Data Pitch

H2020-ICT-2016-1

Project number: 732506

D3.5 Legal and Privacy Toolkit v2

Coordinators: Dr Sophie Stalla-Bourdillon and Dr Laura Carmichael

With contributions from: Dr Pei Zhang (Developer)

Quality reviewer: Open Data Institute (ODI)

Deliverable nature:	Report (R)
Dissemination level: (Confidentiality)	Public (PU)
Nature	Report
Work package	3
Contractual delivery date:	30 June 2018
Actual delivery date:	30 June 2018
Version:	1.1
Keywords:	Law, regulatory compliance, data protection, data flow mapping, personal data, anonymisation assessment, pseudonymisation, re-identification risk, data situation models, GDPR, e-learning tool

Table of Contents

List of figures

Abbreviations

AEPD = Agencia Española de Protección de Datos
CMS = Content Management System
CNIL = Commission Nationale de l'Informatique et des Libertés
DPIA = Data Protection Impact Assessment
EC = European Commission
ENISA = European Union Agency for Network and Information Security
GDPR = General Data Protection Regulation
IAPP = International Association of Privacy Professionals
ICO = Information Commissioner’s Office
ICT = Information and Communications Technologies
ODI = Open Data Institute
PIA = Privacy Impact Assessment
SME = Small or Medium Sized Enterprise
UK = United Kingdom
UKAN = UK Anonymisation Network
v1 = version 1
v2 = version 2

Abstract

The Legal and Privacy Toolkit v2 is conceived as a supplement to the Legal and Privacy Toolkit v1 (available at: https://datapitch.eu/privacytoolkit/), which extends the data protection guidance provided in the first version of the toolkit. The objective is to provide data owners with guidance on the creation of data situation models to be used as part of anonymisation assessment. In particular, it aims to raise-awareness of data flow mapping as an effective and pragmatic approach to the creation of data situation models – and compliance with the General Data Protection Regulation (GDPR). It offers practical guidance that can be understood by non-data protection specialists through devised training materials: (i) three legal decision-trees – an interactive version of these is provided through a prototype e-learning tool; and (ii) a workshop. Part A focuses on some of the key data protection considerations that underpin the robust assessment of anonymisation practices. Part B focuses on the basic components of mapping data flows.

Disclaimer: The content of the Legal and Privacy Toolkit does not constitute legal advice. If in doubt, you should always contact a lawyer.

Executive summary

D3.5 toolkit update: The Legal and Privacy Toolkit ("the toolkit") is a crucial component of the Data Pitch programme. In June 2017, the first version of the toolkit was published on the Data Pitch website: https://datapitch.eu/privacytoolkit/. The toolkit covers a wide-range of key legal considerations that are likely to occur in the course of an open innovation programme – from contractual obligations to intellectual property rights. The Legal and Privacy Toolkit v2 (D3.5) is a toolkit update that extends the data protection guidance provided in the first version of the toolkit.

Objective: In accordance with the Grant Agreement, the central focus of Data Pitch Deliverable 3.5 is as follows:

"A data situation model to assess anonymisation practices of data owners that can be used as a guide by data owners themselves before releasing their data."

Definitions: For the purpose of this deliverable, the terms used by the Grant Agreement objective are defined as follows:

A data owner: An individual and/or organisation who is involved with, and therefore able to make decisions about: the collection, management, release and/or (re)usage of a dataset. E.g. a data provider who releases data within an open innovation programme or an SME who re-uses these data.

Note: this term is widely used in ICT and business sectors – it is not a legal term.

A data situation model: A representation of a (particular version of a) dataset and the relationships with its various environments – past, current and planned – that can be used as a basis for anonymisation assessment.

Note: "data situation" is a term utilised by the UK Anonymisation Network (UKAN) see: [1, p. 130].

An anonymisation practice: A technical and/or organisational method employed to de-identify a dataset (i.e. remove personally identifiable information from a dataset) [2, p. 11]. There are two main approaches to anonymisation [2, pp. 11-19]: (i) *"randomization techniques"* – e.g. *"noise addition"*, *"permutation"* and *"differential privacy"* and (ii) "*generalization techniques*" – e.g. *"aggregation and K-anonymity"* and *"L-diversity and T-Closeness"*.

Note: this term is not to be confused with the higher benchmark set by the legal standard for anonymisation (see definition below).

Assessment of anonymisation practices: A process of evaluation undertaken by data owners to ensure their past, current and planned data practices and activities are legally and ethically compliant.

Legal standard for anonymisation: *"[…] personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable […]"* Source: Recital 26 of the General Data Protection Regulation [3].

The Legal and Privacy Toolkit v2 focuses on data mapping: It aims to equip data owners with guidance on the basics of mapping data flows. This is because data flow mapping is an effective and practical approach to the creation of data situation models. For the purpose of this deliverable, data flow mapping is defined as follows:

Data flow mapping: The creation of a graphical representation that charts the actual and potential movement of a (particular version of) a dataset as it is collected, managed, shared and (re)used across various data environments. An approach employed for the creation of data situation models.

The Legal and Privacy Toolkit v2 is therefore conceived as a supplement to the Legal and Privacy Toolkit v1, which aims to raise-awareness of the important role that data flow mapping can play in responsible data sharing and re-usage within open innovation programmes. Furthermore, it offers practical guidance that can be understood by non-data protection specialists through devised training materials. In particular, data flow mapping is presented as a way in which data owners can demonstrate compliance with the General Data Protection Regulation (GDPR). The Legal and Privacy Toolkit v2 is divided into three parts:

Part A – Understanding the data spectrum: A data flow map is useless for data protection compliance without prior understanding of how personal data is defined by the GDPR. Therefore, effective data flow mapping depends on a solid understanding of (at minimum) the following key data protection considerations:

1. What types of data are likely to fall within and outside the scope of the GDPR.
2. What types of data processing activities are considered as high-risk under the GDPR.
3. What types and levels of measures are required to control the flow of data.

Guidance is therefore given on these three areas of key data protection considerations by: (i) an overview of existing, authoritative guidance; and, (ii) the creation of three legal decision-trees that aim to help raise-awareness.

Part B – The basics of mapping data flows: The Legal and Privacy Toolkit v2 introduces the following key elements of mapping data flows:

- Data flow mapping can be used at enterprise-level and/or dataset-level.
- Data flow mapping has numerous benefits, including gap-identification and risk mitigation.
- The content of a data flow map is most crucial – not the format it takes.
- An effective data flow map will take into consideration both the technical and organisational aspects pertaining to a particular data situation.
- Access to robust provenance information is an advantage for mapping data flow activities.

It further provides three fictional scenarios that data owners can utilise to create data situation models through mapping data flows.

Part C – The development of training materials: Work carried out in Parts A and B – in particular the three legal decision-trees and three fictional scenarios for data flow mapping – led to: (i) a legal training workshop held in May 2018; and (ii) the development of a prototype e-learning tool on data protection and the basics of mapping data flows. The current version of this e-learning tool comprises a series of three interactive legal decision trees (produced in Part A). The workshop hand-out and screen-shots from the interactive e-learning tool are available as annexes to this supplementary report.

Conclusions: The Legal and Privacy Toolkit v2 concludes with two key points for data owners to take forward when (i) assessing their anonymisation practices and (ii) building data situation models to guide these practices before data are released and/or re-used. An effective assessment of anonymisation practices requires understanding of the following:

1. **Context and purpose.** In order to determine whether a particular planned data processing activity is personal or non-personal, this decision will heavily rely on the context and purpose of the specific activity under consideration. Therefore, just because a dataset was used for non-personal purposes in the past does not mean that it cannot be utilised for personal purposes in the future.
2. **An understanding of the basics of mapping data flows.** While data flow mapping is not a panacea for GDPR-compliance, it is a useful tool to employ so that: data owners demonstrate compliance with the GDPR; gaps between the regulatory framework and how data are processed in practice are revealed; (potential) high-risk data processing activities are identified and risks can be mitigated; individuals are well-positioned to make good decisions about data processing; and, the areas that require further legal training and guidance are exposed.

Further toolkit updates: Given the nature of this toolkit 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 https://datapitch.eu/privacytoolkit/. This interactive format reflects the anticipated emerging demands for guidance in specific areas in association with the key features of this toolkit.

1. Introduction

1.1 D3.5: Legal and Privacy Toolkit Update

The Legal and Privacy Toolkit ("the toolkit")[1] [4] is a crucial component of the Data Pitch programme.[2] In June 2017, the first version of the toolkit was published on the Data Pitch website: https://datapitch.eu/privacytoolkit/. The toolkit covers a wide-range of key legal considerations that are likely to occur in the course of an open innovation programme[3] – from contractual obligations to intellectual property rights. The Legal and Privacy Toolkit v2 is a toolkit update that extends the data protection guidance provided in the first version of the toolkit (at the half-way point – M18 – of the programme).

1.2 Objective

In accordance with the Grant Agreement, the central focus of Data Pitch Deliverable 3.5 is as follows:

> *"A data situation model to assess anonymisation practices of data owners that can be used as a guide by data owners themselves before releasing their data."* [Underlining added for emphasis.]

1.2.1 Objective definitions

For the purposes of the deliverable, the terms used by the Grant Agreement objective are defined as follows:

A data owner: An individual and/or organisation who is involved with, and therefore able to make decisions about: the collection, management, release and/or (re)usage of a dataset. E.g. a data provider who releases data within an open innovation programme or an SME who re-uses these data.

Note: this term is widely used in ICT and business sectors – it is not a legal term.[4]

A data situation model: A representation of a (particular version of a) dataset and the relationships with its various environments – past, current and planned – that can be used as a basis for anonymisation assessment.

Note: "data situation" is a term utilised by the UK Anonymisation Network (UKAN) see: Mark Elliot et al. The Anonymisation Decision-Making Framework (2016), p. 130 <http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf> [last accessed 24 May 2018].

A pseudonymisation practice: A technical and/or organisational method employed to de-identify a dataset by replacing an attribute with another attribute [2, p. 20]. Examples of pseudonymous practices include [2, pp. 20-21]: (i) *"encryption with secret key"*; (ii) *"hash*

[1] The principal focus of this toolkit [4] is to ensure that all those involved with Data Pitch: (a) are made aware of the key legal rights that arise in relation to data sharing as part of the programme; and, therefore (b) adhere to any legal obligations that concern these data sharing activities. This legal guidance is achieved through the provision of an overview concerning the legal and regulatory framework that applies to data sharing and data reuse. This overview: (i) sets out the key considerations that govern the data sharing arrangements between the parties involved in the programme; (ii) maps the relevant legal issues that arise in the context of data sharing; and (iii) outlines a methodology for handling these legal issues 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.

[2] The principal motivation for the Data Pitch programme is to support successful applicants (i.e. start-ups) with their high-impact, innovative and data-centric business ideas, products and services that directly respond to the specific challenges defined by the programme [47]. This support is given in numerous forms, mentoring and training services to financial assistance and rights to re-use valuable data that would otherwise remain inaccessible i.e. closed data. For more information about closed, shared and open datasets – the data spectrum – see [46].

