary work copyright in a webpage will subsist in the technical, functional and user specifications and so on, as well as related documentation such as written statements or descriptions. Copyright can also apply to software, certain databases, literary works, music, and films.

Copyright typically arises automatically (so no registration is required) when certain conditions are satisfied. As the name suggests, copyright typically protects an original work against unauthorised copying, and the unauthorised carrying out of other acts protected under copyright law.[16] Notwithstanding, copyright is capable of assignment or licensing that can be scoped in relation to restricted acts, by time, by geography and so on. Future copyright may also be assigned, although the duration of copyright protection is not infinite.[17]

For copyright to subsist in data/information, there must be originality in the work (albeit, in practice, the degree of originality required is typically fairly low).[18] Moreover, there should be a degree of skill and labour in producing the work.

[14] For example, some countries operate a copyright registration requirement, while in others copyright arises automatically by operation of law. At an international level, another example is the fact that the EU-wide 'database right' (as described below) does not apply to databases made in the US.

[15] To note, trademarks are not discussed below as they can apply to data products (like indices), but, generally, not in relation to the actual data. Similarly, patents and rights to inventions are not covered because - while they can apply to software and business processes that manipulate and process data - they do not arise in relation to data itself.

[16] This is not true for entrepreneurial works however. For example, under UK law, there is no requirement of originally in relation to copyright on sound recordings.

[17] Under UK law, for example, it is the life of the author plus seventy years in the case of software, databases and other literary works.

[18] For example, in Infopaq International A/S v Danske Dagblades Forening (Case C-5/08) (Infopaq I), the Court of Justice of the EU (CJEU) held that the storing and printing out of 11-word extracts from newspaper articles would amount to reproduction of a copyright work under the EU Copyright Directive 2001/29/EC if the elements reproduced were the expression of the author's

Despite the existence of copyright (and any contractual rights assigned related thereto), there are a number of permitted acts that can typically be carried out in relation to works to which copyright law applies. For example, under UK law (the Copyright Designs and Patents Act 1988 (CDPA)) various acts are permitted to be carried out in relation to certain types of copyright works, including the following key exceptions:

- **Temporary copying** – copyright will not be infringed by making a temporary copy which is transient or incidental assuming: it is an integral and essential part of a technological process; it has the sole purpose of enabling a transmission of the work in a network between third parties by an intermediary, or a lawful use of the work; and, it has no independent economic significance.
- **Text and data mining** – this applies, for example, in the UK to permit text and data analysis for non-commercial research of copyright works to which a person already has lawful access. Such automated techniques work by bulk copying electronic information, which is then analysed for patterns, trends, and other information.[19]
- **Fair dealing defences** – copyright will not be infringed where copyright works are used for the purposes of: research and private study; criticism or review; reporting current events; quotation; or, parody, caricature, pastiche.
- **Anonymous or pseudonymous works** – copyright will not be infringed by an act either carried out when it is not possible by reasonable enquiry to identify the author, and also reasonable to assume that the copyright has expired or that the author died over 70 years ago.
- **Databases –** copyright will not be infringed if a licensed user of a database carries out acts in the exercise of that right which are necessary to gain access to, and use of, the contents of the database.

Finally, to note, copyright in works created by an employee (in an employment context) are owned automatically by the employer. However, when created by a contractor for an organisation, they will need to be assigned in writing for copyright to be owned by the engaging company.

4.3.2.2 Databases

EU Council Directive 96/9/EC on the legal protection of databases (the Database Directive [7]) introduced a legal definition of a database: essentially, a searchable collection of systematically or methodically arranged works, data or other material.

The Database Directive – subsequently implemented into the national laws of EU Member States – resulted in databases being included in the list of literary works in which copyright subsists.[20] The standard of originality required for database copyright is that the author's 'own intellectual creation' had gone into selecting or arranging the database's contents.

The Database Directive also introduced a new and unique 'database right' subsisting in any database "*if there has been a substantial investment in the obtaining, verifying or presentation of the contents of that database"* where, at the time it was made, the maker was EU-based. Like copyright, therefore, the database right is a formal remedy capable of protecting the way in which information is displayed, rather than protecting investment relating to the creation of the data. The key features of this right include:

intellectual creation. There is some dispute about whether this has altered the originality test applied by, for example, UK courts that the work concerned must not have been copied from elsewhere.

[19] The UK Research and Private Study Regulations 2014 introduced this exemption via a new section (29A) in CDPA with effect from 1 June 2014.

[20] For an example of the application of the Database Directive in an open data context, see https://theodi.org/case-studies/open-addresses-the-story-to-date.

- **Ownership** - the first owner of a database right is the maker, who is the person who takes the initiative in and assumes the risk of obtaining, verifying, or presenting its contents.
- **Duration -** the first generation of a database right lasts for fifteen years from the end of the year when the database was completed, which is effectively refreshed wherever "*any substantial change*" is made.
- **Infringement** – a database right is infringed if a person without the owner's consent "*extracts or re-utilises all or a substantial part of the contents of the database*", or carries out "*repeated and systematic extraction or re-utilisation of insubstantial parts of the contents of a database*". Extraction is defined as the "*permanent or temporary transfer [of the contents] to another medium by any means or in any form" and re-utilisation is defined as "making … available to the public by any means*".

4.3.2.3 EU reform proposals

The law in the area of IPR in relation to data/information will continue to develop in the coming years, particularly in light of the increasing popularity of big data analytics and text mining.

In the EU, the European Commission's Digital Single Market initiative is aimed at removing barriers to online cross-border trade within the bloc, and ultimately at aligning conditions for full alignment of copyright rules across the EU in the form of a single copyright code and copyright entitlement. As part of this initiative, in 2016 the Commission published a proposal for a Directive on Copyright in the Digital Single Market (the draft Copyright Directive [8]). If adopted, this Directive includes a number of proposed EU copyright reforms, intended, according to the Commission, to promote a "*fair, efficient and competitive European copyright based economy*".

Among the proposed measures are mandatory exceptions to copyright for text and data mining, digital and cross-border teaching, and the preservation of cultural heritage, as well as the creation of a structure for the licensing of "*out of commerce*" works. Amongst other things, this would allow EU organisations to process on a large scale copyright works to which they have legal access.

Another Commission initiative published in 2017 was a consultation document (and related papers) asking for feedback from interested stakeholders on a possible future EU framework for facilitating the free flow of non-personal data across Member States, and inviting comments on possible ways to help achieve certain objectives in this respect.[21] One of the ways proposed by the Commission was a new IPR-type right in non-personal, machine-generated (including machine-to-machine, 'M2M') data aimed at improve access, as well as facilitating and incentivising the sharing of such data, in ways that tackle restrictions on its free movement for reasons other than the protection of personal data. Such a right to licence the use of data collected by sensor-equipped machines, tools, or devices was proposed to be awarded exclusively to device manufacturers, or data producers, (or a shared right between them), for licensing to any party they wish. Alternatively, the Commission suggested that the issue of licensing decisions could be left solely to the parties involved as a matter of contractual negotiations. The Commission also asked for views about a possible new obligation to license the reuse of such data under fair, reasonable and non-discriminatory ('FRAND') terms, to facilitate more access to such data with remuneration after anonymisation.[22]

[21] Public consultation on Building the European Data Economy, see: https://ec.europa.eu/digital-single-market/en/news/public-consultation-building-european-data-economy.

[22] In addition, the Commission refers to other options that could be taken forward including developing new guidelines to incentivise businesses to share any non-personal data they have, and granting special rights of access to data to public bodies where this is in the general interest. New default contracts rules are also proposed that could be set to facilitate access to data in accordance with benchmarks that account for the different bargaining positions that businesses in the market have.

4.3.3 EU/national competition laws

National and EU competition authorities have shown increasing interest this decade in analysing data-centric business practices - including mergers and licenses - through the lens of competition law. Sectors investigated include electronic communications and particularly financial markets.

Relevant to the issues that might arise under the Project, EU competition law – implemented into EU Member States' domestic laws - include certain restrictions on information sharing implemented through agreements, or so-called 'concerted practices', between at least two organisations that have the object or effect of restricting competition in horizontal/vertical markets.

For example, in the UK, the Competition and Markets Authority investigating an information sharing practice, would have to establish – on the balance of probabilities - that:

- The parties entered into an agreement or engaged in a decision or concerted practice.
- Which may affect trade within the UK (or part of the UK).
- Which has as its object or effect the restriction, prevention or distortion of competition within the UK.

In broad terms, whether or not information exchange between organisations is anti-competitive will depend on whether it reduces the strategic uncertainty of competitors, consequently diminishing their incentives to compete against one another. In practice, examples of restrictions of competition by 'object' are information exchanges between competitors of individualised data regarding intended future prices or quantities (including intended future sales, market shares, territories or customer lists).

By comparison, restrictions of competition by 'effect' would include information exchanges between any two types of organisations likely to have an appreciable adverse effect on one or more aspects of competition, such as price, output, product quality, product variety or innovation. The likely effects of an information exchange on competition must be analysed on a case-by-case basis, in particular taking into account the economic conditions in the relevant markets and the characteristics of the information exchange arrangement.

4.3.4 Sector-specific regulation

Data regulation is also deepening in many vertical industry sectors. This is not necessarily a novel development; for example, the rules on the confidentiality of client information and privilege have been cornerstones of the legal profession for generations. However, the digitisation of data is changing the regulatory landscape fundamentally in some sectors.

Examples include the financial sector, insurance, air travel (specifically, rules on passenger name record ('PNR') data about an airline customer's itinerary), and healthcare (including rules about aggregating anonymised clinical outcome patient data).

These sector-specific requirements are tending to become more intrusive as regulatory authorities obtain wider supervisory powers to obtain information, investigate business practices and conduct, and audit organisations under their charge.

4.3.5 Other relevant private laws applying at a national level

Data with a quality of confidentiality can also attract legal protection and remedies that can be enforced when confidential information is shared without authority in national courts. In other words, rules governing the confidentiality of information exist in some countries that protect the substance of data that is not generally publicly known.

For example, there is UK case-law suggesting that a right to confidentiality can exist in respect of dataset aggregation despite some of the data not being confidential, alongside legal protection extending to second

and subsequent generation data derived from initially confidential data.[23] Nevertheless, consideration should be given to whether contracts - and websites and other notices - state expressly that data should be considered confidential, and that it is not freely publicly-available. Furthermore, historic data inevitably holds less value and is likely to be more widely disseminated than real-time data.

4.3.6 Contract laws

It is possible to impose explicit obligations and confer rights relating to data between contracting parties where data is intended to be shared between them, including contractual obligation that can be designed to mimic the enforceable rights and obligations associated with different types of IPR. However, contractual provisions are only enforceable between contracting parties.

For data suppliers, contracts should contain express acknowledgements to the effect that relevant rights subsist in the data and are owned by them. Therefore data recipients should undertake to take, allow, or suffer no act inconsistent with rights of the data supplier under the agreement.

Other, common issues in data sharing and reuse legal agreements include:

- **Scope of rights in data being agreed.** Details should be provided in the contract on what data is relevant, plus what rights in relation to that data are being conferred (the terms of use). For example: exclusivity or non-exclusivity; restrictions on how the rights may be used; geographical restrictions on usage; and, the duration of usage.
- **Warranties of compliance with laws and regulation, and indemnities in case of later non-compliance.** These include in respect of data protection; sector specific regulation; and, audit/investigation.
- **Treatment of derived and comingled data.** Questions can arise about the extent to which a recipient of shared data to create derived data from it, and then who owns what rights in the derived data (particularly whether the supplier of the input data has a property interest in the derived data). Commingling data is like deriving data, but with the user taking input data from more than one supplier and creating something different with the commingled data.
- **Post-term use of data.** Another issue is what happens on termination of the contract as regards post-term use. For example, after termination of the agreement, does the data recipient have to delete the data immediately from its systems?

[23] Albert (Prince) v Strange, ([1849] 1 M&G 25); Exchange Telegraph Co. Ltd v Gregory & Co., ([1896] 1 QB 147); Exchange Telegraph Co. Ltd v Central News Ltd ([1897] 2 Ch 48); Weatherby & Sons v International Horse Agency and Exchange Ltd, ([1910] 2 Ch 297).