[3] Open innovation acceleration programmes strive for the development of high impact, cutting-edge products and services. In order to bring these innovative ideas to fruition, participants are often required to share and re-use data. For further information on innovation accelerators see [45] – and the European Commission (EC) online resources on open innovation [70].

[4] The term "data ownership" is commonly used by those in ICT and business sectors to refer to: *"the de facto holder of data, and [who] can therefore decide on the use and trade of these data"* [54, p. 760]. Note that the terms "data holder" or "data steward" may also be used instead of data owner. Data Pitch does not therefore utilise the phrase "data owners" in *"the sense of legal property"* [54, p. 760], but to cover the range of roles involved with the sharing and (re)usage of data in the course of the programme. For further background information on the role of data ownership within the business sector – see the following articles by the Data Governance Institute: [55] and [56]. For legal analysis on whether "big data" requires data ownership rules as a new data property right – see [57].

function”; (iii) *“keyed-hash function with stored key”*; (iv) *“deterministic encryption of keyed-hash function with deletion of the key”*; and (v) *“tokenization”*.[5]

Note: pseudonymisation practices are distinct from anonymisation practices (see definition below).

An anonymisation practice: A technical and/or organisational method employed to de-identify a dataset (i.e. remove personally identifiable information from a dataset) [2, p. 11]. There are two main approaches to anonymisation [2, pp. 11-19]: (i) *“randomization techniques”* – e.g. *“noise addition”*, *“permutation”* and *“differential privacy”* and (ii) “*generalization techniques*” – e.g. *“aggregation and K-anonymity”* and *“L-diversity and T-Closeness”*.[6]

Note: this term is <u>not</u> to be confused with the higher benchmark set by the legal standard for anonymisation (see definition below).[7]

Assessment of anonymisation practices: A process of evaluation undertaken by data owners to ensure their past, current and planned data practices and activities are legally and ethically compliant.

Legal standard for anonymisation: *“[…] personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable […]”* Source: Recital 26 of the General Data Protection Regulation [3].

1.3 Overview: data situation models and data flow mapping

It is anticipated that a robust data situation model will better-position data owners to make effective decisions about their planned data processing activities – a notion that is at the heart of the UKAN Anonymisation Decision-Making Framework. The Anonymisation Decision-making Framework (ADF) is a “practical guide to anonymisation” [1, p. xii] that is principally intended for use by those involved with the sharing and (re)usage of personal data and anonymised data.[8] Through the process of creating a data situation model, data owners should have an improved understanding of (i) the overall context and purpose of the planned data processing activity under consideration and therefore (ii) be better-placed to establish a specially-devised plan for appropriate anonymisation that ensures any planned data processing activity is both legally and ethically compliant [1, pp. 68-69].

The Legal and Privacy Toolkit v2 focuses **on data mapping as an effective and practical approach to the creation of data situation models**, as Mark Elliot et al. [1, p. 69] state in the UKAN Anonymisation Decision-Making Framework: *“by mapping the data flow from the point at which data is collected to the point after which it is shared or released you will be able to define the parameters of your data situation.”*

1.3.1 Define: data flow mapping

For the purpose of this deliverable, data flow mapping is defined as follows:

Data flow mapping: The creation of a graphical representation that charts the actual and potential movement of a (particular version of) a dataset as it is collected, managed, shared and (re)used across various data environments. An approach employed for the creation of data situation models.

[5] See [2, pp. 20-21] for more information on these examples of these pseudonymisation practices.

[6] See the Article 29 Working Party’s Opinion 05/2014 on Anonymisation Techniques [2, pp. 11-19] for more information on these examples of anonymisation practices. For analysis of how this Opinion on Anonymisation Techniques relates to the GDPR see: [75].

[7] In common parlance, the term “anonymise” means: *“[r]emove identifying particulars or details from (something, especially medical test results) for statistical or other purposes”* [58]. However, as James Clark [19, p. 10] states: “*For many organisations which aren't deeply versed in these issues, there is often a gap between on the one hand, the popular notion of what constitutes anonymisation, and its legal meaning in the context of data protection on the other. Inevitably, the latter represents a higher bar.”*

[8] Note that the original UK version of the Anonymisation Decision-Making Framework (ADF) has been adapted for the Australian context*:* *“The De-identification Decision-Making Framework”* [60], [61].

Some key benefits of data flow mapping:

Demonstrate compliance [5, p. 6] – Data flow mapping can help data owners to adhere to the obligation under Recital 82 of the GDPR that requires a controller or processor to *"maintain records of processing activities under its responsibility"*.[9]

Reveal gaps [6, p. 8], [7, p. 7] – Data flow mapping can help to highlight any gaps between the regulatory framework with how data are collected, managed, shared and (re)used in practice.

Risk mitigation [6, p. 8] – It can draw attention to (potential) high-risk data processing activities before data are shared and re-used within an open innovation environment. It can further help to identify the appropriate technical and organisational measures required to assist with the desired level of control over a dataset.

Robust decision-making. It can provide a knowledge-base for robust decision-making about if and how best to share and re-use data.

Legal training [6, p. 8] – By understanding where the gaps between practice and the regulatory framework lie, open acceleration programmes can better-target the areas that require further legal training.

1.3.2 Enhanced guidance

The Anonymisation Decision-making Framework covers a broad-range of *"core anonymisation activities"* that all relate to the assessment of anonymisation practices: (1) *"a data situation audit"*; (2) *"risk analysis and control"*; and, (3) *"impact management"* [1, p. 67]. The objective of this report falls under the first type of core anonymisation activities – (1) *"a data situation audit"*. Furthermore, a data situation audit is comprised of five components: (a) *"describe your data situation"*; (b) *"understand your legal responsibilities"*; (c) *"know your data"*; (d) *"understand the use case"*; and, (e) *"meet your ethical obligations"* [1, p. 67].

In the opinion of IT Governance [8], the three most challenging issues that arise for those who are mapping data flows all relate to legal understanding: (i) *"identifying personal data"*; (ii) *"identifying appropriate technical and organisational safeguards"*; and (iii) *"understanding legal and regulatory obligations".* The Legal and Privacy Toolkit v2 therefore has a much narrower focus than the UKAN Anonymisation Decision-Making Framework, and as such aims to provide further detail on the creation of data situation models for anonymisation assessment by: (i) enhanced focused on key data protection considerations that arise under the GDPR – and are most challenging for data flow mapping; and, (ii) enriched guidance on the basics of mapping data flows.

1.4 Report overview

The Legal and Privacy Toolkit v2 is therefore conceived as a supplement to the Legal and Privacy Toolkit v1, which aims to raise-awareness of the important role that data flow mapping can play in responsible data sharing and re-usage within open innovation programmes. The Legal and Privacy Toolkit v2 aims to equip data owners with guidance on the basics of this approach.

Furthermore, it offers practical guidance that can be understood by non-data protection specialists through devised training materials. In particular, data flow mapping is presented as a way in which data owners can demonstrate compliance with the General Data Protection Regulation (GDPR) 2016/679 [3].[10]

The Legal and Privacy Toolkit v2 is divided into the following parts:

- **Part A – Understanding the data spectrum.** A data owner will only be able to assess their anonymisation practices effectively when they have sufficient knowledge of their legal obligations

[9] Furthermore, note that under certain circumstances, Article 30 of the GDPR identifies the types of information that are required in a record of processing activities for both controllers (see Article 31(1)) and processors (see Article 32(2)). Article 30(5) of the GDPR states: *"The obligations referred to in paragraphs 1 and 2 shall not apply to an enterprise or an organisation employing fewer than 250 persons unless the processing it carries out is likely to result in a risk to the rights and freedoms of data subjects, the processing is not occasional, or the processing includes special categories of data as referred to in Article 9(1) or personal data relating to criminal convictions and offences referred to in Article 10."*

[10] This guidance is very timely as the GDPR entered into force on 25 May 2018 – during the deliverable D3.5.

under the GDPR. Therefore, Part A therefore focuses on some of the key data protection considerations that underpin useful assessment of anonymisation practices:

1. What types of data are likely to fall within and outside the scope of the GDPR.
2. What types of data processing activities are considered as high-risk under the GDPR.
3. What types and levels of measures are required to control the flow of data.

Guidance is therefore given on these three areas of key data protection considerations by: (i) an overview of existing, authoritative guidance; and, (ii) the creation of three legal decision-trees that aim to help raise-awareness by communicating these three key aspects of the GDPR in a simple way to data owners.

- **Part B – The basics of mapping data flows.** Part B outlines the basics of mapping data flows – as a useful approach to the creation of data situation models for GDPR-compliance – which data owners can take forward as part of their anonymisation assessment practices before data are released and/or re-used.

- **Part C – The development of training materials.** Part C explains how Parts A and B led to the development of further legal training materials: (i) a prototype e-learning tool on data protection and the basics of mapping data flows; and, (ii) a workshop – delivered to invited SMEs in May 2018.

This supplementary report then concludes by: (i) summarising the key points from Parts A-C; and (ii) outlining areas for future work.

2. Part A: Understanding the data spectrum

2.1 Brief overview

It is crucial that all those involved with data sharing and re-usage within open acceleration programmes act responsibly by remaining compliant with all applicable legal and ethical obligations (from contractual obligations to intellectual property rights). Non-compliance can have severe consequences – such as litigation and reputational damage.

In order to make sure that data flow mapping is effective, it is important to first outline some of the key data protection considerations that must be considered as part of an anonymisation assessment. Guidance is given to data owners on the following three areas in Part A:

1. What types of data are likely to fall within and outside the scope of the GDPR.
2. What types of data processing activities are considered as high-risk under the GDPR.
3. What types and levels of measures are required to control the flow of data.

This guidance is achieved through review of existing and authoritative guidance, in particular that of the Information Commissioner's Office (ICO)[11] and the Article 29 Working Party.[12] Furthermore, Part A follows the development of three paper-based legal decision-trees (produced by the report authors based on authoritative guidance) that cover these three main areas for consideration.[13] The decision-tree technique is employed as a learning device to further communicate some of the key aspects of the GDPR in a simple way to data owners by representing a key series of concepts and their outcomes.

2.1.1 Disclaimer: important notice to readers

Disclaimer: The content of the Legal and Privacy Toolkit (including any updates) does not constitute legal advice. If in doubt, you should always contact a lawyer.