5 Anonymisation, pseudonymisation, re-identification risk – a changing legal landscape and outline strategy under the Project

This section contains an introductory note to the concept of data anonymisation as it is construed under EU data protection law. It also aims to highlight, in general terms, some of the key risks around the adequacy of the application of anonymisation techniques that are likely to arise under the Project, in the context of considering the legal status of shared data relating to persons.

Current EU data protection law is governed by the Data Protection Directive. Notwithstanding, national laws that implemented the Data Protection Directive (and national data protection authorities, and courts that interpret these rues) have differing approaches to anonymisation, and also the criteria for determining whether data are truly anonymised in a legal sense such that their processing falls outside the scope of data protection law. Compliance with these divergent guidelines is often difficult, in particular for organisations that process data relating to persons in multiple EU Member States. For example, in Belgium and Sweden, 'key-coded' data (where direct identifiers in a dataset – such as name – are replaced by pseudonym 'codes' per subject) are considered personal data if a third party has a key that can be used to re-identify the data subject. In contrast, the UK takes the view that where data recipients are bound by confidentiality obligations, and restrictions on reuse and re-identification, then the risk of the data subject being re-identified from data should be considered low. In such circumstances, key-coded data is less likely to be considered personal data in the UK.

All of these laws will cease to apply from 25 May 2018 when the new GDPR becomes directly applicable in all EU Member States, and it is the higher standards set by the GDPR (rather than the Directive's standards) that are applied in the Project. The impact of the incoming GDPR - in addition to related EU-wide guidelines and best practice recommendations on this topic (including its introduction of the new legal definition of 'pseudonymisation') - are particularly important as they are aimed at unifying the current disparate approaches. A summary of the legal concept of anonymisation is set out below. However, the latest guidelines on these issues will be kept under review and new versions of this toolkit will update these sections in light of material changes.

5.1 The legal concept of anonymisation

The term 'anonymisation' is seen as a term of art under data protection law. It means neither 'nameless' nor completely unidentifiable. Rather, according to Recital 26 of the Data Protective Directive, determining whether a data subject is non-identifiable from data should take into account "*all the means likely reasonably to be used*" by the data controller or by any other person to identify them. Data protection rules do not apply to the processing of data rendered anonymous in this sense (such that the relevant information may no longer be considered to relate to an "*identified or identifiable natural person*").[24]

Therefore, under existing law so interpreted, absolute anonymity is not required for personal data to fall outside data protection rules.[25] Instead, a re-identification risk-based approach is needed to assess whether – taking into account the circumstances of the data environment under consideration – the Recital 26 'means test'

[24] The Directive, Article 2(a)), defines "personal data" as "*any information relating to an identified or identifiable natural person*".

[25] Under national EU Member State laws implementing the Directive, this statement is also often true. For example, as the Project is being led by UK operations, we have also considered the concept of anonymisation under the UK Data Protection Act 1998 (DPA), which implements the Data Protection Directive. Under the DPA, data constitutes personal data if a living individual can be identified *either* from: the information alone; or, with other information which is in the possession of the data controller, or is likely to come into its possession (section 1(1)). Conversely, therefore, data which is anonymous under the DPA means that the data controller does not possess and is not likely to acquire additional information necessary to enable an individual to be identified from that data.

standard described above is satisfied before anonymised data is shared. There is accompanying guidance on how that test should be interpreted in terms of assessing re-identification risk.[26]

Recital 26 GDPR also states that the principles of data protection should not apply to 'anonymous information', which it describes – like the Data Protection Directive – as "*information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable".*[27] The GDPR also describes the concept of 'identifiability' through the use of a 'means test' standard, for assessment upon the facts of the data environment under consideration, similar in language to the Directive. In other words, there is an appreciation that identifiability from data may be possible through means reasonably likely to be used by any person, including potential authorised re-users, but also other third parties that may have an interest in obtaining the data.

However, unlike the Data Protection Directive, Recital 26 GDPR refers explicitly to "singling out" as a means by which someone may be deemed identifiable from data (directly or indirectly); therefore, this possibility should also be taken into account in determining whether anonymisation has been carried out effectively or not (along with other means "reasonably likely to be used" either by the controller or by another person to identify someone from data). [28]

5.2 Singling out and pseudonymisation

In illustration of the fact that singling out someone from data may be deemed a means to their identification (directly or indirectly), Recital 26 GDPR also states that "*personal data which have undergone pseudonymisation, 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*". This is because pseudonymisation is a technique applied to personal data commonly used to refer to a process whereby direct identifiers are removed, albeit that indirect identifiers remain in the data, in turn raising the prospect that the relevant subject may yet be re-identified because they are distinguishable from other people in the data.

Article 4(5) GDPR formally introduces 'pseudonymisation' as a legal term formally defined according to a more exacting definition: "*the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person*". For example, to satisfy this legal definition, any 'key' necessary to identify data subjects from key-coded data must be kept separately,

[26] For example, factors to consider include the cost of conducting the identification, the intended purpose of processing, the way processing is structured, the advantage expected by the data controller, the interests of the data subject and any risk of organisational and technical failure.

[27] The GDPR, Article 4(1), defines 'personal data' as *"any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person".*

[28] The GDPR, Recital 26, states as following: *The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, 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. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.*

and subject to technical and organisational security measures to prevent inadvertent re-identification from the coded data.

Importantly, the GDPR refers to pseudonymisation as a means by which personal identification from data may be assumed to be *still* possible, despite it previously being thought of as an effective form of anonymisation technique. In other words, there appears to be an automatic legal *presumption* that data subject to pseudonymisation processes as defined by the GDPR should still be treated as personal data. This is because pseudonymised data can be attributed to a natural person by the use of additional information allowing for re-identification of the singled out individual.

However, it is still considered possible for this presumption to be overturned, where robust technical, legal, and organisational measures are taken so that the risk of the data subject being re-identified becomes 'reasonably' impossible. This topic is particularly relevant, in theory, to issues arising under the Project because of the possibility that data relating to persons that have been subject to pseudonymisation are proposed to be shared with Participating SMEs.[29]

5.3 Anonymisation/pseudonymisation strategy under the Project

Two approaches have been considered under the project in relation to the sharing of data relating to persons by the Data Providers to the Participating SMEs, as outlined below:

Approach 1

One approach open to the Consortium considered at the planning stage would have been to require that all data relating to persons shared under the Project would be aggregated (pre-sharing) to the extent that no person-specific data (from which individuals can be singled out) remains in the datasets available for analysis by the Participating SMEs.

While this was considered as a possibility, the implicit trade-off between data perturbation and latent data value (the greater the degree of the former, the less the potential for the latter) is worth acknowledging. Thus, while minimising the risk of negative impact upon individuals that might arise when anonymised data is later processed, anonymisation techniques are typically applied to data in ways that allow value to still be extracted from it post-anonymisation. In reality, this means that more data value can often be extracted when individual-level granularity is left in datasets for experimentation.

Approach 2

An alternative approach open to the Consortium which we believe to be legitimate - as a means to balance the requirements of data protection law with the need for preserving data utility, and therefore provide opportunity for value extraction from the secondary reuse of data relating to persons - is to allow pseudonymised data to be shared under the Project in certain cases that justify this approach. However, as mentioned, data relating to persons from which direct identifiers have been removed through pseudonymisation are still likely to be characterised as personal data under EU data protection law, unless effective measures are put in place to overturn this presumption.

Therefore, pseudonymised data that is proposed to be shared under the Project would only be acceptable for use in the Project when accompanied by the putting into place of appropriate safeguards (legal, technical, and organisational steps) for reducing re-identification risk to an adequately safe level. In particular, these should reduce the risk of Participating SMEs from re-identifying the data subjects in the datasets shared with them to

[29] This may be because Data Providers believe that the application of this technique provides a guarantee of legal anonymity. This is clearly not the case under the GDPR regime and possibly also not under the Data Protection Directive. Further legal guidance on this issue has been provided by the pan-EU Article 29 Working Party in an Opinion published in 2014 on Anonymisation Techniques. A summary of this Opinion is set out in **Annex C** to this Deliverable. See also Stalla-Bourdillon, Sophie, and Alison Knight. "Anonymous data v. Personal data–A false debate: An EU perspective on anonymisation, pseudonymisation and personal data." *Wis. Int'l LJ* (2016).

an extremely low level. Putting in place such re-identification risk mitigatory measures will also facilitate compliance with the GDPR if it were later deemed to apply by a national data protection authority or court.[30]

At the same time, it is recognised that – under this alternative approach - given the residual risk of re-identification, the act of anonymisation/pseudonymisation and supporting measures cannot be treated as a one-off exercise. As re-identification risks can change over time, re-identification risk assessments and management of the results of such assessments should be applied iteratively during the life of the Project. In other words, a 'release and forget' approach to previously personal data that is not appropriate insofar as safeguards to mitigate re-identification risk must be kept in place during the lifetime of the experimental phase of the Project (after which time Participating SMEs would be required to delete data shared with them).

Accordingly, and considering the diversity of potential situations, a 10-point multi-factor strategy under this approach has been developed and is set out below. The strategy is to be read alongside the next section which contains practical guidance more generally for both Data Providers and Participating SMEs, including in scenarios where pseudonymised data is not relevant. This strategy will help to ensure both that re-identification becomes 'reasonably' impossible from any pseudonymised datasets for sharing, and also that any secondary processing of such datasets by a SME recipient would be compliant with the GDPR:

1) Technical and organisational measures should be put in place to make sure Participating SMEs do not have access to the additional information required for recovering direct identifiers. For example, Participating SMEs will be bound by confidentiality obligations and restrictions on reuse and re-identification.

2) The pseudonymised datasets for sharing with the Participating SMEs should be stored on company-secure servers, or on the servers of the University of Southampton. Where 'keys' that would enable the linking back of real-world identities to pseudonyms are retained, these should always be safely secured by the Data Provider using organisational and technical safeguards (to reduce the risk of illegitimate access to such information).[31]

3) Data Providers should only share data relating to persons to the extent that it is necessary to achieve a previously-delineated purpose or purposes.

4) Moreover, any indirect identifiers in the pseudonymised datasets for sharing should be removed or masked where these would not be strictly necessary for the participating SME to achieve the specified purpose(s). Similarly, possibilities for linking or inferring new information about data subjects from analysing the datasets – which could increase their risk of re-identification by the SME – should be muted as far as strictly necessary relative to the specified purpose(s).

5) The Participating SMEs should only be permitted to process the pseudonymised datasets for sharing for a specified analytics-driven purpose(s).

6) Where that specified purpose(s) does not relate to scientific research (technological development and demonstration, fundamental research, applied research, or privately funded research), it should be compatible with the initial purpose(s) for which the data was originally collected. Obtaining a clear description of the initial purpose and the legal basis justifying the initial collection and the challenge to be solved will help in assessing this compatibility.

[30] In other words, this approach also takes into account that, while pseudonymisation (alone) might not transform personal data into non-personal data, it serves a useful security purpose (recognised by the GDPR) as a legitimate way to minimise the likely privacy harm that might befall data subjects when their personal data are processed. Therefore, its use is to be encouraged generally as part of a data protection by design approach as described in section 4.

[31] This security mechanism would be strengthened through the legal terms of the contract agreed with them (e.g. prohibiting them contractually from sharing the key under any circumstances).

7) As data subjects have a right to object to secondary processing on data relating to them in certain circumstances, data subjects should be informed of what is being proposed where the scope of the initial consent obtained from them does not extend to the specified purpose(s). Otherwise, another legal basis would be required to justify the processing, such as the 'legitimate interest' basis. This would requiring analysis by the Data Provider regarding whether the processing is necessary for the purposes of the legitimate interests pursued by them, and are not overridden by the interests or fundamental rights and freedoms of the data subject which require personal data protection. The purpose of this balancing exercise is to prevent disproportionate impact on individuals. In practice, carrying out this exercise will also require a full assessment of the facts and context of each case as relevant.[32]

8) If the specified purpose(s) would involve data subjects being subject to 'profiling' analytics – and measures or decisions are subsequently planned to be taken vis-a-vis individual customers or groups of customers, based on profiles that might be created by the Participating SME – we advise that Data Providers carry out a DPIA. This would again require a full assessments of the facts and context of each case, and should be done before the measures or decisions are taken (and, if possible, even earlier before the Participating SME receives the relevant data from them).[33] Any high risks of impact to data subjects identified under a DPIA should then be addressed through the Data Provider taking different kinds of safeguards to mitigate such risks in proportion to the level of harm predicted. This might include technical and organisational measures taken to ensure 'functional separation' through data silo'-ing (i.e. data used for research purposes not being made available to support decisions that could be taken with regard to individual data subjects). Notwithstanding, when an organisation specifically wants to analyse or predict personal preferences, behaviour and attitudes of individuals from data they propose sharing - which will subsequently inform 'measures or decisions' that are taken with regard to those customers - free, specific, informed and unambiguous 'opt-in' consent should be obtained.[34] Ultimately, if these safeguards cannot be met, the Data Provider would have to rethink the data being made available for reuse and/or the specified purposes of that reuse.

9) Participating SMEs will be required to comply with, and provided training on compliance with, data protection law under the GDPR. They will be required to comply with data protection law and prohibited from re-identifying data subjects.

10) Participating SMEs will also be required to destroy the data at the end of the acceleration stage of the Project in which they are involved. Further, publication of research results should, as a rule, be possible in such a way that only aggregated (and/or otherwise fully anonymised) data will be disclosed.

5.4 Mosaic effects

So-called 'mosaic effects' denotes a broad concept underpinned by the idea that data may be linked to other information (to create a 'mosaic') in ways that increase privacy risks (potential 'effects') to those individuals about whom the data relates. For example, data techniques can reveal intimate personal details because of opportunities for data fusion – the merging of data, in particular across disparate datasets and information sources. Like joining together different pieces of a jigsaw puzzle, for example, data analytic techniques are specifically aimed at merging multiple data sets to reveal complex patterns and infer new knowledge. Where

[32] To note, that depending on their capacity, some national data protection authorities are happy to assist with, or assess results of, this exercise.

[33] In other words, it is recommended that such a robust and detailed impact assessment should be completed *prior* to the disclosure of information and making it available for reuse.

[34] For the consent to be informed, and to ensure transparency, data subjects should ultimately also be given access to 'profiles' relating to them, as well as to the logic of the algorithm that led to the development of the profile. Furthermore, the source of the data that led to the creation of the profile should also be disclosed and the decisional criteria.

data relates to persons, such fusion might permit the revelation and inference of new information in terms of what can be concluded about those people (potential **privacy attribute disclosures**).

In the context of data to which techniques have been applied to personal data to hide the identity of the data subject, the possibility of data fusion also has relevance because it may also be possible to work out who that person is by joining together the modified data with other auxiliary information (potential **identity attribute disclosures**). In that context, mosaic effects has been defined as "*the process of combining anonymous data with auxiliary data in order to reconstruct identifiers linking data to the individual it relates to"*.[35] Thus, personal details may be discerned or risk becoming discernible even from ostensibly 'anonymous' data through linking and combining multiple data points about the same person. The underlying presumption in such scenarios being that individual identifiers in these datasets would not otherwise allow a data subject to be re-identified. This is especially relevant to large scale data sharing and repurposing because seemingly 'anonymous' data sets can often be combined and analysed to reveal restricted or sensitive information. In other words, rather than the information shared being sensitive, sensitivity may lie in the inferences that are drawn from its processing and the way in which those inferences are drawn, potentially giving cause for concern.

Such concerns are especially topical in the context of research around 'big data' analytics and re-identification risk, such that this theme has been considered to deserve explicit consideration and treatment in the Project. Such trending concerns are because large scale data analytics involving sharing and repurposing of seemingly 'anonymous' datasets can often result in the revelation of new personal details about a person to whom they relate. This is despite the fact that individual identifiers in information that has been subject to anonymisation techniques would not – in and of themselves – enable its subject to be re-identified.

In legal terms, as alluded to, this issue of a potential mosaic effect arises alongside the issue of determining whether data protection law – which applies to any processing of personal data relating to living persons – would in fact apply to a specific data processing activity. There is concern particularly for data analytics that personal data may be discerned from data despite efforts being made to anonymise personal data by stripping it of identifiers. Consequently, data protection rules would then apply to the processing of this data set, which may not have been envisaged when it was planned to be processed. In broad terms, this would mean that organisations processing de-identified data typically could not use that data for purposes beyond those for which it was originally obtained where these purposes are incompatible, and that data could not be kept indefinitely.

5.5 Mosaic effects mitigation strategy under the Project

For the Project, concerns could be raised that mosaic effects might arise either as a result of integrating different data sets being shared under the Project and finding commonalities, or by linkage of individual datasets to other sources of information that are outside the control of the Consortium. As mentioned, in particular, there is a risk that data protection law might consequently apply in relation to all challenges where the processing of data relating to persons is involved where individuals can be singled out from the relevant data shared. Special concern is also reserved for the possibility that sensitive personal data might be inferable from specific datasets. As mentioned, categories of sensitive personal data raise a high level of concern, such that a stricter standard of data protection obligations should always apply to their processing.

To explore such concerns further, it is useful to set out the main activity element of the Project in outline:

[35] Maude, F., 2012. Open Data White Paper-Unleashing the potential. *The Stationary Office Limited on behalf of HM Government, Cabinet Office, London, United Kingdom*, glossary. Compare a similar non-UK specific definition, "*whereby personally identifiable information can be derived or inferred from datasets that do not even include personal identifiers, bringing into focus a picture of who an individual is and what he or she likes*" (United States. Executive Office of the President and Podesta, J., 2014. *Big data: Seizing opportunities, preserving values*, p.8).

- Each Participating SME will be able to access and process data shared by a specific Data Provider after they have been selected to take part in the main acceleration stage of the Project.
- A single Participating SME will be selected to work on a single challenge (and only one, although they will not be precluded from applying for another, unrelated challenge upon the second call for participants).
- Such data will have been collected externally to the Project (either by the Data Provider themselves, or exceptionally by third parties who have made such data available to the Data Provider). Such data will typically be 'closed' data (that is, not data which is not currently in the public domain).
- Each dataset shared by a Data Provider will be assigned a Project 'challenge' linked to revealing undiscovered value from them. Challenges will either be defined by the Data Provider with the help of the Consortium linked to the reason why they would like the data analysed (such as a key issue facing the Data Provider that they believe can be addressed by the data – a **'provider-driven' challenge**); or, set by the Consortium in consultation with others around general 'high-impact' problems related to specific types of data in specific sectors that might be addressed through a dataset's exploration (an **'exploratory' challenge**). Although the latter class of challenge will be wider (less focused) in scope in terms of what output might be achievable than the former class, it will still be accompanied by restrictions – in particular, the data shared will only be available for use by the participating SMEs within the terms of that challenge and no other.
- It may be possible that exploratory challenges will also be set that are not attached to a specific dataset shared by a Data Provider. In particular, it could be proposed by the Participating SMEs themselves using any combination of data that they provide themselves (e.g. in combination with other open data).

Specifically, the following risk-scenarios from which mosaic effects might flow from these procedures are concerning:[36]

1. The possibility of combining data about people within a Data Provider dataset with external sources of information, including information brought to the Project for analysis in combination with the Data Provider dataset. This could include data collected by the Participating SME, data that has been shared with the Participating SMEs by third party data providers, and/or data made publicly available, in relation to the same singled out individual.
2. The possibility of combining data about people within a dataset provided by a Participating SME with external sources of information, including data publicly available[37] as well as closed datasets made available to the Participating SMEs by third party data providers, in relation to the same singled out individual.

The strategy adopted to address the possibility of 'mosaic effects' arising at the acceleration stages of the project, as described, encompasses anonymisation and pseudonymisation best measures as outlined above, the latter to be applied to data from which individuals can be singled out under the Project in any event.[38] Following this approach should reduce some types of mosaic effects to a low level, including in particular the imposition of contractual obligations, such as confidentiality, on the Participating SMEs.

[36] The separate concern that about combining data about people within the same Data Provider dataset whereby linkages could be made to multiple data points in relation to a singled out individual has been dealt with in the above section under pseudonymisation strategy.

[37] To note, the mere fact that data relating to persons has been made publicly available does not lead to an exemption from data protection law. The reuse of such data remains subject, in principle, to data protection law if it is personal data.

[38] In other words, data that has been pseudonymised that is proposed to be shared under the Project would only be acceptable when accompanied by the putting into place of appropriate safeguards (legal, technical, and organisational steps) for reducing re-identification risk to an adequately safe level.

Also, as part of the overall, comprehensive risk-management approach adopted under the Project to reduce – not just re-identification risk but also - the likeliness and severity of any harm potentially caused by the secondary processing of data to its subjects, the following notification procedure is in place. In the event that a Participating SME wishes to provide datasets for use under the Project, it will be obliged to provide a description of each such dataset to the Consortium. Such description will include: the original and ownership of such data; the rights that the Participating SME has to use such data; and, whether such data has been subject to any anonymisation or pseudonymisation procedures. Following such notification, the Consortium will assess the potential for mosaic effects arising from the combination of such data and the Data Provider's dataset, or arising from the combination of discrete datasets that the Participating SMEs wishes to self-supply. The notification will only be rejected if such effects are deemed reasonable likely to arise (e.g. if such data sources might conceivably relate to the same individuals).

6 Turning theory into practice

As mentioned, the legal framework for sharing and re-using data presents as a complex picture, giving rise to challenges in assessing and managing associated risks. In particular, the legal implications of a third party using data in relation to which rights already exist – and for purposes other than that for which it was originally obtained – are key issues. This is because:

- Different types of law act concurrently in relation to a data sharing arrangement (e.g. IPR as database right or copyright), contractual rights and duties, and data protection regulation where personal data is involved or likely to be involved in secondary processing of data relating to persons). In other words, legal rights and duties arise in a multi-layered way and may differ between countries (including within the EU bloc).
- The requirements for legal and regulatory compliance, on the one hand, and innovative efficacy on the other hand, are sometimes hard to reconcile.

The identification of a separate data handling protocol for Data Providers and Participating SMEs is valuable and this is set out in guidance form in sub-sections 6.1 and 6.2 below. In this first version of the toolkit, the issue of data protection risk mitigation is the key focus, supplemented by contractual provisions (whereby the Consortium imposes contractual requirements on Data Providers and Participating SMEs to ensure that they act in compliance with legal analysis already carried out based on EU best practices).

To note for the Project, data will only be permitted to be shared with Participating SMEs once a legal review has been carried out, and a legal agreement signed by, Data Providers. This will require individual discussions with the Consortium about defining the conditions for the data's reuse in light of legal compliance, and setting out the scope of the purposes for which each dataset will be processed (and only processed) under the Project.

6.1 Data sharing methodology for Data Providers - managing the legal risks associated with data sharing in practice

Data Providers will need to review the types of data that they propose to share with the SMEs in advance of sharing to assess what types of legal issues arise in relation to that data – in particular, whether or not such data falls within the definition of personal data or not. When anonymising personal data, organisations should assess the risks that the data could be re-identified given the nature of the data, the context in which the data will be used, and the resources available to those with access to the data.

Where they consider that they data types do fall within the definition of personal data, or there is a reasonable risk that they might be construed as personal data, consideration must be given to:

- How to reduce the risk of re-identification of the data to be shared; and,
- If it is not possible to reduce that risk, or where pseudonymised data must be shared in order to achieve the challenge they set for the SME in relation to the data they provide, how to ensure that the requirements of the GDPR would be met upon reuse and how to minimise any risk of harm to data subjects flowing from the secondary processing. These could include compliance with extra requirements, such as performing a DPIA.

Best practice recommendations are suggested in the next sections to be followed by Data Providers in respect of the stages of consideration of data protection risks.

6.1.1 Assessing the effectiveness of – and managing - anonymisation/pseudonymisation techniques

Before the acceleration stage of the Project where the SME start to utilise the data, there are two, interlinked areas being assessed under the Project to reduce re-identification and harm risk flowing from the further processing of shared data when it relates to persons. These involve the Data Providers: ensuring the implementation of effective anonymisation practices (bearing in mind that no anonymisation technique is a fail-safe solution to the problem of re-identification risk); as well as putting in place adequate non-technical elements (organisational and legal measures).