2.2 Define: personal and non-personal data

2.2.1 Context and purpose

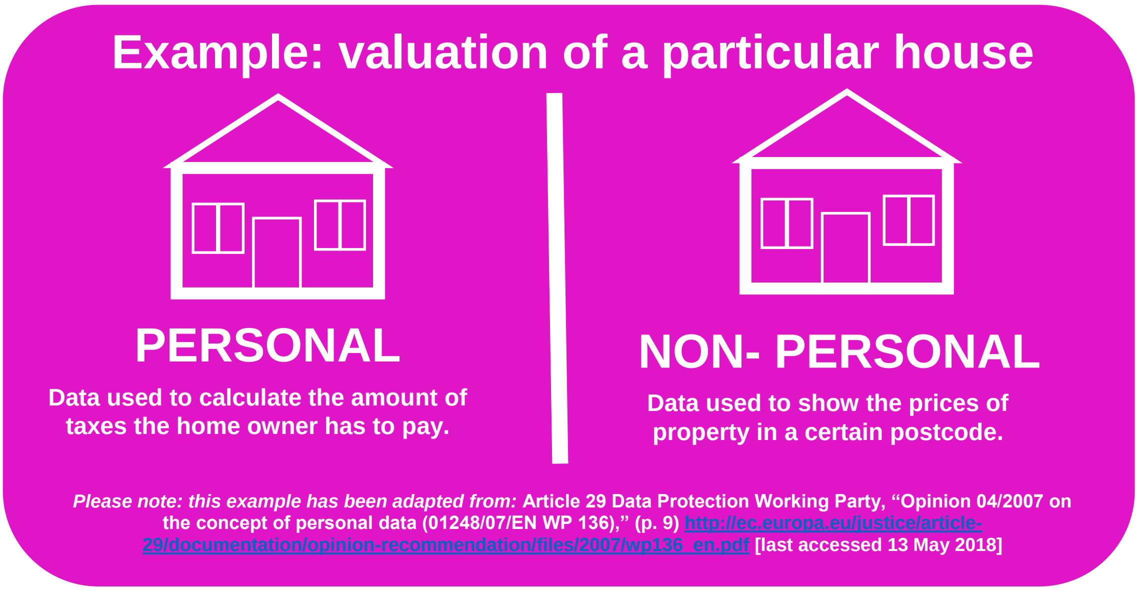
Personal data and non-personal data are not binary concepts – data exist on a spectrum.[14] For instance, one use of a particular dataset could be personal, but another use of the exact same dataset could be non-personal.

[11] ICO is *"the UK's independent authority set up to uphold information rights in the public interest, promoting openness by public bodies and data privacy for individuals"* [78].

[12] The Article 29 Working Party is an independent body that provides impartial advice to the European Commission on data protection and aims to facilitate policy harmonisation across the EU member states [63]. For further background information about the Article 29 Working Party see: Article 29 of the Data Protection Directive [64] and the Article 29 Working Party Newsroom [65].

[13] An interactive version of this series of three legal decision-trees has also been created, see Part C of this report for further information.

[14] *"Privacy has traditionally worked along a spectrum that's context dependent. Is it personal data? Well, that depends."* [49, p. 2]. For instance, Boris Lubarsky [41] represents data identifiability as a staircase. For more information about the identifiabillity data spectrum see: [9], [71] and [72]. Note: this report utilises the term "data spectrum" in the context of data protection, this usage should not be confused with other uses of the term, e.g. the ODI's data spectrum [46] focuses on the level of access to data, i.e. closed, shared and open.



Example: valuation of a particular house

PERSONAL

Data used to calculate the amount of taxes the home owner has to pay.

NON- PERSONAL

Data used to show the prices of property in a certain postcode.

Please note: this example has been adapted from: **Article 29 Data Protection Working Party, "Opinion 04/2007 on the concept of personal data (01248/07/EN WP 136)," (p. 9) http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf [last accessed 13 May 2018]**

Furthermore, there is a *"fluid line"* [9, p. 38] between properly anonymised data (for further information see section 2.2.3 of this report) and personal data as: *"anonymized data can always become personal data again depending upon the evolution of the data environment"* [9, p. 38].

Therefore, the context and purpose of each planned data processing activity (e.g. sharing data with a third party) determines whether data fall under or outside the scope of the GDPR. In the words of ICO guidance [10, p. 14]: *"It is important to remember that the same piece of data may be personal data in one party's hands while it may not be personal data in another party's hands."* Furthermore, Mark Elliot et al. [1, p. 75] state: *"A person is identifiable where the conditions exist to identify them. […] we consider whether a person is identifiable or not to be heavily contextualised."*

2.2.2 Does this dataset fall inside the scope of the GDPR?

Legal definition of personal data

In order to assess whether a planned data processing activity (e.g. a data release or re-usage) of a specific dataset falls under the GDPR, the data owner must first refer to the legal definition of personal data:

Legal definition of personal data:
"'personal data' means 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"
Source: Article 4(1) of the General Data Protection Regulation (GDPR) [3]

A fundamental aspect of this legal definition is that information relates to an identified or identifiable person i.e. a data subject.

How can data relate to individuals?

Data can relate to an individual in a number of ways. In most instances, a name combined with further information (e.g. address or telephone number) will be an adequate amount of information to identify a person [10, p. 7]. Nonetheless, it is important to note the following: (1) a name will not always be sufficient to distinguish a person from other members of a group, e.g. if two members of a group have exactly the same name [10, p. 7]; and (2) a person can be identified without reference to a name [10, p. 8].

The Information Commissioner's Office (ICO) [11, pp. 4-6] outlines the six most common ways in which data relate to an individual as follows (see original document for full information):

ICO six commons ways data relate to individuals:

1. Data are *"obviously about a particular individual"*.
2. Data are *"linked to an individual"*.
3. Data are *"used […] to inform and/or influence actions and decisions affecting an identifiable individual".*
4. Data have *"biographical significance"*.
5. Data *"focus or concentrate on the individual as its central theme".*
6. Data *"have the potential to impact on an individual".*

Source: ICO guidance [11, pp. 4-6]

It is obvious

In some cases, it is extremely obvious that a dataset relates to a data subject, because that specific dataset is about individuals [12, p. 9], [11, p. 4]. In other words, individuals are the unmistakable *"'focus' of the information"* [11, p. 5] as *"the data units are people"* [1, p. 9]. The Article 29 Working Party [12, p. 9] offers three illustrative examples of where data are clearly about individuals: *"the data registered in one's individual file in the personnel office […] the data on the results of a patient's medical test contained in his medical records, or the image of a person filmed on a video interview of that person."* The Information Commissioner's Office (ICO) [11, p. 4] provides four further examples of data that obviously relate to a data subject: *"medical history, criminal record, record of […] work or […] achievements in a sporting activity."*

In consequence, the first step towards an assessment over whether a planned data processing involves personal data is the consideration of the following question:

Legal decision-tree 1: determine whether the planned processing involves personal data
Step 1 of 4: Consider the focus of the data.
Q: Is it obvious that the data you intend to process are about individuals (i.e. the data units are people)?
A: Yes/No/I do not know

Identifiability and identified persons

The following legal decision-tree question has been adapted (by the report authors) from the following definition of anonymous data: *"information that does not identify any individuals, either in isolation or when cross referenced with other data already in the public domain"* [13].

Legal decision-tree 1: determine whether the planned processing involves personal data
Step 2 of 4: Assess whether individuals are identifiable from the data you intend to process.
(i) Identifiability in the immediate processing environment:
Q: Can you identify individual(s) from the data you intend to process in isolation (i.e. without reference to any other data and/or information)?
A: Yes/No/I do not know

The next legal decision-tree question has been adapted (by the report authors) from a question posed by ICO [10, p. 7] in its guidance on determining what is personal data: *"Can a living individual be identified from the data, or, from the data and other information in the possession of, or likely to come into the possession of, the data controller?"*

Legal decision-tree 1: determine whether the planned processing involves personal data
Step 2 of 4: Assess whether individuals are identifiable from the data you intend to process.
(i) Identifiability in the immediate processing environment:

Q: Can you identify individual(s) from the data you intend to process in isolation (i.e. without reference to any other data and/or information)?
A: Yes/No/I do not know

Again, the following legal decision-tree question has been adapted (by the report authors) from the following definition of anonymous data: *"information that does not identify any individuals, either in isolation or when cross referenced with other data already in the public domain"* [13].

Legal decision-tree 1: determine whether the planned processing involves personal data
Step 2 of 4: Assess whether individuals are identifiable from the data you intend to process.
(iii) Identifiability in the public domain:
Q: Is there more than a remote possibility that you can identify individuals(s) from these data when cross referenced with other data and/or information which is: (a) already in the public domain; and (b) likely to come into the public domain?
A: Yes/No/I do not know

It is not so obvious: content, purpose and result

In other cases, it is less obvious whether a dataset relates to a data subject, where at first glance it may appear that a particular dataset is not about individuals i.e. where the data units are objects, processes, events or other non-people entities [12, p. 9]. In other words, a dataset does not have to relate to individuals through content, but by the purpose or the result of a particular processing activity.

The Article 29 Working Party [12, p. 11] offers the following example of personal data by purpose (paraphrased from original document): a telephone call log could be used by a company to provide information about the number of callers or to learn something about the employee responsible for that phone line. The following legal decision-tree question therefore covers the purpose element of how data may relate to an individual. It has been adapted from the description of the purpose element provided by the Article 29 Data Protection Working Party Opinion (04/2007) on the concept of personal data [12, p. 10]:

> *"[…] a* ***"purpose"*** *element can be responsible for the fact that information "relates" to a certain person. That "purpose" element can be considered to exist when the data are likely to be used, taking into account all the circumstances surrounding the precise case, with the purpose to evaluate, treat in a certain way or influence the status or behaviour of an individual."*

Legal decision-tree 1: determine whether the planned processing involves personal data
Step 3 of 4: Consider whether the reasons behind your planned data processing relates to individuals.
Q: Does the reason for carrying out the planned data process relate to (at least) one of the following: (a) to learn about individuals; (b) to evaluate individuals; (c) to make a decision that affects individuals; (d) to treat individuals in a certain manner; and/or (e) to influence the status or behaviour of an individual?
A: Yes/No/I do not know

The Article 29 Working Party [12, p. 11] also provides the following example of personal data by result (paraphrased from original document): a taxi company could use location-monitoring data to make their service more efficient which would in turn impact on the taxi drivers. The following legal decision-tree question therefore covers the result element of how data may relate to an individual. It has been adapted from the description of the purpose element provided by the Article 29 Data Protection Working Party Opinion (04/2007) on the concept of personal data [12, p. 11]:

> *"[…] data can be considered to "relate" to an individual because their use is likely to have an impact on a certain person's rights and interests, taking into account all the circumstances*

surrounding the precise case. It should be noted that it is not necessary that the potential result be a major impact."

Legal decision-tree 1: determine whether the planned processing involves personal data
Step 4 of 4: Evaluate whether the consequences of your planned data processing are likely to impact on the rights and freedoms of individuals.
Q: The outcome of this planned data processing is likely to have an impact on the rights and freedoms of individuals?
A: Yes/No/I do not know

2.2.3 Does this dataset fall outside the scope of the GDPR?

There are two main types of non-personal data processing activities that fall outside the scope of the GDPR: (1) data that are properly anonymised; and (2) data that are apersonal.

Data that are properly anonymised

The legal standard for anonymisation

In order to comply with data protection law and wider ethical obligations,[15] it is of paramount importance that all those involved with the sharing and (re)usage of data are fully cognisant of the legal standard for anonymisation. Data are properly anonymised when the legal standard of anonymisation is reached.