Each Data Provider will be asked to complete a Consortium-prepared questionnaire (a copy of which can be found in **Annex B** below), aimed at, inter alia, determining what data relating to persons is intended to be shared and the extent of the risks that might arise if such data were processed under the Project. A key section of that questionnaire relates to determining what anonymisation techniques have been applied to data relating to persons intended for sharing by the Data Providers, together with any other steps they have taken to reduce re-identification risk (such as the secure storage of any 'key' that would be required to back-track a pseudonym to a recognisable individual).

Responses to such questionnaires are analysed by the relevant Consortium member against EU guidance on best standards for ensuring the effective anonymisation of personal data and associated steps that can be taken to reduce the risk of data subjects being re-identified from transformed data (taking into account the 'means test' standard for determining 'identifiability' as set out in Recitals 26 of the Directive and the GDPR (see sub-section 5.1)).

In particular, as described by the Art.29 WP in 2014 in its Opinion on Anonymisation Techniques ('WP216' [9], summarised in **Annex C** below), the adequacy of the anonymisation solution applied to data by Data Providers will be considered in terms of the risks of any future data recipients being able to:

- single out an individual in a dataset;
- link two records within a dataset with respect to the same individual; or,

- infer any information in such dataset about an individual (in particular, the inference of sensitive personal data about them).

This analysis also addresses the possibility of mosaic effects that might arise where different data sets may be combined, or shared data linked to other unknown external sources that are out of control of the Data Providers. Another recommended piece of regulatory guidance is advice given by the UK ICO in particular in its Anonymisation Code of Practice ([10], summarised in **Annex D** below).

Data Providers should consider the above guidance on a case-by-case basis relevant to the datasets they propose to share and the pre-set challenges (specified objectives) attached to each dataset, alongside possible harm mitigation measures that can be taken. In particular, the amount of linkability compatible with a challenge determines the anonymisation (or pseudonymisation) technique appropriate to be applied.

We also recommend a dynamic approach to anonymisation/pseudonymisation that involves consideration of the relevant data environment. In other words, Data Providers should assess the risks that the data could be re-identified given the nature of the data, the context in which the data will be used, and the resources available to those with access to the data. In particular, the purposes to be achieved by way of the processing of modified dataset should be clearly set out not least because they play a key role in determining the likelihood of residual re-identification risk arising. Such risk should then be reduced to an acceptable level.

It is an important consideration that the weaker the level of anonymisation techniques applied to personal data (so-called 'sanitisation', or indeed where only pseudonymisation is applied to such data), the stronger organisational and legal measures are needed to ensure that the overall risk levels associated with secondary processing remain low.

6.1.2 An overview of different types of de-identification methods

It is very important to take great care, at the initial stage of producing, disclosing and making available for reuse, any information derived from personal data. Several techniques can be considered for application to de-identify personal datasets and reduce any risks of re-identification of an individual prior to making the data available for reuse. As discussed above, full or partial anonymisation can be relevant to the safe use or sharing of data between organisations. When full anonymisation and use of aggregated data (at a sufficiently high level of aggregation – the most definitive solution to minimising the risks of inadvertent disclosure) is not possible, data will often at least need to be partially anonymised and additional safeguards may also be required, as discussed below.

It is helpful to distinguish different scenarios for further analysis ranked according to the strength of privacy protection provided (from the weakest to the strongest):

- **Scenario 1:** situations where directly identifiable personal data are needed due to the nature of the research and other solutions are not possible without frustrating the purpose of the processing, and further provided that other appropriate and effective safeguards are in place). This scenario is not outside the scope of the Project.
- **Scenario 2**: situations involving indirectly identifiable personal data: lower level of aggregation, partial anonymisation, pseudonymisation or key-coded data.
- **Scenario 3:** unidentifiable personal data: data are anonymised or aggregated in such a way that there is no remaining possibility to (reasonably) identify the data subjects.

The robustness of each technique is considered taking into account the three re-identification risk categories described by the Art.29 WP WP216 reformulated as the following questions:

(i) is it still possible to single out an individual?

(ii) is it still possible to link records relating to an individual?

(iii) can information be inferred concerning an individual?

Scenario 2

This scenario covers situations of partial de-identification where full anonymisation is not practically feasible due to the nature of the processing (e.g. where there is a need to use more granular data which as a side effect, may allow indirect identification). It takes into account that full anonymisation is increasingly difficult to achieve with the advance of modern computer technology and the ubiquitous availability of information. Thus, re-identification of individuals is an increasingly common and present threat.

In practice, there is a very significant grey area, where a data controller may believe a dataset is anonymised, but a motivated third party will still be able to identify at least some of the individuals from the information released. Furthermore, data controllers must be aware that the risk of re-identification can change over time, (e.g. powerful data analysis techniques that were once rare are now commonplace). Addressing and regularly revisiting the risk of re-identification, including identifying residual risks following the application of the following techniques, therefore remains an important element of any solid approach in this area. Good practice requires organisations to carry out a periodic review of their policy on the release of data and of the techniques used to anonymise it, based on current and foreseeable future threats. Notwithstanding a good practice approach, there remains the risk that data protection law continues to apply where such datasets are processed and this risk will need to mitigated by additional safeguards beyond the de-identification techniques described under the next sub-heading.

Removing identifiers

The first step of de-identification is to remove clear identifying variables from the data (name, date of birth or address).

Although some identifiers are removed, datasets may still retain a relatively high potential for re-identification as the data still exists on an individual level and other, potentially identifying, information has been retained. For example, some address postcodes have very small populations and combining this data with other publicly available information, can make re-identification of data subjects living in such postcodes a relatively easy task.

Key-coding

The technique involves consistently replacing recognisable identifiers with artificially generated identifiers, such as a coded reference or pseudonym (e.g. a randomly selected number). This allows for different information about an individual, often in different datasets, to be correlated.

This method has a relatively high potential for re-identification, as the data exists on an individual level with other potentially identifying information being retained. Also, because key-coding is generally used when an individual is tracked over more than one dataset, if re-identification does occur more personal information will be revealed concerning the individual. For that reason, key-coding is not considered a method of anonymisation. It merely reduces the linkability of a dataset with the original identity of a data subject, and is accordingly a useful security measure.

Reducing the precision of the data

Rendering personally identifiable information less precise can reduce the possibility of re-identification. For example, dates of birth or ages can be replaced by age groups.

Related techniques include suppression of cells with low values or conducting statistical analysis to determine whether particular values can be correlated to individuals. In such cases it may be necessary to apply the frequency rule by setting a threshold for the minimum number of units contributing to any cell. Common threshold values are 3, 5 and 10.

Introducing random values (‘adding noise’) is more advanced and may also include altering the underlying data in a small way so that original values cannot be known with certainty but the aggregate results are unaffected.

Various other techniques (such as keyed-hashing, using rotating salts, removing outliers, and replacing unique pseudonyms) may also be used to reduce the risk that data subjects can be re-identified, and subsequently, that any measures or decisions can be taken in their regard.

Scenario 3

Full anonymisation (including a high level of aggregation) is the most definitive solution. It implies that there is no more processing of personal data and that data protection law is no longer applicable. However, it is a high standard to meet, in particular because it requires that any reasonable possibility of establishing a link with data from other sources with a view to re-identification be excluded.

Aggregation

Individual data can be combined to provide information about groups or populations. The larger the group and the less specific the data is about them, the less potential there will be for identifying an individual within the group. An example is aggregating postcodes at a regional level.

6.1.3 Importance of additional risk assessment and mitigation measures (safeguards and controls) beyond de-identification techniques

The above analysis shows that applying techniques to personal data is a key tool in achieving different levels of de-identification, but most of these procedures have their challenges and limits. The analysis also shows that once a first assessment has been completed in terms of the possibilities and limits of effective de-identification, a second step of assessing the need to complement these techniques with additional (organisational and technical) safeguards will often need to followed.

This second stage is about applying these other safeguards in order to adequately protect the data subjects. Risks flowing from secondary processing of data relating to persons must be reduced to an acceptable level after they are identified to prevent re-identification and reduce the future likelihood of harm. Safeguards may involve technological, legal, and administrative measures, and often a combination of each.

As an essential guide, it should be kept in mind that the easier it is for the data subject to be identified, the more additional safeguards will be needed. Furthermore, the more consequential potential adverse impact on the data subject if identified would be, the more should be done to limit the possibilities of re-identification and the more additional safeguards may be required.

Among the appropriate safeguards which may bring additional protection to the data subjects, the following could be considered and applied to the data provided by each Data Provider in the Project:

- taking specific additional security measures (such as encryption);
- in case of removing identifiers and key-coding data, making sure that data enabling the linking of information to a data subject (the 'keys') are themselves also coded or encrypted and stored separately to reduce the risk of illegitimate access to such information.[39]
- entering into a trusted third party arrangement; this model is increasingly being used to facilitate large-scale research using data collected by a number of organisations each wanting to anonymise the personal data they hold for use in a collaborative project. Trusted third parties can be used to link datasets from separate organisations, and then create anonymised records for researchers to work on.
- restricting access to personal data only on a need-to-know basis, carefully balancing the benefits of wider dissemination against the risks of inadvertent disclosure of personal data to unauthorised

[39] As mentioned, the formal legal definition of 'pseudonymisation' under the GDPR (Article 4(5) describes it to mean, "'*the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organisational measures to ensure non-attribution to an identified or identifiable person*".

persons. This may include, for example, allowing read-only access on controlled premises. Alternatively, arrangements could be made for limited disclosure in a secure local environment to properly constituted closed communities. Legally enforceable confidentiality obligations placed on the recipients of the data, including prohibiting publication of identifiable information, are also important.

To note, in high-risk situations, where the inadvertent disclosure of personal data would have serious or harmful consequences for individuals, even the strongest type of restriction may not be suitable. That is because, even with safeguards in place, these would not be considered adequate to prevent undue impact on the data subjects.

6.1.4 De-identification and impact mitigation checklist

The following points should be considered:

- Determining of any specific unique (such as national identity number), or quasi-unique (such as names, date of birth), identifiers in the relevant data.
- Cross-referencing to determine unique combinations like age, gender, and postcode.
- Acquiring knowledge of other publicly available datasets and information that could be used for list matching.

Even with such variables missing other factors should be considered:

- Motivation to attempt identification
- Level of details (the more detail the more likely identification becomes)
- Presence of rare characteristics
- Presence of other information (a dataset itself may not include any data that can identify an individual, but it may include enough variables that can be matched with other information).
- What a potential breach of privacy could mean for related individuals. More specifically, the level of potential impact will be dependent on the likelihood that identification could occur from the sharing of data but also the consequences of such sharing in terms of what it will be reused for. Factors that will help assess the level of likely impact to result from a processing activity include consideration of: the purposes of the proposed secondary processing; the proportionality of the secondary processing operations proposed in relation to such purposes; an assessment of the risks to the rights and freedoms of data subjects; and, any measures envisaged appropriate to address such risks.
- Any challenge set by a Data Provider requiring data analytics involving profiling should be considered carefully. In practice, it would need to be presumed that any secondary processing activities on data relating to persons involving profiling would require not just compliance with the GDPR (as presumed wherever there is a risk that personal data are to be processed under the Project), but also extra caution. In particular, obligations regarding the provision of information to the data subject in respect of the carrying out of profiling activities are more onerous compared with other processing activities (as described in section 4.3.1.1 above). Thus, consents and notices for profiling should be reviewed carefully. Moreover, the GDPR requires controllers to conduct a DPIA any time "*a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, **including profiling**, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person*." In this context, a data controller subject to the obligation to carry out the DPIA *"shall maintain a record of processing activities under its responsibility"* including the purposes of processing, a description of the categories

of data and recipients of the data and *"where possible, a general description of the technical and organisational security measures"*. Data Providers will have to carefully consider whether a DPIA is needed for this Project from 25 May 2018.

- **What profiling activity could mean for related group interests.** Although this may not be so relevant to this project, for completeness, if data controllers plan to make decisions at a group-level (i.e. potentially affecting all individuals within a group), likely impact on collective interests should also be considered by them.[40] Even without decisions being planned in respect of a singled-out group, there may still be inherent risks of discrimination, or other negative impact liable to flow, from the making of profiling assumptions (that is, the detecting of patterns in data between a particular type of person grouped into various categories by profilers, and a type of behaviour or characteristic). We recommend such risks of harm also be considered in carrying out DPIAs, such as the possibility of incorrect assumptions being made about individuals categorised in line with an inferred group profile.

6.1.5 Other safeguards under the Project where data relating to individuals are to be shared

Where personal data are processed, data protection law requires that such data must be obtained only for specified lawful purposes and not further processed in a manner incompatible with those (original) purposes. Data Providers will be asked to share details in response to the Consortium's questionnaire about the lawful purpose for which they currently process any personal data to be anonymised/pseudonymised and shared under the Project. Potentially personal data that does not have a lawful purpose (under the Data Protection Directive (Article 7) and the GDPR (Article 6)) will not be shared under the Project. Regarding the lawful reuse of anonymised/pseudonymised data (assuming it were in fact deemed to be personal data, such as by a data protection authority in a relevant EU Member State, or a court), where relevant any consents given by the data subjects to the original processing activity will be considered.

However, relying upon consent as legal basis for secondary purposes under the Project is recognised as challenging. This is in light of the restricted definition of 'consent' under the GDPR – in reminder (see footnote 5), it is defined as meaning a "*freely given, specific, informed and unambiguous indication of the data subject's wishes" signifying agreement to the processing of personal data relating to him or her"* (Article 4(11)). Whereas, consent will *not* be presumed to be freely given according to Recital 43 GDPR unless separate consent opportunities for different personal data processing operations have been proffered. These requirements are unlikely to be satisfied where pre-obtained data sets are repurposed to unlock yet unrevealed value (such as in the case of exploratory challenges). In other words, it is not easy to implement the principle of purpose limitation effectively in cases where processing purposes cannot be clearly defined or clearly foreseen as consent will not be presumed. An alternate legal basis to justify secondary personal data processing must be considered (see the related discussion in section 4.3.1.1 above), along with the establishment of appropriate data protection safeguards.

For the avoidance of any doubts, in the event that data protection law applies to the processing carried out during the acceleration stage, the contracts entered into for the Project will also require that the Data Providers acknowledge their obligations as data controllers, including in respect of those who process (potentially) personal data on their behalf, and the Participating SMEs acknowledge that they may be construed as data processors (or, indeed, as joint data controllers of such data).

[40] For guidance on how privacy and data protection may be interpreted as referring to collective interests, see e.g. Mantelero, A., 2017. From Group Privacy to Collective Privacy: Towards a New Dimension of Privacy and Data Protection in the Big Data Era. In Group Privacy (pp. 139-158). Springer International Publishing.

6.1.6 Assessing the adequacy of – and managing - data security measures

Article 32(1) GDPR follows a risk-based approach regarding imposing an obligation on data controllers to ensure the security of personal data processing. When assessing the appropriate level of security required, the data controller must take into account the risks that are presented by a relevant processing activity (in particular from accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to personal data transmitted, stored or otherwise processed). This reflects the new risk-based approach taken by the GDPR in general to data protection law compliance.

Article 32(1) also sets out a number of specific "appropriate technical and organisational measures" controllers and processors could take, by way of suggestions for the contents of a data security policy. These are:

- The pseudonymisation and encryption of personal data.
- Measures ensuring the ability to ensure the ongoing confidentiality, integrity, availability and resilience of systems and services processing personal data.
- Measures ensuring the ability to restore the availability and access to data in a timely manner in the event of a physical or technical incident.
- A process for regularly testing, assessing and evaluating the effectiveness of technical and organisational measures for ensuring the security of the processing.

Regarding security measures to be put in place under the Project, Data Providers will be asked to complete a technical questionnaire so that the Consortium can suggest the most appropriate option for the hosting of their (personal or non-personal) data. Such suggestions will be framed in light of the relevant contextual elements under consideration from which an assessment of any attendant risks (their type, likelihood and severity) will be carried out.

Four possible options for storing and providing access to data sets shared by the Data Providers by the Participating SMEs have been considered:

5) The University of Southampton hosts the data. Multiple levels of security (data hosting facilities using secure infrastructure managed by the Consortium at the University of Southampton) can be offered depending on what is deemed appropriate to the level of re-identification assessed to be present. These include: connection to the public internet through Janet, the UK university network; or, defining a secure zone Data Pitch via a standard proxy - an academic infrastructure hosting all the academic services and laboratories (this can be supplemented by the addition of a further proxy to create an isolated and secure zone for Data Pitch processing which is managed independently).[41]
6) The Data Provider hosts the data.
7) The commercial cloud hosts the data (under the direction of the Data Provider).
8) The participating SME (chosen to process the relevant data) hosts it.

Ultimately, the Consortium realises that the option of choice for the secure provision of data relating to individual to be shared by the Data Providers will be led by their preferences and their current legal compliance measures taken as data controllers. However, the Consortium will advise the Data Providers in their choices

[41] The most secure level - hosting on a separate, BIL3 and ISO27001-compliant network in the University's Data Centre with a requirement for physical authentication for network access – has been discarded. Fulfilling this requirement is considered inappropriate because the Participating SMEs will be based around Europe. Furthermore, this type of physical-only access (and workstation isolation) is deemed suitable for the processing of highly sensitive data, of the type which will not be allowed to be shared under the Project.

as far as possible, in particular when pseudonymised data is shared, to ensure that they are aware of the extent of their legal compliance obligations under existing and incoming data protection rules.

Options (1) or (2) are the Consortium's preferred option for hosting data relating to persons under the Project.[42] This is because of the risks in the other options as set out below:

- **Option (3) -** Hosting potentially personal data in a commercial Cloud carries security and other data protection risks.[43] While the Data Provider may consider that such risks are mitigated sufficiently when they manage the private Cloud storage themselves – in accordance with their duties as data controllers of such data under data protection law – the risks associated with the use of professional Cloud providers are normally not so mitigatable. For that reason, we are less inclined to recommend that Data Providers share access to potentially personal data via – for example non-EU (e.g. US) - cloud providers, in particular where it can be assumed that they use servers based outside the EU. Alternatively, recommend that the data provider has carried out a DPIA in this respect which they can evidence to the Consortium, in particular in respect of where data may be stored outside the EU (including regarding data 'adequacy' arrangements with non EU countries).
- **Option (4) –** This is discouraged as a means to host potentially personal data, unless the chosen SME already has a strong data analytics infrastructure, and a DPIA exercise has been carried out and can be evidenced to the Consortium with respects to the adequacy of the security over data currently hosted. At the very minimum, the Participating SMEs must have control mechanisms in place and evidence an organisational culture that encourages compliance with data protection law (see the next section on Participating SME requirements in relation to data protection law).

Where the Consortium feels that the re-identification risk presented in a particular case justifies caution, the Consortium's advice regarding the use of options (3) and (4) will be strict, based on a number of factors relating to technical and legal constraints, as well as the possibility of external attacks bearing in mind the nature of the data intended for sharing and its potential sensitivity. Furthermore, as described in section 5, this advice will be backed up by the 10-point strategy approach where pseudonymised data is shared.

Finally, at no time will any data relating to individuals be released (in whole or part) in the public domain under the auspices of the Project except in accordance with data protection law. Moreover, a reasonable level of documentation will be required to demonstrate that adequate security controls in place.

---

[42] Under option (2), the Consortium also recommend that access to data relating to persons only be given to authorised persons at the Participating SMEs, as well as the adoption of passwords and other relevant further security mechanisms against data leaks or intrusion, such as access solely through APIs and encryption. To note, the ICO Blog: Why encryption is important to data security provides a useful discussion on how encryption works and the importance of identifying suitable forms of encryption techniques, bearing in mind the sensitivity of the data and how it will be stored and processed. It cites - by way of example - that the encryption techniques used for data storage will differ to those used for transferring personal data. To note, as security, like privacy, is an on-going process, not a one-time project, it is conceivable that a Data Provider will be required to re-evaluate access conditions during the course of the Project.

[43] The Art.29 WP has made a number of recommendations, which it describes as "a checklist for data protection compliance by cloud clients and cloud providers" (Opinion 5/2012 on Cloud Computing, p.19). It explains that these arise primarily from the fact that cloud computing clients relinquish exclusive control of their data, which can lead to: lack of portability and interoperability; loss of integrity of the data due to sharing of cloud resources; disclosure of data to law enforcement agencies outside the EU in contravention of EU data protection principles; loss of ability to intervene owing to the complex chain of outsourcing; inability of the cloud provider to help the controller respond to data subjects' requests; possibility that the cloud provider might link personal data from different clients; and, lack of transparency about how the cloud service intends to process data, so that the controller cannot take proper measures to ensure data protection compliance. The danger is considered increased if the client remains unaware that the cloud service involves a chain of processors (each subcontracting to the next), that processing is being conducted in different countries (which will determine the applicable law in the event of a dispute), or that the service will result in the transfer of data to countries outside the European Economic Area. Related useful guidance has been produced by the ICO here: https://ico.org.uk/for-organisations/guide-to-data-protection/principle-7-security/.

6.1.7 Intellectual property law considerations

To be completed in version 2.0 of the toolkit.

6.1.8 Case studies

To be completed in version 2.0 of the toolkit.

6.2 Data reuse methodology for Participating SMEs – managing the legal risks associated with data reuse in practice

Participating SMEs will be required to show understanding of the different types of legal issues that might arise in relation to the data that is to be shared with them, as well as sign a legal agreement with the Consortium obliging them to observe conditions for the data's reuse in light of legal compliance and setting out the scope of the research purposes for which each dataset will be processed (and only processed) under the Project.

Set out below are the Project's best practice recommendations in respect of this exercise. In this first version of the toolkit, they focus on data protection law compliance and related risk mitigation.

6.2.1 Data protection law compliance

Steps will be taken by the Consortium – via this toolkit, associated support, and the legal agreement that they will sign – to raise Participating SMEs' awareness of EU data protection law requirements in case it becomes relevant.

Furthermore, the Consortium will require all SMEs applying to take participate in the Project to sign an Ethics Statement and Declaration of Honour as part of the application process in line with H2020 guidelines. These will refer to the principles that apply to the processing of personal data that must be followed. Such declarations will also form part of the contract signed with the Consortium by those Participating SMEs who are successful in their applications, which can later be checked during the milestone reviews or in the final graduation.

6.2.1.1 Data controllers or data processors?

Data relating to persons shared by the Data Providers with Participating SMEs must first be subject to anonymisation/pseudonymisation techniques. However, there is a risk that even data so modified may yet still be deemed personal data.

For the avoidance of any doubts, in case data protection law applies to the processing of such data carried out during the acceleration stage, the Participating SMEs must acknowledge that they may be construed as data processors,[44] but also potentially – in the alternate - as joint data controllers of any personal data shared with them. Therefore, the responsibilities of both processors and controllers must be understood.

Where a participating SME is acting in a data processor role for a Data Provider, the Data Provider will be obliged under the GDPR to enter into a contract directly with them to ensure that they follow their data protection law obligations. For example, SMEs must offer sufficient guarantees to implement appropriate technical and organisational measures when they may be dealing with personal data.[45]

Higher standards of data protection are required when the Participating SME would be acting in the role of data controller. This role may be assumed, in the context of the Project, in relation to any personal data that they are permitted (by the Consortium) to analyse alongside the shared data during the acceleration stage. To this end, Participating SMEs should consider whether they have assessed the legal risks (and compliance obligations beholden on them) associated with any data relating to persons that they currently hold. In particular, subject to some exceptions, organisations currently processing personal data should be registered (publicly) with the relevant data protection authority (such as the ICO in the UK) first. Details of such

[44] To note, unlike existing law, the GDPR introduces direct compliance obligations for processors. Whereas under the Data Protective Directive processors generally are not subject to fines or other penalties, under the GDPR they may be found liable to pay fines (including, as mentioned, fines of up to €20 million or 4% of global annual turnover for the preceding financial year, whichever greater, where they are directly responsible for a data protection law breach).

[45] Article 28(1), GDPR states: "*Where processing is to be carried out on behalf of a controller, the controller shall use only processors providing sufficient guarantees to implement appropriate technical and organisational measures in such a manner that processing will meet the requirements of this Regulation and ensure the protection of the rights of the data subject".*

registration will be required by the Consortium within the legal agreement entered into between the Consortium and the Participating SMEs.