Legal standard of anonymisation:
"[…] personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable […]"
Source: Recital 26 of the General Data Protection Regulation

According to guidance provided by ICO [14], while absolute anonymity[16] is not required, the risk of re-identification must be mitigated *"until it is remote"* [14, p. 6].[17] As part of this authoritative guidance, ICO [14, p. 48] further defines anonymised data as: *"[d]ata in a form that does not identify individuals and where identification through its combination with other data is not likely to take place".* Moreover, Steve Wood [13] defines anonymous data as: *"information that does not identify any individuals, either in isolation or when cross referenced with other data already in the public domain."*

The risk of re-identification

The risk of re-identification is defined as: (1) the probability in which an individual/individuals could

[15] For instance, Luciano Floridi and Mariarosaria Taddeo [79] consider issues relating to anonymisation, such as the sharing and (re)use of large-scale data and the potential for re-identification of individuals to be a key part of data ethics. Floridi and Taddeo [79] define "data ethics" as: *"a new branch of ethics that studies and evaluates moral problems related to data (including generation, recording, curation, processing, dissemination, sharing and use), algorithms (including artificial intelligence, artificial agents, machine learning and robots) and corresponding practices (including responsible innovation, programming, hacking and professional codes), in order to formulate and support morally good solutions (e.g. right conducts or right values)."*

[16] The position of the Article 29 Working Party is that the de-identification of personal data must be "irreversible" for data to reach the legal standard of anonymity [2, pp. 5, 7]. In other words, this view appears to support absolute anonymity, i.e. personal data are only considered as rendered anonymous where there is zero risk of data subjects being re-identified from an anonymised dataset. In practice, absolute anonymity (i.e. where there is zero risk that individuals can be identified from an anonymised dataset) is extremely difficult to achieve when you take into consideration: the utility of data, technological advancements and new data releases (e.g. open data – mosaic effect) – seen and unforeseen. Note that it appears that the Article 29 Working Party has faced some criticism in the past *"for being too conservative on data protection issues and for setting out positions that are commercially impractical"* [80].

[17] Given that absolute anonymity is impracticable, it is unsurprising that various authoritative data protection bodies disagree with the Article 29 Working Party's strict interpretation of legal standard for anonymity (see previous footnote) by favouring a more pragmatic approach. For instance, the German Data Protection Authority in Hamburg – that is *"known for its strong stance on privacy issues"* [66] – also acknowledges that absolute anonymity cannot be realised in all cases: *""German privacy law defines 'rendering anonymous' as 'the alteration of personal data so that information concerning personal or material circumstances cannot be attributed to an identified or identifiable natural person or that such attribution would require a disproportionate amount of time, expense and effort'"* [66]. Furthermore, the Handbook on European Data Protection Law [69, p. 44] considers data to be anonymised as follows: *"Data are anonymised if all identifying elements have been eliminated from a set of personal data. No element may be left in the information which could, by exercising reasonable effort, serve to re-identify the person(s) concerned."* Outside the EU, the Information and Privacy Commissioner of Ontario Canada [17, p. 3] accepts that de-identification methods are unable to completely mitigate the risks of re-identification: "*While it is not possible to guarantee that de-identification will work 100 per cent of the time, it is still an essential tool that drastically reduces the risk that personal information will be used or disclosed for unauthorized or malicious purposes."*

be identified from specific data [15, p. 26], [16, p. 24]; and, (2) the likely resultant harm or impact if re-identification were to occur [16, p. 24].[18] The Article 29 Working Party [2, pp. 11-12] identifies three principal ways in which individuals may be re-identified, by: (1) singling-out an individual; (2) linking records relating to an individual; and (3) inferring information concerning an individual.[19]

It is asserted that where personal data are *"properly de-identified"* [17, p. 4], the risk of re-identification will be *"extremely low"* [17, p. 4]. However, you cannot *"leave and forget"* – with the technological advances need to ensure anonymity is sustainable [18, p. 3].[20] Given these difficulties, it is unsurprising that the legal standard for anonymisation is described as *"very high"* [2, p. 6] Furthermore, in view of this uncertainty, it is again unsurprising that James Clark [19, p. 11] describes the *"required standard for anonymisation as a moving target"*. It will be of considerable interest to observe how the required standard for anonymisation develops over the coming years.

It is important to note that in May 2018 [20], the re-identification of de-identified personal data (without an appropriate defence) has become a criminal offence under Section 171(1) of the UK Data Protection Act [21]:*"It is an offence for a person knowingly or recklessly to re-identify information that is de-identified personal data without the consent of the controller responsible for de-identifying the personal data".*[21]

Data that are apersonal

Apersonal data are the second type of non-personal data – these data do not relate to individuals through content, purpose or result. Elliot et al. [1, p. 10] provide the following examples of apersonal data: *"Astronomical data, meteorological data, food nutrition data, bus timetables, seismological data, stress readings for the Humber Bridge and lists of endangered species".*

2.2.4 Legal decision-tree 1: determine whether the planned processing involves personal data

Instructions

Use the following decision-tree (on the next page) to determine whether a planned data processing activity is likely to involve data that relates to individuals.

You may wish to think about data you intend to process. You may also find it useful to utilise the following fictional scenarios:

- A company wants to use data collected from its loyalty card scheme to target new products at specific customers.
- Researchers want to re-use patient-monitoring data for a specific research study.
- An organisation wants to publish a dataset about levels of traffic in a particular city on its website.
- A dataset about levels of traffic is re-used by a taxi firm to monitor its employees.
- The locations and exact numbers of endangered species in a named location.
- The publication of census data.

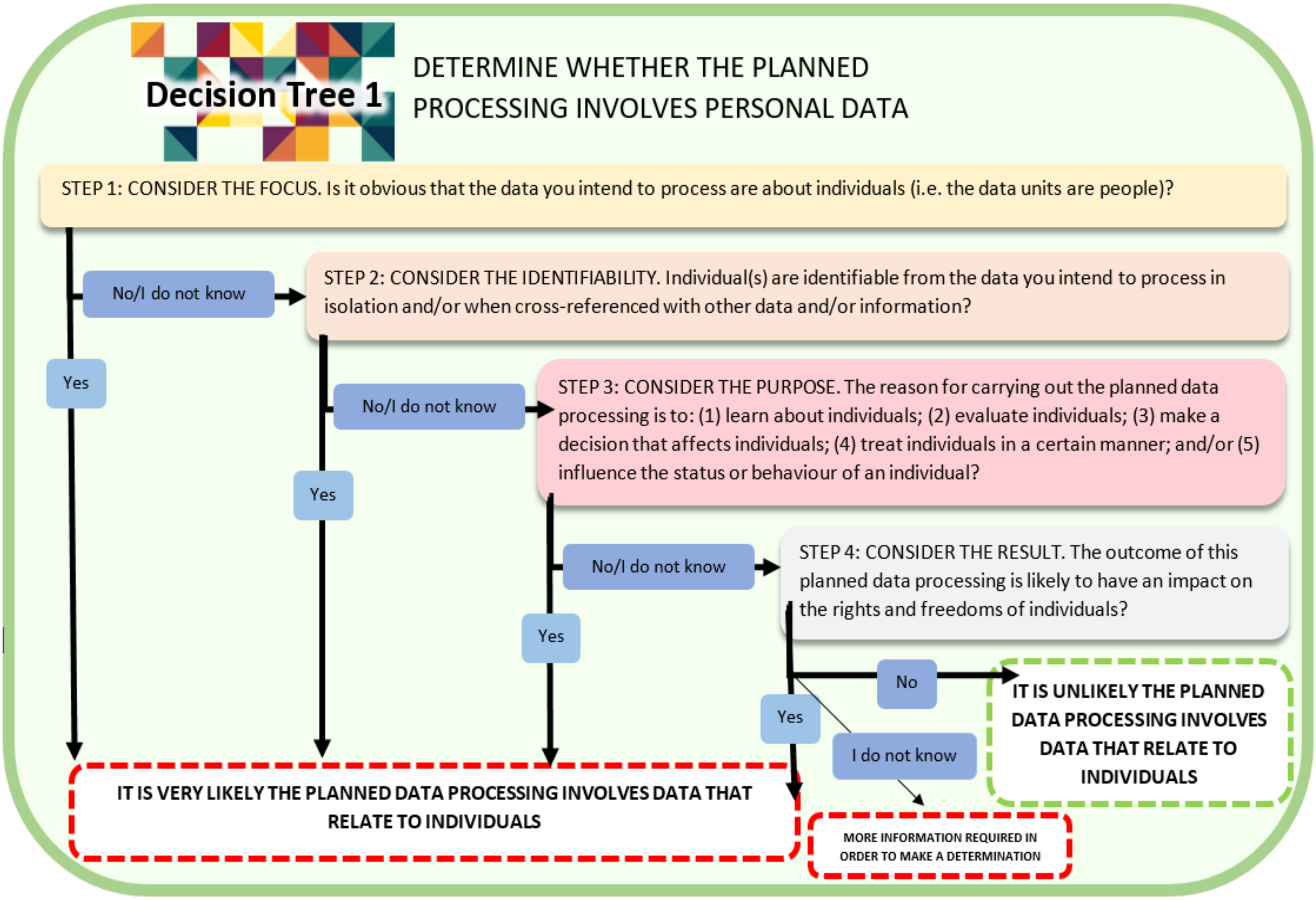
[18] In the words of Marion Oswald [16, p. 24], it is an assessment about *"the possibility of something bad happening"*.

[19] Moreover, Boris Lubarsky [41] outlines the three principal methods of re-identification: (a) *"insufficient de-identification"*; (b) *"pseudonym reversal"*; and, (c) *"combining datasets".*

[20] Kate Brimsted [18, p. 3] states: *"sever the link between the individuals and their data, and you can side step the Gordian complexity of EU data protection law. […] [/] However, […] anonymised data are starting to exhibit troubling signs of becoming 're-identified' or at least 're-identifiable'. […] it is clear that -- due to the speed of technological advances in analytics -- the data reversibility clock is ticking. Anonymisation -- the solution and the price to pay to unlock the broader utility present in massive data sets -- is getting more and more difficult to carry out with confidence."*

[21] For further background information on this offence under Section 171(1) of the Data Protection Act 2018 see: [59], [67], [68]. For an extensive overview on the history and debates surrounding the *"criminal prohibition of wrongful re-identification"* of anonymised data see the following article [62] published by Mark Philips et al.

Figure 1 Legal Decision-Tree 1: Determine whether the planned processing involves personal data



Decision Tree 1

DETERMINE WHETHER THE PLANNED PROCESSING INVOLVES PERSONAL DATA

STEP 1: CONSIDER THE FOCUS. Is it obvious that the data you intend to process are about individuals (i.e. the data units are people)?

STEP 2: CONSIDER THE IDENTIFIABILITY. Individual(s) are identifiable from the data you intend to process in isolation and/or when cross-referenced with other data and/or information?

STEP 3: CONSIDER THE PURPOSE. The reason for carrying out the planned data processing is to: (1) learn about individuals; (2) evaluate individuals; (3) make a decision that affects individuals; (4) treat individuals in a certain manner; and/or (5) influence the status or behaviour of an individual?

STEP 4: CONSIDER THE RESULT. The outcome of this planned data processing is likely to have an impact on the rights and freedoms of individuals?

IT IS VERY LIKELY THE PLANNED DATA PROCESSING INVOLVES DATA THAT RELATE TO INDIVIDUALS

IT IS UNLIKELY THE PLANNED DATA PROCESSING INVOLVES DATA THAT RELATE TO INDIVIDUALS

MORE INFORMATION REQUIRED IN ORDER TO MAKE A DETERMINATION

NOTE –this decision-tree has been derived by the toolkit authors from the following sources: [1], [10], [11], [12] and [13]. Refer to sections 2.2.2 and 2.2.3 for further information.

2.3 High-risk data processing

2.3.1 GDPR: a risk-based approach

The General Data Protection Regulation (GDPR) 2016/679 [3] marks an important change in the overall legal approach to EU data protection law – from *“an administrative process based on a priori controls to a risk-based accountability”* [22, p. xiii]. Article 35(3) of the GDPR provides three examples of where data processing is likely to result in a high risk [23, p. 8] (see section 2.3.2 below), and therefore requires a mandatory data protection impact assessment (DPIA) to assess *“the impact of the envisaged processing operations on the protection of personal data”* (Article 35(1) of the GDPR):

Security of processing:
“Taking into account the state of the art, 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, the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk […]”
Source: Article 32(1) of the General Data Protection Regulation (GDPR) [3]

Once again, the likelihood and severity of such risks to the rights and freedoms of natural persons are dependent on the specific context and purpose of the planned data processing activity under scrutiny.[22] An advantage of using data flow mapping is that it can draw attention to (potential) high-risk data processing activities before data are shared and re-used within an open innovation environment. It can further help to identify the appropriate technical and organisational measures required to assist with the desired level of control over a dataset.

2.3.2 GDPR: types of high-risk processing

Therefore, a crucial part of anonymisation assessment is for data owners to determine whether the planned data processing is likely to constitute a high risk to the rights and freedoms of data subjects [23], [24, p. 13]. Despite no *“definitive”* [25] list for high-risk data processing activities, three examples of high-risk data processing activities (see below) are outlined by Article 35(3) of the GDPR. Furthermore, under Article 35(4) of the GDPR the supervisory authority can release a public list of high-risk processing examples that require a mandatory DPIA.[23] For instance, the Article 29 Data Protection Working Party [23] and ICO [25] have both released such lists of examples.

Article 35 of the GDPR

Article 35(3) provides examples of where data processing is likely to result in a high risk [23, p. 8]:

Examples of high-risk data processing activities:
“A data protection impact assessment referred to in paragraph 1 shall in particular be required in the case of: [/] (a) 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; [/] (b) processing on a large scale of special categories of data referred to in Article 9(1), or of personal data relating to criminal convictions and offences referred to in Article 10; or [/] (c) a systematic monitoring of a publicly accessible area on a large scale.”
Source: Article 35(3) of the General Data Protection Regulation (GDPR) [3]

[22] Richard Thomas [48, p. 4] provides three types of harm that may result from the processing of personal data: “*This is potentially controversial territory, but I suggest that three main types of harm can be identified and should be set out in suitable terms in the [GDPR] legislation: [/] Material/tangible harm to individuals: For example, damage to health, financial interests, liberty or freedom of movement. [/] Moral/non-tangible harm to individuals: For example, damage to reputation or to expectations of privacy and family life. [/] Societal harm: For example, threats to the democratic values of a free society or the prospect of excessive State power.”*

[23] Article 35(4) of the GDPR [3] states: *“The supervisory authority shall establish and make public a list of the kind of processing operations which are subject to the requirement for a data protection impact assessment pursuant to paragraph 1. The supervisory authority shall communicate those lists to the Board referred to in Article 68.”*

In consequence, the first step towards an assessment over whether a planned data processing is high risk requires consideration of the following question:

Legal Decision-Tree 2: Determine whether the planned processing is likely to be a high risk to individuals?
Step 1 of 3: Assess whether the planned processing is one of the three types of processing operations that are considered as high risk under Article 35(5) of the GDPR.
Q: Does the planned processing involve (at least one of the following):
(i) Systematic and extensive profiling of individuals (e.g. profiling and prediction) with significant effects?
(ii) Large scale use of sensitive data.
(iii) Public monitoring.
A: Yes/No/I do not know

Information Commissioner's Office (ICO) guidance

National data protection authorities also have an important role to issue further guidance to data owners on the types of high-risk processing (aside from the three examples explicitly featured in the GDPR) that would require a mandatory DPIA. For instance, in the UK, ICO has released a list of ten high-processing examples (the following is paraphrased – see original webpage [25] for full information – including the entire list of examples): [24]

ICO ten examples of processing likely to result in high risk:
(i) *"New technologies"* – e.g. *"Artificial intelligence, machine learning and deep learning".*
(ii) *"Denial of service"* – e.g. *"Credit checks".*
(iii) *"Large-scale profiling"* – e.g. *"Social media networks"*
(iv) *"Biometrics"* – e.g. *"Facial recognition systems".*
(v) *"Genetic data"* – e.g. "*Medical diagnosis".*
(vi) *"Data matching"* – e.g. *"Fraud prevention".*
(vii) *"Invisible processing"* – e.g. *"Online advertising".*
(viii) *"Tracking"* – e.g. *"Data processing in the context of home and remote working".*
(ix) *"Targeting of children or other vulnerable adults"* – e.g. *"Connected toys".*
(x) *"Risk of physical harm"* – e.g. *"Whistle-blowing/compliant procedures".*

Source: ICO guidance online [25]

Therefore, the second step towards an assessment over whether a planned data processing is high risk requires consideration of further high-risk processing examples provided by national authoritative bodies, such as ICO:

Legal Decision-Tree 2: Determine whether the planned processing is likely to be a high risk to individuals?
Step 2 of 3: Assess whether the planned processing is one of the ten types of processing operations that are considered as high risk by the Information Commissioner's Office (ICO).
Q: Does the planned processing involve (at least one of the following): (i) new technologies; (ii) denial of service; (iii) large-scale profiling; (iv) biometrics; (v) genetic data; (vi) data matching; (vii) invisible processing; (viii) tracking; (ix) targeting of children or other vulnerable adults; and/or (x) risk of physical harm.
A: Yes/No/I do not know

[24] It is important to note that other national data protection authorities within the EU also issue guidance on high-risk processing under the GDPR. For instance, the Agencia Española de Protección de Datos (AEPD) [44] outlines the following four high-risk scenarios [24, p. 13]: (i) decision-making; (ii) profiling; (iii) predictive analysis; and, (iv) health-related services, monitoring, control and observation of persons (monitoring). [This has been translated into English via Google Translate – original text [24, p. 13] in Spanish: *"Finalidades del tratamiento: Se deben identificar cada una de las finalidades del tratamiento y analizar si estas derivan en un alto riesgo. Por ejemplo, si la finalidad incluye: [/] Toma de decisiones [/] Elaboración de perfiles [/] Análisis predictivo [/] Prestación de servicios relacionados con la salud Seguimiento, control y observación de personas (monitorización)".*]

Article 29 Working Party guidance

The Article 29 Working Party [23, pp. 9-11] provides the following nine point criteria to be taken into account when considering whether a (planned) data processing activity is likely to result in a high risk to the rights and freedoms of individuals (the following is paraphrased – see original guidelines [23, pp. 9-11] for full information – including the entire list of examples):

Article 29 Working Party nine point criteria of processing likely to result in high risk:

(i) *"Evaluation or scoring"* – e.g. *"a company building […] marketing profiles based on usage or navigation of its website".*

(ii) *"Automated-decision making with legal or similar legal effect"* – e.g. *"the processing may lead to the exclusion or discrimination against individuals".*

(iii) *"Systematic monitoring"* – e.g. *"data collected through networks".*

(iv) *"Sensitive data or data of a highly personal nature" – e.g. "a general hospital keeping patients' medical records".*

(v) *"Data processed on a large-scale"* – In order to determine whether data are processed on a large-scale, the following factors must be considered: *"number of subjects", "volume" and/or "range", "duration" and/or "permanence"*, and *"geographical extent"*.

(vi) *"Matching or combining datasets".*

(vii) *"Data concerning vulnerable data subjects"* – e.g. "*children"* and *"employees"*.

(viii) *"Innovative use or applying new technological or organisational solutions"* – E.g. "*combining use of finger print and face recognition for improved physical access control".*

(ix) *"When the processing in itself "prevents data subjects from exercising a right or using a service or contract"* – e.g. *"a bank screens its customers against a credit reference database in order to decide whether to offer them a loan".*

Source: Article 29 Working Party Opinion guidelines on DPIA [23, pp. 9-11]

Following this approach, a DPIA is required when two criteria are present in a planned data processing activity. A DPIA is recommended if only one criterion is existent. Hence, the third step towards an assessment over whether a planned data processing is high risk requires consideration of further high-risk criteria provided by the Article 29 Working Party:

Legal Decision-Tree 2: Determine whether the planned processing is likely to be a high risk to individuals?

Step 3 of 3: Assess whether the planned processing is falls under the nine-point criteria of processing operations that are considered as high risk by the Article 29 Working Party.

Q: Does the planned processing involve (at least one of the following): (i) evaluation or scoring; (ii) automated-decision making with legal or similar legal effect; (iii) systematic monitoring; (iv) sensitive data or data of a highly personal nature; (v) data processed on a large-scale; (vi) matching or combining datasets; (vii) data concerning vulnerable subjects; (viii) innovative use or applying new technological or organisational solutions; and (ix) preventing data subjects from exercising a right or using a service or contract.

A: Yes/No/I do not know

2.3.3 Legal Decision-Tree 2: Is the planned data processing likely to result in a high risk to the rights and freedoms of individuals?