6.2.1.2 Key issues for consideration

The processing of personal data must comply with the data protection law principles set out in section 4 above and with the contractual terms set out in the legal agreement signed with the Consortium. These include conditions regarding what Participating SMEs are allowed to do in the reuse of data shared with them, such as the fact that the data shared under the Project must only be processed for the purposes of the Project. Furthermore, Participating SMEs must agree not to attempt to identify any data shared with them that has been subject to anonymisation/pseudonymisation processes.

Any data relating to persons will also only be shared within the context of specific data experiments where the broad objectives for the secondary processing has been scoped by a defined purpose or purposes (a challenge). These challenges will not require the Participating SMEs to take any decisions or measures regarding any particular individuals, and this activity should be avoided at all costs. If such decisions or measures were taken, the Participating SMEs will be required to inform the Consortium immediately.

It should be noted that, any data controller intending to carry out processing activities which are likely to result in a "*high risk*" to data subjects (taking into account the nature, scope, context and purposes of the processing), should carry out a DPIA. In certain circumstances a Participating SME may be a data controller and in that case the requirement for a DPIA should be assessed carefully. To do this a consideration of Article 35(3)(a) GDPR should be made which requires data controllers to conduct a DPIA any time "*a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person*."

Any processing by Participating SMEs of personal data would also need to be justifiable in light of a specified lawful purpose (legal basis) for its reuse, at least where such use would be deemed incompatible with the original purposes for which personal data was processed. Furthermore, all appropriate safeguards for the rights and freedoms of the data subject should be put in place, including technical and organisational measures proportionate to the processing aim pursued, to safeguard the fundamental rights and interests of the data subject.

Finally, all Participating SMEs will also be required to enter into a confidentiality agreement, either separately or as part of the legal agreement with the Consortium whereby they will not be permitted to disclose any data identified as confidential that is shared with them.

6.2.1.3 Managing data security

Assessing the level of security required

As previously stated, organisations should adopt a risk-based approach to deciding what level of security is needed when they process personal data. Thus, Participating SMEs acting as data controllers or data processors must ensure a level of security appropriate to both:

- the harm (damage or distress) that might result if there was a data breach (the unauthorised or unlawful processing or accidental loss, destruction or damage of personal data); and,
- the nature of the personal data held (for example, how valuable, sensitive or confidential it is).

To assess this, Participating SMEs should carry out risk assessments by way of a formal process to identify, document and manage security risks.[46] For example, such assessments would typically also include a review of the purposes for which personal data are processed internally, or to be shared via external mechanisms, as well as taking into account:

- The nature and extent of an organisation's premises and computer systems.
- The number of staff and the extent of their access to personal data.
- The state of technical development.
- The cost of implementing security measures.

Managing security risks

Once security risks have been assessed, Participating SMEs should take reasonable steps to mitigate significant risks. This requires ensuring, not only the correct physical infrastructure and technical tools, but also that appropriate data security policies and procedures are in place. To this end, Participating SMEs that are also data controllers are obliged to take reasonable steps to ensure the reliability of any employees who have access to personal data. In addition, Participating SMEs should identify internal individuals who are responsible for ensuring data security.

As part of any general registration requirement under data protection law, Participating SMEs that are also data controllers may be required to notify their national data protection agency of the security measures that they have put in place. This requirement for registration would be limited to one year. The Participating SMEs should also be able to respond to any data security breach swiftly and effectively.

Finally, further obligations are also placed on Participating SMEs that are also data controllers and who employ a data processor. As mentioned, they must choose a data processor providing sufficient guarantees in respect of the technical and organisational security measures governing the processing to be carried out, and take reasonable steps to ensure compliance with those measures. However, it is unlikely although not impossible for Participating SMEs to utilise a data processor as part of the Project.

6.2.2 Intellectual property law considerations

To be completed in version 2.0 of the toolkit.

6.2.3 Case studies

To be completed in version 2.0 of the toolkit.

[46] An example is the risk assessment framework has been published by NIST (NIST 800-30: Risk Management Guide for Information Technology Systems), which sets out a methodology to evaluate and govern risk.

7 Key messages at a glance: a quick checklist to help confirm whether data sharing and reuse is lawful

What to consider when you are sharing and reusing data:

- Make sure you check the legal agreement you have signed. In the case of Participating SMEs, check the nature of the limitations included which could reduce your planned re-usage of the data being shared with you and in particular, the allocation of intellectual property rights between the data provider and the data recipient. One key distinction to bear in mind in order to adequately allocate intellectual property rights is the distinction between the algorithm produced to process the shared data and the output of the processing activity through the means of the algorithm, which could be described as enriched or output data.
- Make sure that the data does not allow individuals to be identified if combined with other information available to you. To note, it is not always necessary to have biographical details in data - such as a requirement for an individual to be named – for it to be deemed personal data. Identification (and re-identification) may be effected in other ways, notably through singling out individuals from others in a dataset via indirect identifiers.
- The risk of re-identification through data linkage is unpredictable because it can never be assessed with certainty what data are already available or what data may be released in the future. On the one hand, while it may not be possible to determine with absolute certainty that no individual will ever be identified as a result of the disclosure of anonymised data, a certain amount of pragmatism needs to be adopted. It involves more than making an educated guess that information is about someone.
- The likelihood and severity of re-identification risk occurring can also change over time (e.g. with the introduction of new technologies that can link data) and, therefore, re-assessments should be carried out periodically and any new risks managed. This should include trying to determine what additional information - personal data or not – could become available that could be linked to the data to result in re-identification.
- General and data-specific safeguards are set out in the legal agreements governing the reuse of data shared. Above and beyond fulfilling their contractual obligations, there is no ‘one size fits all’ solution to data security – each organisation should adopt a risk-based approach to its assessment and management in conjunction with implementing advice given by the Consortium. Different measures (and combinations of measures - legal, technological, and organisational) may be appropriate depending on the processing activity and other data environment contextual factors.
- Any risk of harm to individuals who are the subjects of anonymised/pseudonymised data should be avoided. This includes consideration of harm that might result from accidental loss or unauthorised processing by others of data shared under the Project, bearing in mind the nature of the data being processed and any possible sensitivity.
- Any breach of data protection law must be reported immediately to the Consortium and the relevant Data Provider must act accordingly as soon as possible to mitigate any risk of harm to data subjects.

8 A note on data ethics

Compliance with data ethics is complementary to the rest of the toolkit. Requirements for ethical conduct under the Project are also enshrined in an Ethical Statement that the Participating SMEs will be required to sign in advance of taking part, as well as a Declaration of Honour requiring self-disclosure of matters such as past professional misconduct - as well as a guarantee of future compliance with ethical and legal principles under the Project. Copies of these documents can be found in **Annexes E and F** below.

All participants in the Project are expected to consider ethics issues throughout its lifecycle - before, during, but also after the acceleration period in relation to knowledge exchange and impact activities (such as reporting and publication). While such issues may often overlap with legal obligations upon participants (such as in relation to data confidentiality), they should also be considered in their own right.

Below are a set of core ethical principles providing general guidance in relation to the carrying out of research:[47]

- research should aim to maximise benefit for individuals and society and minimise risk and harm;
- the rights and dignity of individuals and groups should be respected;
- wherever possible, participation should be voluntary and appropriately informed;
- research should be conducted with integrity and transparency;
- lines of responsibility and accountability should be clearly defined; and
- independence of research should be maintained and where conflicts of interest cannot be avoided they should be made explicit.

Regarding the ethics of data sharing, specifically, the following should be considered over and above previously-discussed considerations of data anonymisation/pseudonymised with respect to data relating to persons. In other words, such considerations apply to data that does not relate to persons, as well as data that does:

- Does the relevant data contain confidential or sensitive (personally or commercially) information?
- If so, have you discussed data sharing with those from whom you collected the data?
- Have you gained consent from relevant others with an interest in that data if you plan to share it outside the remit of the Project?
- Was that consent informed (i.e. adequately descriptive of what you wanted to do with the data)?
- Did any consent obtained relate to a single purpose or multiple purposes of re-usage in the future clearly defined?
- Is the relevant data being safely handled at all times?
- Is access control required now and, if this is an on-going requirement, how will this be managed?
- If you are required, or deem it necessary, to carry out a DPIA in respect of processing activities that might be carried out in association with the Project, might ethical considerations also be taken into account in assessing potential impact?

[47] Economic and Social Research Council, Six Key Principles for Ethical Research, available at: http://www.esrc.ac.uk/funding/guidance-for-applicants/research-ethics/our-core-principles/

References

[1] EU Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data (Data Protection Directive), available at http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:31995L0046

[2] 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), available at http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf

[3] Council Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector, as amended in 2009 by the Citizens' Rights Directive 2009/136/EC (E-Privacy Directive), available at http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:en:HTML

[4] Proposal for a Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications) COM(2017) 10 final (draft E-Privacy Regulation), available at http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=41241

[5] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (the Database Directive), available at http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31996L0009

[6] Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market COM(2016)593 (the draft Copyright Directive), available at http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=17200

[7] Article 29 Working Party Opinion, Opinion 03/2013 on purpose limitation (WP203), issued 2 April 2013, available at http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

[8] Article 29 Working Party Opinion, Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is “likely to result in a high risk” for the purposes of Regulation 2016/679 (WP248), issued 4 April 2017, available at http://ec.europa.eu/newsroom/document.cfm?doc_id=44137

[9] Article 29 Working Party Opinion, Opinion 05/2014 on anonymisation techniques (WP216), issued April 2014, available at http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[10] UK Information Commissioner’s Office, Anonymisation: Managing Data Protection Risk Code of Practice, issued November 2012, available at https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf

Annex A EU Article 29 Working Party Guidelines on DPIAs (2017)

In April 2017, the Art.29 WP adopted an initial version of guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679 (the GDPR).

A DPIA is mandatory under the GDPR (Article 35) where processing is likely to result in a 'high risk' to the rights of individuals and is particularly relevant where new data processing technology is being introduced.

The main body of the guidelines cover the following issues in great detail and also provide recommendations:

- What does a DPIA address?
- Which processing operations are subject to a DPIA?
- How to carry out a DPIA?
- When shall the supervisory authority be consulted?

In particular, the guidelines offer a general rule of thumb that processing operations meeting at least two of a list of criteria will require a DPIA, albeit a processing operation meeting only one criterion may require a DPIA depending on the circumstance. The criteria are as follows:

- Are you doing evaluation or scoring (including profiling and predicting) of aspects specific to the data subject?
- Does the processing involve automated decision making that produces significant effect on the data subject?
- Are you performing systematic monitoring of data subjects, including in a publicly accessible area?
- Does the processing involve sensitive data (special categories of data as defined in Article 9 GDPR and data regarding criminal offences)?
- Is the data being processed on a large scale?
- Have datasets been matched or combined ("*for example originating from two or more data processing operations performed for different purposes and/or by different data controllers in a way that would exceed the reasonable expectations of the data subject*").
- Does the data concern vulnerable data subjects (as laid out in Recital 75 GDPR)?
- Is this an innovative use or does it apply technological or organizational solutions (for example, combining use of finger print and facial recognition)?
- Are you transferring data outside the European Union?
- Will the processing itself prevent data subjects from exercising a right or using a service or a contract?

The WP is happy to leave it to controllers to choose the exact form and structure of each DPIA to fit within each controller's existing working practices, as long as they align their analyses with established impact assessment methodologies. At a minimum, however, the WP specifies four features that a DPIA must include that will help assess the level of likely impact to result from a processing activity:

- A systematic description of the envisaged processing operations, the purposes of the processing and, if applicable, the legitimate interests pursued by the controller;
- An assessment of the necessity and proportionality of the processing operations in relation to such purposes;
- An assessment of the risks to the rights and freedoms of data subjects; and,

- The measures envisaged to address such risks (compare, Recital 84 GDPR: "*The outcome of the assessment should be taken into account when determining the appropriate measures to be taken in order to demonstrate that the processing of personal data complies with this Regulation*").

Annex 1 to the guidelines contain a list of links to examples of guidance on existing PIA/DPIA frameworks (both generic and sector-specific), and to international standards containing DPIA methodologies, for reference. These include the following:

- Commission Recommendation of 12 May 2009 on the implementation of privacy and data protection principles in applications supported by radio-frequency identification. https://ec.europa.eu/digital-single-market/en/news/commission-recommendation-12-may-2009-implementation-privacy-and-data-protection-principles
- Opinion 9/2011 on the revised Industry Proposal for a Privacy and Data Protection Impact Assessment Framework for RFID Applications. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2011/wp180_en.pdf
- Opinion 07/2013 on the Data Protection Impact Assessment Template for Smart Grid and Smart Metering Systems prepared by Expert Group 2 of the Commission's Smart Grid Task Force. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp209_en.pdf
- Conducting privacy impact assessments code of practice, Information Commissioner's Office (ICO), 2014. https://ico.org.uk/media/for-organisations/documents/1595/pia-code-of-practice.pdf
- ISO/IEC 29134 (project), Information technology – Security techniques – Privacy impact assessment – Guidelines, International Organization for Standardization (ISO). https://www.iso.org/standard/62289.html

Annex 2, by comparison, sets out the criteria for an acceptable DPIA with reference to the relevant GPDR provisions:

- a systematic description of the processing is provided (Article 35(7)(a)):
  - nature, scope, context and purposes of the processing are taken into account (Recital 90);
  - personal data, recipients and period for which the personal data will be stored are recorded;
  - a functional description of the processing operation is provided;
  - the assets on which personal data rely (hardware, software, networks, people, paper or paper transmission channels) are identified;
  - compliance with approved codes of conduct is taken into account (Article 35(8));
- necessity and proportionality are assessed (Article 35(7)(b)):
  - measures envisaged to comply with the Regulation are determined (Article 35(7)(d) and Recital 90), taking into account:
    - measures contributing to the proportionality and the necessity of the processing on the basis of:
      - specified, explicit and legitimate purpose(s) (Article 5(1)(b));
      - lawfulness of processing (Article 6);
      - adequate, relevant and limited to what is necessary data (Article 5(1)(c));
      - limited storage duration (Article 5(1)(e));
    - measures contributing to the rights of the data subjects:

- information provided to the data subject (Articles 12, 13 and 14);
- right of access and portability (Articles 15 and 20);
- right to rectify, erase, object, restriction of processing (Article 16 to 19 and 21);
- recipients;
- processor(s) (Article 28);
- safeguards surrounding international transfer(s) (Chapter V);
- prior consultation (Article 36).

- risks to the rights and freedoms of data subjects are managed (Article 35(7)(c)):
  - origin, nature, particularity and severity of the risks are appreciated (cf. recital 84) or, more specifically, for each risk (illegitimate access, undesired modification, and disappearance of data) from the perspective of the data subjects:
    - risks sources are taken into account (recital 90);
    - potential impacts to the rights and freedoms of data subjects are identified in case of illegitimate access, undesired modification and disappearance of data;
    - threats that could lead to illegitimate access, undesired modification and disappearance of data are identified;
    - likelihood and severity are estimated (recital 90);
  - measures envisaged to treat those risks are determined (Article 35(7)(d) and recital 90);
- interested parties are involved:
  - the advice of the DPO is sought (Article 35(2));
  - the views of data subjects or their representatives are sought (Article 35(9)).

Annex B Data Provider Questionnaire

The datasets to be provided are likely to span many different types of data types, including proprietary data, associated with various levels of sensitivity.

Therefore, we would like some brief information about the type of data that your organisation would be happy to provide to the project, and possible degrees of data sensitivity in that overall data bundle. Using this information we can better assess what types of risks might arise and need to be managed.

ABOUT THE DATASETS YOU ARE CONSIDERING SHARING

1. What type and format of data would you be happy to share under the project?
2. How were these datasets created/collected?
3. What do they relate to (e.g. sector, product/service, territory, etc.)?
4. In relation to each dataset you would be happy to share under the project, would you currently classify them as (i) '**closed'** (internal access only); (ii) '**shared'** (third party access only); or (iii) '**open**' (where anyone can access).
5. Have you any constraints/strong will on data storage (e.g. in relation to where you would prefer data for sharing under the project to be stored, and around data access)?

To note, our data hosting and experimentation facilities are designed to manage datasets of all types, and will provide clear details about which organisation owns each dataset, usage restrictions, and the specific security level required for it.

In this context, we propose using pre-defined rule sets to specify the use, sharing, and anonymisation requirements (or pseudonymisation processes), of each dataset based on its sensitivity and level of openness as preferred by your organisation.

LEGAL OVERVIEW

Please note that any organisation that is considering entering into a data sharing arrangement with us must also do the following in respect of such data to be provided, and in respect of any national legal systems (including EU law) that apply:

- Highlight any regulatory requirements that might arise upon such data disclosure. In particular, this includes compliance with data protection law (for guidance, see e.g. the UK's Information Commissioner's Office Data Sharing Code of Practice).
- Highlight whether any intellectual property rights apply (for example, copyright restrictions, or data originally provided under a duty of confidence) that might arise upon such data disclosure.
- Highlight whether any other legal issues might arise upon such data disclosure (for example, because it is subject to contractual legal restrictions governing its usage).

Therefore:

- **If any of the data is likely to contain personal data, please answer the questions in the next section compliance with data protection law.**
- **If intellectual property (IP) rights are likely to apply to any of the data, please answer the questions in the following section regarding IP law.**
- **If you think that any other legal issues might become relevant if the data were shared as envisioned under this project, please answer the questions in the final section regarding contractual restrictions upon data usage, or any other possible legal issues that might be raised upon disclosure.**

DATA PROTECTION AND PRIVACY LAWS

Please provide details as follows:

1. Does any of the data contain personal data? In other words, according to EU law, does it contain *"any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person"*?

Guidance: In relation to determining whether someone is identifiable from data (e.g. where it is not immediately obvious because the names and other identifiers have been removed), the key question for consideration is whether there are means that could likely reasonably be used to identify that person. In particular, under the new EU General Data protection Regulation (GDPR) coming into effect on 25 May 2018, the following guidance is provided (Recital 26, our emphasis regarding important, new changes from the existing EU Data Protection Directive):

> *"The principles of data protection should apply to any information concerning an identified or identifiable natural person.* ***Personal data which have undergone pseudonymisation, 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****. ['Pseudonymisation' is defined at Article 4(5) as, "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or 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. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.* *The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes".*

If no to question 1, please answer question 10 below.

If yes, to question 1, please answer questions 2-9 below:

2. Please describe the type of personal data (e.g. names, addresses, telephone numbers, or financial information?) and how it was collected.
3. Where are the personal data currently stored and processed (including, if a third party is relied upon for these activities, where are they based)?
4. At a national level, what country's data protection laws applies to the processing of such personal data?
5. Under the Project, we want to mitigate the risks that personal data would be shared. Therefore, what specific measures will you put in place to mitigate the risk of identification from the data? Please give details. Would the data still retain individual-level elements – such that the data subject could still be singled out from the data via indirect identifiers?

6. Was data subject consent obtained when such personal data was gathered, or was another legal basis used to justify its collection,[48] and for what purpose?
7. If consent was obtained at the initial point of personal data collection, was consent also obtained to re-use that data for other purposes (e.g. for research purposes)?
8. If the personal data was collected from a third party, did that third party give a warranty or other appropriate form of assurance as to their compliance with data protection/privacy laws?
9. Does any of this personal data contain or imply sensitive personal data? In other words, according to EU law, does it contain information about the data subject's "*racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation*"? If yes, please give details.
10. If the data you wish to share is not believed to be personal data (but it does relate to persons – that is, it refers to them, is used to determine or influence the way in which an individual is treated, or is likely to have an impact on their rights and interests), please specify why you think that those persons are not identifiable from the data. We are particularly interested in this answer if the data contains individual-level elements – such that individuals could still be singled out from the data by the SMEs analysing the data under the Project?

INTELLECTUAL PROPERTY (IP) LAWS

IP is an umbrella term which is used to describe a range of legal rights that attach to certain types of information and ideas and to their particular forms of expression. IP rights fall into two general categories:

- **Registered rights** that are granted on application to an official body, and include **patents, registered trademarks and registered designs**; and,
- **Unregistered rights** that arise automatically, give protection against copying or using the right, and include **copyright, database rights, unregistered design rights, rights in unregistered trademarks and confidential information**.

Please provide details of the following, including the relevant legal jurisdiction:

1. Any IP rights applying to the data you would be happy to share under the project. If so, please specify what type they are and what they apply to.
2. Ownership of such IP rights (including territory of right)
3. Any registrations of such IP rights (including the territory of registration and the relevant market of registration)
4. Any anticipated difficulty in granting licences (or sub-licences) to use such IP rights under the project (for example, on-going litigation)?

[48] The list of legal grounds that can justify the processing of personal data under the GDPR (Article 6) are: ***(a)*** *the data subject has given consent to the processing of his or her personal data for one or more specific purposes;* ***(b)*** *processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;* ***(c)*** *processing is necessary for compliance with a legal obligation to which the controller is subject;* ***(d)*** *processing is necessary in order to protect the vital interests of the data subject or of another natural person; (e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;* ***(f)*** *processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.*

5. Whether any of the data you would be happy to share under the project was given in confidence. Under UK law, for example, a common law duty of confidence will arise where the information in question has a quality of confidence about it and was given in circumstances importing an obligation of confidence.

CONTRACTUAL RESTRICTIONS AROUND DATA USAGE, OR ANY OTHER LEGAL ISSUES THAT MIGHT ARISE UPON DATA PROVISION AND SUBSEQUENT USAGE

1. Please provide details of any contractual restrictions around the usage of the data that you would be happy to provide under the project, or any other possible legal issues related to such data usage, and specify the relevant territory. For example, if the data has been provided to your organisation by another party, please provide a brief overview of the contractual terms or service level agreements (SLAs) used for that purpose.
2. Are there any strong requirements you have regarding the further usage of any of your data sets dictated by law or for other reasons? What are these, in relation to what data, and what usage?

Annex C EU Article 29 Working Party Opinion on Anonymisation Technologies (2014)

The Art.29 WP adopted an Opinion on anonymisation techniques in the context of EU data protection law in April 2014. It discusses the anonymisation definition in the EU Data Protection Directive (95/46/EC) and references in the E-Privacy Directive (2002/58/EC) to anonymisation. The Opinion examines current anonymisation techniques and makes recommendations on using them in the light of the risk of re-identification of individuals.

Although the Art.29 WP's opinions are not binding on the member states' national data protection authorities, they are normally drafted and approved by a representative of those authorities and thus tend to represent an EU-wide consensus on how the Data Protection Directive should be interpreted and enforced at national level.

The Opinion describes anonymisation as a technique applied to personal data to achieve irreversible de-identification. The Art.29 WP notes that anonymisation may be a useful means to maintain the benefits of open data in society, while protecting individuals by mitigating privacy risks. However, it warns of the difficulty of creating a truly anonymous dataset where much underlying information is retained, as combining an anonymised dataset with another dataset may lead to re-identification.

The Art.29 WP discusses randomisation and generalisation (the main anonymisation techniques). Noise addition, permutation, differential privacy, aggregation, k-anonymity, l-diversity and t-closeness are also examined. Strengths and weaknesses of techniques are highlighted along with common mistakes and failures, which should assist data controllers with designing an anonymisation process.

The Opinion clarifies that pseudonymisation (one element in a record being replaced for another element, such as a name being replaced by a number) is *not* a method of anonymisation, but merely a useful security measure. To explain its thinking on this point, the Art.29 WP states its belief that data will remain regulated personal data in the event *any party* is capable of associating it with a living individual (using reasonable means). Therefore, according to the Opinion:

- Removing directly or indirectly identifying elements in itself is not enough to ensure that identification of the data subject is no longer possible. In cases where it is not immediately obvious whether a person can be identified, the question whether or not the individual is nevertheless identifiable will depend on the means the data controller is likely reasonably to use to identify that person.
- Data controllers cannot render data anonymous if it is retained at an "event-level", a term the Art.29 WP uses to contrast with data held on an aggregated basis, provided they keep a copy of the raw data. By contrast, an effective anonymisation solution prevents all parties from singling out an individual in a dataset, from linking two records within a dataset (or between two separate datasets) and from inferring any information in such dataset.
- Unless the original data is destroyed, the data controller continues to have the ability to attribute an element to a relevant individual, either directly or by inference. The inability of the recipient of the anonymised data to do this and/or the intentions of the data controller are irrelevant; the data remains personal data. [To note, this statement by the Art.29 WP suggests that anonymised should always be deemed personal data as long as the raw data exists in the data controller's possession. In other words, the Art.29 WP rejects a relativistic approach to the concept of identifiability.[49] On the other hand, there are statements by the Art.29 WP elsewhere that appear to contradict this approach: in particular, endorsing the view that anonymised data may be deemed non-personal data for a recipient of it, even

[49] For the meaning of this term, and related discussion, see https://peepbeep.wordpress.com/2016/06/17/mind-the-caveats-cjeu-advocate-general-opines-that-dynamic-ip-addresses-can-be-personal-data-sometimes/.

if that is not the case for the original data controller who oversaw its anonymisation. The Consortium adopts this latter view in believing that anonymised data could be considered non-personal data under the 'means test' even if the original data controller retains the raw data. This is in line with UK case-law.]