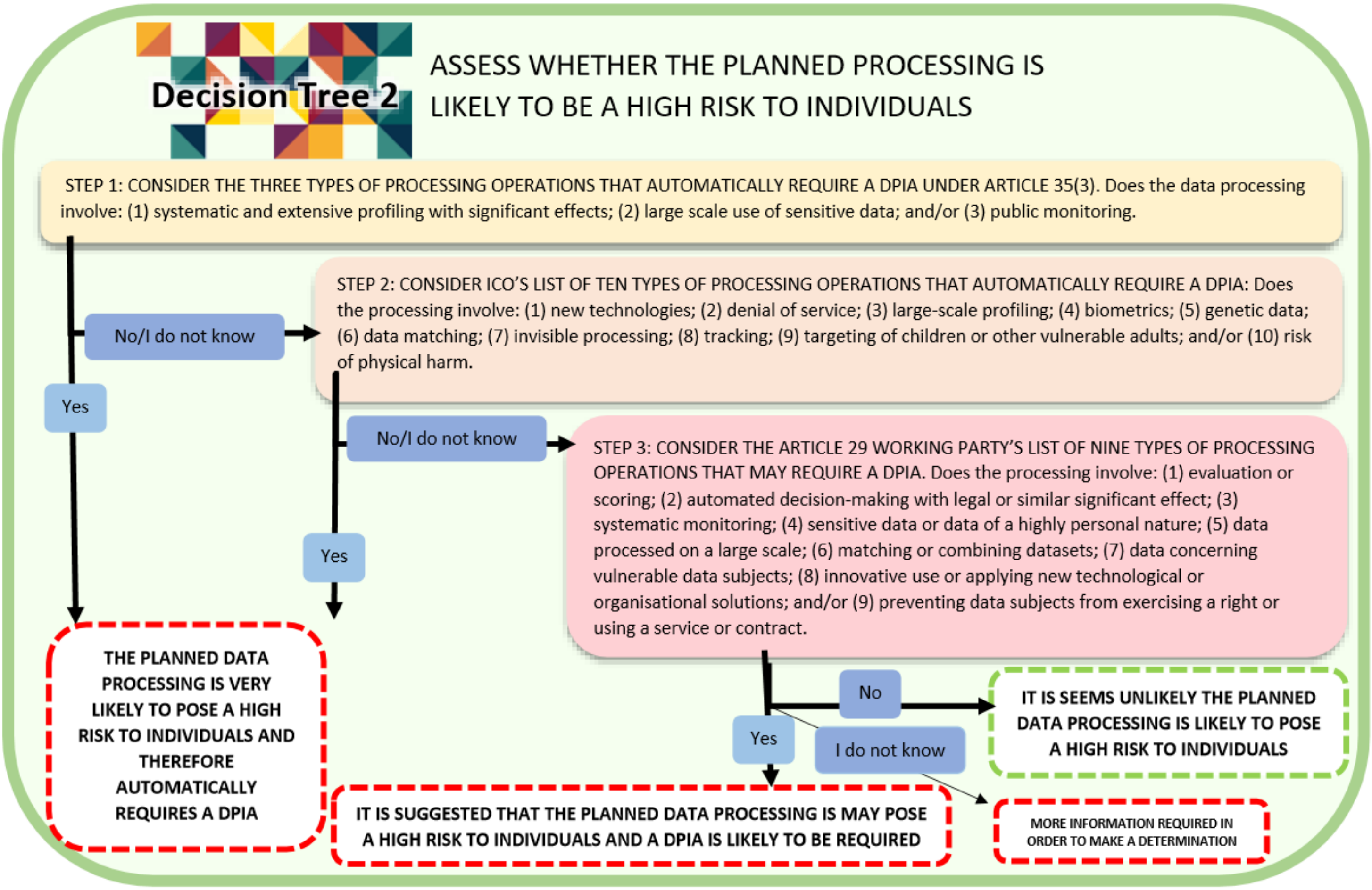
Instructions

Use the following decision-tree (on the next page) to determine whether a planned data processing activity is likely to be high risk based on examples provided by Article 35(3) of the GDPR, ICO [25], and the Article 29 Working Party [23]. It is important to note that this Legal Decision Tree is indicative of high-risk processing – it is not a definitive list of high-risk processing activities.

You may wish to think about data you intend to process. You may also find it useful to utilise the following fictional scenarios:

- A company wants to use data collected from its loyalty card scheme to target new products at specific customers.
- Researchers want to re-use patient-monitoring data for a specific research study.
- An organisation wants to publish a dataset about levels of traffic in a particular city on its website.
- A dataset about levels of traffic is re-used by a taxi firm to monitor its employees.
- The locations and exact numbers of endangered species in a named location.
- The publication of census data.

Figure 2 Legal Decision-Tree 2: Assess whether the planned processing is likely to be a high risk to individuals



ASSESS WHETHER THE PLANNED PROCESSING IS LIKELY TO BE A HIGH RISK TO INDIVIDUALS

STEP 1: CONSIDER THE THREE TYPES OF PROCESSING OPERATIONS THAT AUTOMATICALLY REQUIRE A DPIA UNDER ARTICLE 35(3). Does the data processing involve: (1) systematic and extensive profiling with significant effects; (2) large scale use of sensitive data; and/or (3) public monitoring.

STEP 2: CONSIDER ICO'S LIST OF TEN TYPES OF PROCESSING OPERATIONS THAT AUTOMATICALLY REQUIRE A DPIA: Does the processing involve: (1) new technologies; (2) denial of service; (3) large-scale profiling; (4) biometrics; (5) genetic data; (6) data matching; (7) invisible processing; (8) tracking; (9) targeting of children or other vulnerable adults; and/or (10) risk of physical harm.

STEP 3: CONSIDER THE ARTICLE 29 WORKING PARTY'S LIST OF NINE TYPES OF PROCESSING OPERATIONS THAT MAY REQUIRE A DPIA. Does the processing involve: (1) evaluation or scoring; (2) automated decision-making with legal or similar significant effect; (3) systematic monitoring; (4) sensitive data or data of a highly personal nature; (5) data processed on a large scale; (6) matching or combining datasets; (7) data concerning vulnerable data subjects; (8) innovative use or applying new technological or organisational solutions; and/or (9) preventing data subjects from exercising a right or using a service or contract.

THE PLANNED DATA PROCESSING IS VERY LIKELY TO POSE A HIGH RISK TO INDIVIDUALS AND THEREFORE AUTOMATICALLY REQUIRES A DPIA

IT IS SUGGESTED THAT THE PLANNED DATA PROCESSING IS MAY POSE A HIGH RISK TO INDIVIDUALS AND A DPIA IS LIKELY TO BE REQUIRED

IT IS SEEMS UNLIKELY THE PLANNED DATA PROCESSING IS LIKELY TO POSE A HIGH RISK TO INDIVIDUALS

MORE INFORMATION REQUIRED IN ORDER TO MAKE A DETERMINATION

NOTE – If you answer "I don't know" to any of the questions – overall outcome = more information is required.

This decision-tree has been derived by the Toolkit authors from the following sources: Article 35(3) of the GDPR, [23] and [25]. Refer to sections 2.3.1 and 2.3.2 for further information.

2.4 What levels of control are required?

2.4.1 Data protection by default and design

Pursuant to Article 25(1) of the GDPR [3], controllers need to ensure that any planned data processing preserves the rights and freedoms of data subjects by design and default. This protection is achieved by: (i) implementing appropriate legal and technical measures to uphold data protection principles (see Article 5 of the GDPR); and (ii) integrating safeguards. According to Article 25(1), any decision taken on these measures and safeguards must take: *"into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing".*

2.4.2 Data type definitions

According to Roger Clark [26], the function of big data analytics can be split into the following two categories: (1) individual-focused data analytics – *"concerned about individual instances within populations"*; and, (2) population-focused data analytics – *"focus on populations and sub-populations"*. To help data (re)users to better-review the potential risks of the planned data processing, there must therefore be an assessment of whether this processing is individual-focused or population-focused. For the purpose of the Legal and Privacy Toolkit v2, the following data types are defined:

Data at individual-level: Data are recorded for each specific person [26] i.e. the data units are people [1, p. 9]. E.g. data relate to a particular customer, patient or survey participant [26], [27], [28].[25]

Data at aggregate-level: Data are recorded in a summary form (e.g. statistics) about sub-populations and populations i.e. the data units are about groups of people [26].[26] E.g. statistics about customer preferences.[27]

Apersonal data: The data units are non-people entities such as events or processes. E.g. a bus timetable.[28]

2.4.3 Dependant on context and purpose

Different levels of access and control may be applied to different versions of the particular (version of a) dataset that a data owner plans to process based on the specific circumstances of each data processing activity.

Example:
Version of dataset *y* – Raw data in closed environment. A medical professional collects sensitive personal information about a patient in order to identify and action the most effective course of treatment for medical condition *x*.
Version of dataset *y* – Pseudonymised data in restricted environment. A research team is given permission to (re)use a pseudonymous form of this data as part of a large-scale study into the treatments for medical condition *x*.
Version of dataset *y* – Aggregated data in public domain with an open licence. The hospital releases statistics about the number of patients treated for the medical condition *x* over the past year.

[25] A further example of individual-level data are census microdata, refer to [73] for more information.

[26] For further definitions of data at an aggregated level see the following references. (i) The Organisation for Economic Co-operation and Development (OECD) [42] defines the term aggregation in its glossary of statistical terms as*: "the combination of related categories, usually within a common branch of a hierarchy, to provide information at a broader level to that at which detailed observations are taken."* (ii) Margaret Rouse [43] defines data aggregation as: *"any process in which information is gathered and expressed in a summary form, for purposes such as statistical analysis. A common aggregation purpose is to get more information about particular groups based on specific variables such as age, profession, or income."*

[27] A further example of aggregate-level data are census aggregate data, refer to [74] for more information.

[28] For further examples of apersonal data see [1, p. 10].

For data processing that falls under the scope of the GDPR, data owners will have to take appropriate technical and organisational measures to ensure that they stay in control of the level of agreed access and re-usage. Furthermore, where data are properly anonymised, data owners will also have to take appropriate technical and organisational measures to ensure that they stay in control of the level of agreed access and (re)usage. The GDPR focuses on the roles of controllers and processors:

(i) **Controller.** A controller is defined by Article 4(7) of the GDPR as: *“the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data […]”*. According to Article 24(1) of the GDPR, the controller is responsible for the implementation of *“[…] appropriate technical and organisational measures […]”.*

(ii) **Processor.** A processor is defined by Article 4(8) of the GDPR as: *“a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller*”. Article 28(1) of the GDPR explains the responsibilities of the processor, including the following requirement: *“[…] sufficient guarantees [to the controller] to implement appropriate technical and organisational measures […]”.*