The Opinion advises that anonymisation should be planned on a case-by-case basis, using a variety of techniques and factoring in the Opinion's recommendations. It will often be necessary to take additional measures to prevent identification, once again depending on the context and purposes of the processing for which the anonymised data are intended. The Art.29 WP seems to exclude reliance solely upon contractual measures to make data anonymous vis-a-vis the recipient.

Data controllers are advised not to treat anonymisation as a one-off exercise. Rather, regular risk assessment should continue in the light of the residual risk of identification. Even if the data received by the recipient is truly anonymous now, does not mean that it will always remain anonymous. The data controller entity will have to take into account the context and circumstances of its processing operations to evaluate the risk of future identifiability, for example, because of associations or connections that may be forged with new, 'anonymous' data sets.

Annex D UK Anonymisation Code of Practice (2012)

In 2012, the ICO published a Code of Practice, 'Anonymisation: managing data protection risk' (the Code). The Code explains the data protection implications of anonymising personal data, together with the ICO's recommendations about anonymising personal data and assessing the risks associated with producing and publishing anonymised data. In the event of the ICO investigating an issue arising from the anonymisation of personal data, its officials will take the good practice advice in the Code into account.

In the Code, the ICO addresses the question of what it regards should be taken into account in assessing whether or not a person is re-identifiable from anonymised data under the DPA and, within this assessment, it indicates factors such as the cost-effectiveness of any available means that would enable re-identification in light of new technological developments or changes to the public availability of certain records. It also refers to additional factors deemed relevant for helping to make this determination:

> *"The test in the DPA for determining whether information relating to a living individual is personal data is based entirely on the identification or likely identification of the individual. The risk posed to individuals by disclosure, or the public benefit of this, are not factors that the DPA allows to be taken into account when determining whether or not information is personal data. In reality though, some types of data will be more attractive to a motivated intruder than others and more consequential for individuals. In reality these factors should also inform an organisation's approach to disclosure. Clearly the identification of an individual can have a range of consequences depending on the nature of the data, the context in which it is disclosed and who it is about. The Information Commissioner would certainly be more concerned about a disclosure of personal data that is detrimental to an individual, than about an inconsequential one. The Information Commissioner will take the effect or potential effect into account should a case of reidentification or inappropriate data disclosure come to his attention. In borderline cases where the consequences of re-identification could be significant eg because they would leave an individual open to damage, distress or financial loss, organisations should: seek data subject consent for the disclosure of the data, explaining its possible consequences; adopt a more rigorous form of risk analysis and anonymisation. In some scenarios, data should only be disclosed within a properly constituted closed community and with specific safeguards in place. In some particularly high-risk situations, it may not even be possible to share within a closed community."*[50]

In other words, the ICO believes it is good practice, when releasing anonymised data, to try to assess "*what the consequences of re-identification are likely to be, if any, for the data subject concerned*".[51] Notwithstanding, the ICO acknowledges that assessing likely consequences, "*can be difficult to assess in practice and a member of the public's sensitivity may be different from yours*".[52] To carry out the assessment, the ICO has put forward a test based on the existence of a 'motivated intruder'. The test essentially involves considering whether someone would be able to achieve re-identification from anonymised data if motivated to attempt this. The approach assumes that the 'motivated intruder' is competent, diligent and has access to resources commensurate with the motivation the intruder may have for the re-identification. In other words, would a person who starts without any prior knowledge but who wishes to identify an individual be able to access resources and investigative techniques to de-anonymise the data? The motivated intruder is not, however, assumed to resort to criminality or have specialist equipment or skills.

A difficult technical issue for organisations will be whether anonymised data could be combined with information by a third party to re-identify the individual To this end, data controllers should assess whether an

[50] The Code, page 20.

[51] The Code, page 25.

[52] Ibid.

individual can be reasonably re-identified from the anonymised data to be released for reuse – either in itself or in combination with other available information. The ICO's position is that the risk of identification "must be greater than remote, and reasonably likely" in order for the data to be considered to be personal data for the purpose of falling under the DPA. That assessment should form the basis for determining which data it is safe to release. To note, the ICO considers that pseudonymised data - where individuals are distinguished by the use of a unique identifier which does not reveal their real identity - pose a high level of re-identification risk.

Annex E Declaration of Honour for Participating SMEs to sign

1. As legal representative of [insert legal entity name], I declare that the entity is not:
   a) bankrupt or being wound up, is having its affairs administered by the courts, has entered into an arrangement with creditors, has suspended business activities, is the subject of proceedings concerning those matters, or is in any analogous situation arising from a similar procedure provided for in national legislation or regulations;
   b) having powers of representation, decision making or controlling personnel being convicted of, or having been convicted of an offence concerning their professional conduct by a judgment which has the force of res judicata;
   c) having been guilty of grave professional misconduct proven by any means which the contracting authority can justify including by decisions of the European Investment Bank and international organisations
   d) failing to be compliant with obligations relating to the payment of social security contributions or the payment of taxes in accordance with the legal provisions of the country in which it is established or with those of the country of the contracting authority or those of the country where the contract is to be performed;
   e) having powers of representation, decision making or controlling personnel having been the subject of a judgment which has the force of res judicata for fraud, corruption, involvement in a criminal organisation or any other illegal activity, where such illegal activity is detrimental to the Union's financial interests;
   f) subject to an administrative penalty for being guilty of misrepresenting the information required by the contracting authority as a condition of participation in a grant award procedure or another procurement procedure or failing to supply this information, or having been declared to be in serious breach of its obligations under contracts or grants covered by the Union's budget.
2. I declare that the natural persons with power of representation, decision-making or control over the aforementioned legal entity are not in the situations referred to in b) and e) above.
3. I declare that I
   a) am not subject to a conflict of interest and will take all reasonable measures to prevent any situation where the objectives of the Data Pitch project might be compromised due to undeclared shared interests;
   b) have not made false declarations in supplying the required information to the project formally detailed as Data Pitch, and have not failed to supply the required information;
   c) am not in one of the situations of exclusion, referred to in the abovementioned points a) to f).
4. I certify that I:
   a) am committed to participate in the aforementioned project as part of the legal entity detailed above;
   b) have stable and sufficient sources of funding to maintain its activity throughout its participation in the aforementioned project, and will provide any counterpart funding necessary;
   c) have or will have the necessary resources as and when needed to carry out its involvement in the above mentioned project.
   d) will comply with my responsibilities and obligations under the Data Pitch project, including those set out in the Data Sharing Agreement.

e) will respect any third party rights in relation to data provided for processing under the Data Pitch project.

f) will abide by international, EU and national laws and regulations that might apply to the substance, or outcome, of data sharing arrangements as relevant to activities that I/my entity will be involved in under the Data Pitch project.

g) will not share or disseminate data received through the Data Pitch project without the explicit prior consent of the data provider and any others with proprietary rights in relation to that data.

h) will take all reasonable measures to safeguard data provided to me/my entity for use in the Data Pitch project against possible misuse and unauthorised access.

i) will abide by international, EU and national laws imposing privacy and data protection requirements (including, in anticipation for its coming into effect, the *General Data Protection Regulation* **(***GDPR***)** (Regulation (EU) 2016/679)) as relevant. In particular, personal data shared under the Data Pitch project will not be re-used for purposes outside the project without the explicit prior consent of the data controller.

j) will act in good faith as far as reasonably possible under the Project and fully apply the principles of the Ethics Statement.

5. I declare that, to the best of my knowledge, I am eligible to apply for the Data Pitch call and all the information I have provided is true.

Name	
Signature	
Date	

Annex F Ethics Statement for Participating SMEs to sign

This Ethics Statement underpins the Data Pitch project in setting out specific rules and standards of conduct expected from recipients of Data Pitch funding. Ethical conduct means acting consistently in a way that is ethical and fair and encouraging others to do likewise.

The standard of behaviour expected is additional to compliance with relevant legal rights and obligations arising automatically by virtue of law applying to each participant. It is also not intended to exclude or replace responsibilities agreed under contract with the Data Pitch consortium (in case your application is successful), as well as the certifications/declarations set out in the Declaration of Honour.

As legal representative of [insert legal entity name], I certify that [insert legal entity name] will adhere to the following principles as far as reasonably possible under the Data Pitch project:

1. act in good faith;
2. respect human rights;
3. ensure research quality and integrity;
4. be able to show that our findings are independent and non-discriminatory to any groups of individuals;
5. not misrepresent credentials;
6. demonstrate authenticity and validity of authorship;
7. respect confidential information;
8. secure any confidential information provided to prevent its misuse or unauthorised access;
9. only share confidential information where necessary and only where the prior informed consent of anyone potentially affected by the disclosure of such information has been received;
10. respect the privacy of any people identified from the findings of the Data Pitch project as far as possible;
11. avoid any conduct that may cause anyone harm, and seek relevant individuals' informed consent for any activities that might affect them directly;
12. determine the applicable laws that may apply to our activities under the Data Pitch project and plan our activities in accordance with such laws as early as possible;
13. not collect or otherwise process any personal or sensitive data not essential for our Data Pitch activities;
14. be fully transparent to the Data Pitch consortium about the purpose, methods and intended possible uses of our Data Pitch activities, and what risks, if any, are involved.
15. seek advice promptly from the Data Pitch consortium where we believe ethical and/or legal risks may be raised by our activities.

Name	
Signature	
Date	

Annex G Organisations providing useful guidance and notable publications

- COMPETITION AND MARKETS AUTHORITY (CMA, UK) https://www.gov.uk/government/organisations/competition-and-markets-authority
- EUROPEAN COMMISSION
    - http://ec.europa.eu/justice/data-protection/
    - http://ec.europa.eu/competition/antitrust/overview_en.html
    - https://ec.europa.eu/digital-single-market/en/
    - https://ec.europa.eu/programmes/horizon2020/
- EUROPEAN DATA PROTECTION SUPERVISOR (EDPS) https://edps.europa.eu/
- EU ARTICLE 29 WORKING PARTY (ART.29 WP) http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=50083
    - See, in particular, Opinion on Anonymisation Techniques (2014)
- EU INTELLECTUAL PROPERTY OFFICE (EUIPO) https://euipo.europa.eu/ohimportal/en
- INSTITUTE FOR THE LAW AND THE WEB, UNIVERSITY OF SOUTHAMPTON (ILAWS, UK) http://www.southampton.ac.uk/ilaws/index.page
    - See, in particular, 'Anonymous data v. Personal data–A false debate: An EU perspective on anonymisation, pseudonymisation and personal data'. *Wis. Int'l LJ* (2016).
- INFORMATION COMMISSIONER'S OFFICE (ICO, UK) https://ico.org.uk/
    - See, in particular, 'Anonymisation Code of Practice' (2012) and 'Data Sharing Code of Practice' (2011)
- INTELLECTUAL PROPERTY OFFICE (IPO, UK) https://www.gov.uk/government/organisations/intellectual-property-office
- OPEN DATA INSTITUTE (ODI, UK) https://theodi.org
    - See, in particular, 'Handling personal data – a checklist for organisations' (2017)
- ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD) http://www.oecd.org/sti/ieconomy/data-driven-innovation.htm
- UKANON (UK) http://ukanon.net/externalresources/
    - See, in particular, 'Anonymisation decision-making framework' (2016)

In version 2.0, it is intended to extend this list with reference to useful guidance from data protection authorities in countries beyond the UK. In the meantime, it may be useful to refer to additional guidance such as the following: https://united-kingdom.taylorwessing.com/en/global-data-protection-guide.