2.4.4 Legal Decision-Tree 3: Determine the status of control over current and planned data situations

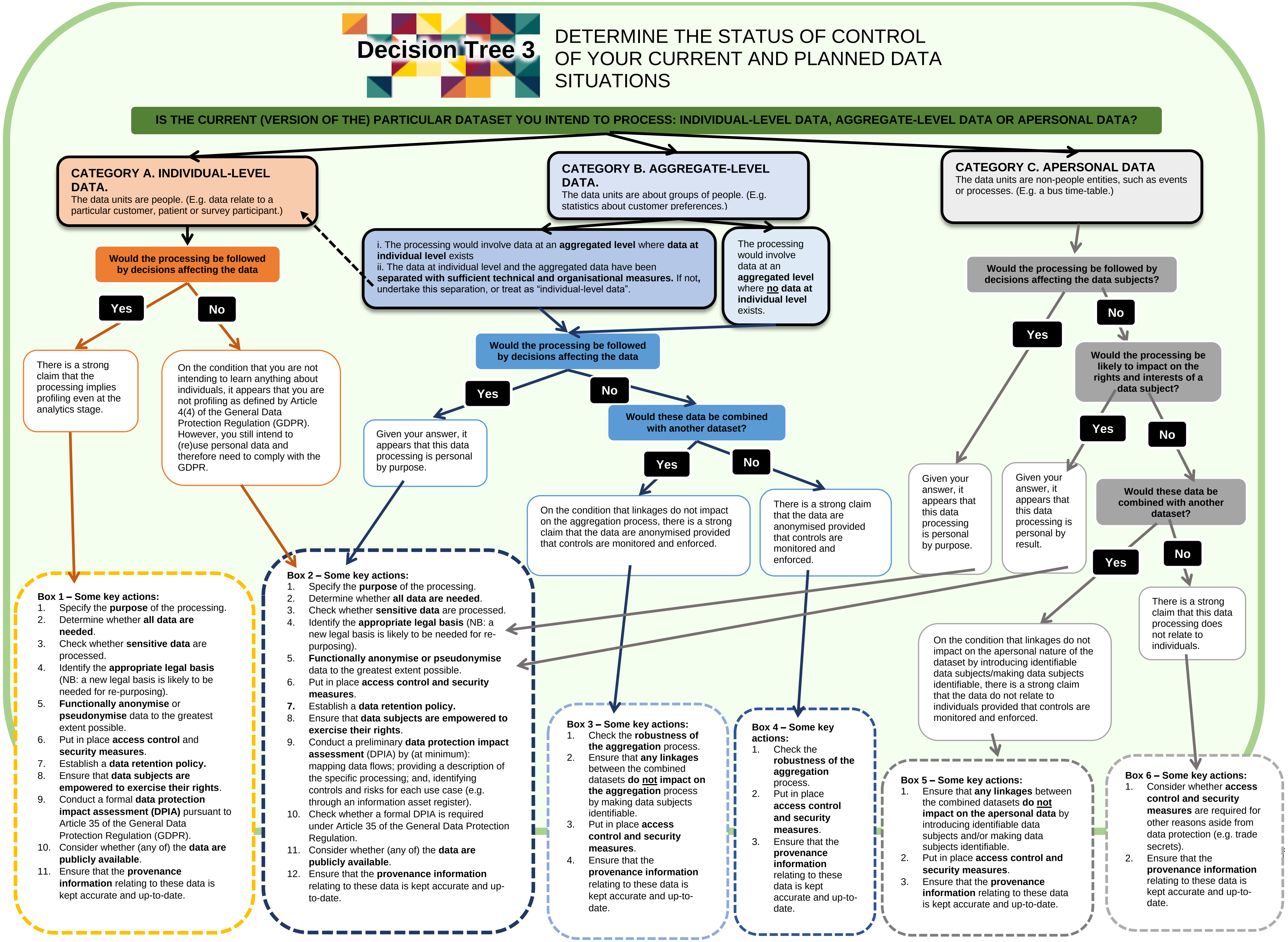
The following legal decision-tree provides an overview of the likely types of controls required for the following three categories of data: (i) data at an individual-level; (ii) data at an aggregate-level; and, (iii) apersonal data. The accompanying table (located directly after this decision-tree) offers further descriptive information about the key actions outlined by this decision-tree. Again, this legal decision-tree is indicative of the types of controls required for different types of data. It is the responsibility of data owners to take into account the individual circumstances of their planned processing activity as part of their anonymisation assessment.

Instructions

Use the following decision-tree to determine the likely level of controls required for a planned data processing activity. You may wish to think about data you intend to process. You may also find it useful to utilise the following fictional scenarios:

- A company wants to use data collected from its loyalty card scheme to target new products at specific customers.
- Researchers want to re-use patient-monitoring data for a specific research study.
- An organisation wants to publish a dataset about levels of traffic in a particular city on its website.
- A dataset about levels of traffic is re-used by a taxi firm to monitor its employees.
- The locations and exact numbers of endangered species in a named location.
- The publication of census data.

Figure 3 Legal Decision-Tree 3: Determine the status of control of your current and planned data situations



Decision Tree 3

DETERMINE THE STATUS OF CONTROL OF YOUR CURRENT AND PLANNED DATA SITUATIONS

IS THE CURRENT (VERSION OF THE) PARTICULAR DATASET YOU INTEND TO PROCESS: INDIVIDUAL-LEVEL DATA, AGGREGATE-LEVEL DATA OR APERSONAL DATA?

CATEGORY A. INDIVIDUAL-LEVEL DATA.
The data units are people. (E.g. data relate to a particular customer, patient or survey participant.)

Would the processing be followed by decisions affecting the data

Yes

There is a strong claim that the processing implies profiling even at the analytics stage.

No

On the condition that you are not intending to learn anything about individuals, it appears that you are not profiling as defined by Article 4(4) of the General Data Protection Regulation (GDPR). However, you still intend to (re)use personal data and therefore need to comply with the GDPR.

CATEGORY B. AGGREGATE-LEVEL DATA.
The data units are about groups of people. (E.g. statistics about customer preferences.)

i. The processing would involve data at an **aggregated level** where **data at individual level** exists
ii. The data at individual level and the aggregated data have been **separated with sufficient technical and organisational measures.** If not, undertake this separation, or treat as “individual-level data”.

The processing would involve data at an **aggregated level** where **no data at individual level** exists.

Would the processing be followed by decisions affecting the data

Yes

Given your answer, it appears that this data processing is personal by purpose.

No

Would these data be combined with another dataset?

Yes

On the condition that linkages do not impact on the aggregation process, there is a strong claim that the data are anonymised provided that controls are monitored and enforced.

No

There is a strong claim that the data are anonymised provided that controls are monitored and enforced.

CATEGORY C. APERSONAL DATA
The data units are non-people entities, such as events or processes. (E.g. a bus time-table.)

Would the processing be followed by decisions affecting the data subjects?

Yes

Given your answer, it appears that this data processing is personal by purpose.

No

Would the processing be likely to impact on the rights and interests of a data subject?

Yes

Given your answer, it appears that this data processing is personal by result.

No

Would these data be combined with another dataset?

Yes

On the condition that linkages do not impact on the apersonal nature of the dataset by introducing identifiable data subjects/making data subjects identifiable, there is a strong claim that the data do not relate to individuals provided that controls are monitored and enforced.

No

There is a strong claim that this data processing does not relate to individuals.

Box 1 – Some key actions:

1. Specify the **purpose** of the processing.
2. Determine whether **all data are needed**.
3. Check whether **sensitive data** are processed.
4. Identify the **appropriate legal basis** (NB: a new legal basis is likely to be needed for re-purposing).
5. **Functionally anonymise** or **pseudonymise** data to the greatest extent possible.
6. Put in place **access control** and **security measures**.
7. Establish a **data retention policy.**
8. Ensure that **data subjects are empowered to exercise their rights**.
9. Conduct a formal **data protection impact assessment (DPIA)** pursuant to Article 35 of the General Data Protection Regulation (GDPR).
10. Consider whether (any of) the **data are publicly available**.
11. Ensure that the **provenance information** relating to these data is kept accurate and up-to-date.

Box 2 – Some key actions:

1. Specify the **purpose** of the processing.
2. Determine whether **all data are needed**.
3. Check whether **sensitive data** are processed.
4. Identify the **appropriate legal basis** (NB: a new legal basis is likely to be needed for re-purposing).
5. **Functionally anonymise or pseudonymise** data to the greatest extent possible.
6. Put in place **access control and security measures**.
7. Establish a **data retention policy.**
8. Ensure that **data subjects are empowered to exercise their rights**.
9. Conduct a preliminary **data protection impact assessment** (DPIA) by (at minimum): mapping data flows; providing a description of the specific processing; and, identifying controls and risks for each use case (e.g. through an information asset register).
10. Check whether a formal DPIA is required under Article 35 of the General Data Protection Regulation.
11. Consider whether (any of) the **data are publicly available**.
12. Ensure that the **provenance information** relating to these data is kept accurate and up-to-date.

Box 3 – Some key actions:

1. Check the **robustness of the aggregation** process.
2. Ensure that **any linkages** between the combined datasets **do not impact on the aggregation** process by making data subjects identifiable.
3. Put in place **access control and security measures**.
4. Ensure that the **provenance information** relating to these data is kept accurate and up-to-date.

Box 4 – Some key actions:

1. Check the **robustness of the aggregation** process.
2. Put in place **access control and security measures**.
3. Ensure that the **provenance information** relating to these data is kept accurate and up-to-date.

Box 5 – Some key actions:

1. Ensure that **any linkages** between the combined datasets **do not impact on the apersonal data** by introducing identifiable data subjects and/or making data subjects identifiable.
2. Put in place **access control and security measures**.
3. Ensure that the **provenance information** relating to these data is kept accurate and up-to-date.

Box 6 – Some key actions:

1. Consider whether **access control and security measures** are required for other reasons aside from data protection (e.g. trade secrets).
2. Ensure that the **provenance information** relating to these data is kept accurate and up-to-date.

2.4.5 Legal decision-tree 3: key actions in brief

Legal Decision-Tree 3 aims to show data owners how the type of data they intend to process may affect the types of technical and organisational measures and safeguards, which are required to control the flow of data. The following table provides further information about the key actions outlined in Boxes 1-6:

Figure 4 Table 1: Legal Decision-Tree 3 - Further information on the key actions outlined in Boxes 1-6

Key action	Box(es)	Brief description
Specify the purpose of the processing.	1 and 2	The principle of purpose limitation is enshrined by Article 5(1)(b) of the GDPR: *“Personal data shall be […] collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes (‘purpose limitation’)”.* Data owners therefore need to ensure that they specify the purpose of the planned processing.
Determine whether all data are needed.	1 and 2	The principle of data minimisation is enshrined by Article 5(1)(c) of the GDPR: *“Personal data shall be […] adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (‘data minimisation’)”.* Data owners therefore need to ensure that all data are necessary for the planned processing activity.
Check whether sensitive data are processed.	1 and 2	It is imperative for the data owner to review whether the planned data processing activity would involve any sensitive data. This is because special categories of data – defined by Article 9 of the GDPR – may only be processed if an exemption applies (see Article 9(2)(a)-(j) for these exemptions): *““Processing of personal data revealing 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 shall be prohibited.”*
Identify the appropriate legal basis (NB: a new legal basis is likely to be needed for re-purposing).	1 and 2	It is crucial for the data owner to determine the lawful basis for the planned data processing activity before processing begins. Article 6 of the GDPR provides six lawful bases: (i) consent, (ii) contract, (iii) legal obligation, (iv) vital interests, (v) public task, and (vi) legitimate interests. The context and purpose of specific, planned data processing activity will impact the legal basis/bases that is/are selected [29]. In addition to detailed guidance on lawful basis for processing [29], ICO also provide a lawful basis interactive tool [30] on their website – to offer *“tailored guidance”* [29].
Functionally anonymise or pseudonymise data to the greatest extent possible.	1 and 2	Where possible, personal data should be anonymised [14, p. 13]. Functional anonymisation takes into consideration the data situation of the planned data processing – see [1, pp. 21-22] for further information. In some cases, it may not be possible to fully anonymise data, because it would adversely affect the utility of the specific data in question [14, p. 13], [31]. In this instance, data should be pseudonymised to the greatest extent possible.
Put in place access control and security measures.	1, 2, 3, 4 and 5	The principle of integrity and confidentiality is enshrined by Article 5(1)(f) of the GDPR: *“Personal data shall be […] processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or*

		unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures ('integrity and confidentiality')." Furthermore, Article 25(2) states: *"The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed."* Data owners therefore need to establish appropriate access and security measures for their planned data processing activities. For instance, the European Union Agency for Network and Information Security (ENISA) provides: a Handbook on Security of Personal Data Processing [32].
Consider whether access control and security measures are required for other reasons aside from data protection (e.g. trade secrets).	6	While this report focuses on control measures for GDPR-compliance, data owners remain cognisant of other legal and ethical obligations. Data owners must therefore assess what other factors may require the use of access and security measures – e.g. commercial confidentiality.
Establish a data retention policy.	1 and 2	The principle of storage limitation is enshrined by Article 5(1)(e) of the GDPR*: "Personal data shall be […] kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed […] ('storage limitation')".*[29] A data retention policy (e.g. a document that specifies how data are stored, managed and deleted) is an essential way of adhering to this principle.
Ensure that data subjects are empowered to exercise their rights.	1 and 2	Chapter III of the GDPR describes the rights of data subjects – these rights cover four main areas: (i) *"transparency and modalities"* (Article 12); (ii) *"information and access to personal data"* (Articles 13-15); (iii) *"rectification and erasure"* (Articles 16-20); and, (iv) *"right to object and automated individual decision-making"* (Articles 21-22). It is critical that data owners are cognisant of the rights of data subjects, and therefore ensure that data subjects are empowered to exercise such rights.
Conduct a formal data protection impact assessment (DPIA) pursuant to Article 35 of the General Data Protection Regulation (GDPR).	1	As aforementioned, a DPIA is mandatory for a (planned) data processing activity that is likely to result in a high-risk to the rights and freedoms of natural persons (see section 2.3 of this report for further information). A number of national data authorities provide guidance information about DPIAs, (including ICO – see [33], and Commission Nationale de l'Informatique et des Libertés (CNIL) which provides an open source PIA software to help users carry out DPIAs – see [34]).
Conduct a preliminary data protection impact assessment (DPIA)	2	While the GDPR has made DPIAs a legal requirement in certain circumstances, DPIAs (also known as privacy impact assessments) are not new and are an established part of best practice. It is therefore important for data owners to conduct a preliminary impact assessment by (at minimum): mapping data flows; providing a description of the specific processing; and, identifying controls and risks for each use case (e.g. through an information asset register).

[29] Note that under Article 5(1)(e) of the GDPR personal data that is *"processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes"* can be stored for longer periods in accordance with the conditions specified by this Article.

Check whether a formal DPIA is required under Article 35 of the General Data Protection Regulation.	2	Data owners must make sure that the planned data processing does not require a formal DPIA, i.e. that it is unlikely to result in a high-risk to the rights and freedoms of data subjects.
Consider whether (any of) the data are publicly available.	1 and 2	The principle of data minimisation is enshrined by Article 5(1)(c) of the GDPR – if these data are already available for re-use, the planned data processing activity may contravene this principle.
Ensure that the provenance information relating to these data is kept accurate and up-to-date.	All	Refer to section 3.2.4 of this report for further information.
Check the robustness of the aggregation process.	3 and 4	Data owners need to ensure that the aggregation process meets the legal standard for anonymisation now and in the future [2, p. 24]. In view of the *"residual risk of identification"*, anonymised data cannot be released and forgotten [2, p. 24]. The residual risk of identification should therefore be actively assessed, monitored and controlled [2, p. 24].
Ensure that any linkages between the combined datasets do <u>not</u> impact on the aggregation process by making data subjects identifiable.	3	Anonymised data cannot be released and forgotten [2, p. 24], data owners must actively monitor and manage any risks of re-identification. For instance, there is a possibility for re-identification to occur following a combining of datasets.
Ensure that any linkages between the combined datasets do <u>not</u> impact on the apersonal data by introducing identifiable data subjects and/or making data subjects identifiable.	5	While a dataset may be apersonal in isolation, it may be personal in combination with another dataset. E.g. a bus timetable is apersonal in isolation, but when combined with passenger data – for the purpose of monitoring the route and length of journeys taken over a specific period – it is likely to be personal data. Again, it is critical to take into account the specific context and purpose of the planned data processing.

2.5 Summary: data spectrum

After reviewing Part A, it is anticipated that the user of this guidance document should now have a better or re-affirmed understanding of (i) how personal and non-personal data are legally defined, including (ii) what types of data processing are high risk, and (iii) what appropriate control measures should be implemented. This legal understanding is essential for effective data flow mapping as part of an overall approach to GDPR-compliance.

Any decision that focuses on the appropriate level and type of measures to control a data flow (e.g. a specific data sharing activity) will heavily rely on the individual circumstances that surround the particular activity under consideration. Since open innovation programmes are highly likely to traverse a multitude of sectors – e.g. from hospitality to health care – it is not possible to provide a data situation model that covers all these possibilities. In consequence, the next part of this report (Part B) presents the basics of data flow mapping as a useful approach data owners can employ to create such data situation models for assessment of anonymisation practices.

3. Part B – the basics of mapping data flows

3.1 Brief overview

Data flow mapping is one way in which data owners can demonstrate compliance with the GDPR through adherence to Recital 82 of the GDPR [5, p. 6]:

Demonstrating compliance with the GDPR:
"In order to demonstrate compliance with this Regulation, the controller or processor should maintain records of processing activities under its responsibility. Each controller and processor should be obliged to cooperate with the supervisory authority and make those records, on request, available to it, so that it might serve for of the General Data Protection Regulation (GDPR) [3]

Data flow mapping can be used at enterprise-level [35], [5], [36, p. 4] (i.e. to chart the movement of all data through an organisation and/or network of organisations) and at dataset-level [1, p. 69] (i.e. to chart the movement of a particular dataset).[30] This guidance focuses on data flow mapping at dataset-level. Data flow mapping can also be undertaken at technical and organisational levels [7, p. 7].

3.2 The basics

3.2.1 The key benefits

In order to manage data flows effectively in accordance with the GDPR, those involved with the processing of data need to first understand these data flows. As James Graves [37] states: *"It's much easier to describe how you are protecting the confidentiality, integrity and availability of your data when you can provide details such as where it is going, how it is getting there and who is sending and receiving it."* As aforementioned, data flow mapping has a number of benefits for those involved with data processing.

Some key benefits of data flow mapping:

Demonstrate compliance [5, p. 6] – Data flow mapping can help data owners to adhere to the obligation under Recital 82 of the GDPR that requires a controller or processor to *"maintain records of processing activities under its responsibility"*.

Reveal gaps [6, p. 8], [7, p. 7] – Data flow mapping can help to highlight any gaps between the regulatory framework with how data are collected, managed, shared and (re)used in practice.

Risk mitigation [6, p. 8] – It can draw attention to (potential) high-risk data processing activities before data are shared and re-used within an open innovation environment. It can further help to identify the appropriate technical and organisational measures required to assist with the desired level of control over a dataset.

Robust decision-making. It can provide a knowledge-base for robust decision-making about if and how best to share and re-use data.

Legal training [6, p. 8] – By understanding where the gaps between practice and the regulatory framework lie, open acceleration programmes can better-target the areas that require further legal training.

3.2.2 Approach

The most important part of data flow mapping is the content not the format [6, p. 14]. While some approaches adhere to formal standards (e.g. LINDDUN Privacy Threat Modelling [38]) and are software-based (e.g. the Data Flow Mapping Tool [39]), other approaches do not insist on strict formatting rules (e.g. the ODI's Mapping Data Ecosystems Methodology [36]). In the words of Nicola Fulford and Krysia Oastler [5, p. 7]: *"There is no one right way to carry out data mapping".*

[30] For instance, the ICO [35] provide documentation templates to document processing activities within an organisation.

Data situation approach

An essential part of effective data flow mapping is an understanding of the data situation(s) that pertain to the particular (version of the) dataset you plan to process. A data situation is defined by UKAN's Anonymisation Decision-Making Framework [1, p. 130] as: *"**Data Situation:** The relationship between some data and their environment"* [Bold emphasis in original]. Each data environment should be considered as a distinct *"configuration"* of *"people, other data, infrastructure and governance structures"* [1, p. 69]. For example, where a data provider shares data with a start-up as part of an innovation acceleration programme – there could be four different data environments linked together in a data flow: (1) the data provider environment; (2) the start-up environment; (3) the intermediary environment (i.e. where data are hosted); and (4) the innovation acceleration programme environment. [See Annex 4 – for further information].[31]

Example of technical content

James Graves [37] outlines three key technical areas to focus on during data flow mapping (the following has been paraphrased from the original – see [37] for full information):

1. *"Policy and standards"* – Consider the rules that govern how and what types of data are permitted to flow: (i) externally: in and out of the organisation; and (ii) internally: via *"various zones"*. Moreover, examine the standards utilised in order to configure data flows. [37]
2. "*Architecture and design"* – Ascertain where data are situated *"at rest"* and *"in motion"* by identifying *"databases", "applications", "users and groups",* and *"current controls"*. [37]
3. *"Technical controls"* – Focus on the technical controls that: (a) monitor data *"at rest"* and *"in motion"*; and, (b) manage the data flows through access restrictions. [37]

Example of organisational content

Nicola Fulford and Krysia Oastler [5, pp. 7-8] highlight six key organisational areas to focus on during data flow mapping (the following has been paraphrased from the original – see [5, pp. 7-8] for full information):

1. *"The reasons/purposes for processing personal data"* – Consider the legal basis for processing personal data. [5, p. 7]
2. *"Key stakeholders"* – Identify the key stakeholders and their roles in a data processing activity. [5, p. 7]
3. *"Types/categories and sub-categories of personal data"* – Review details about datasets including their attributes. [5, p. 7]
4. *"Source and location of personal data"* – For example, examine the *"data entry point"* and data storage arrangements.
5. *"[W]ith whom personal data are shared or disclosed"* – Document the types of data sharing relationships between the organisation and other third parties, e.g. *"group companies, with suppliers and services providers, with public/official authorities (such as law enforcement and tax authorities), regulators and with business partners/sponsors."* [5, p. 8]
6. *"[D]ocument the relevant retention periods or retention criteria for that data"***.** [5, p. 8]

3.2.3 Key stakeholders

It is crucial to identify the key stakeholders that are involved with a data processing activity [5, p. 7] and involve them in the data mapping process. Where it is possible, the person overseeing data flow mapping should be the designated data protection officer [7, p. 8].

3.2.4 Role of provenance

Access to robust provenance information is an advantage for those who are mapping data flows.[32] Provenance information can be defined as data about data – e.g. information pertaining to the

[31] For further examples of data flow maps see [36, pp. 6-8].

[32] For further information on how to "document your data" see: [76]; for further background information about provenance see: [77].

origins, licensing, versioning, (non-)personal nature and quality of a particular dataset. For many datasets, their status as personal and non-personal may change throughout their lifecycle as they enter different environments, undergo various (re)uses and a number of transformations.

Therefore, provenance information should be able to facilitate the mapping of past and current data flows that relate to the particular (version of a) dataset that a data owner plans to share or re-use. By understanding the provenance, data owners will be able to assess factors such as:

- Information about any previous anonymisation assessment(s) and/or data impact assessments conducted internally and/or by third parties.
- What versions of the dataset exist and in what form (e.g. does the raw personal data exist as well as a pseudonymised version of the dataset).
- Whether and how the particular dataset has been shared and (re-)used before and by whom.
- If there are plans to share versions of the current dataset as different products.
- The history of control over the dataset, e.g. who has had access and (re)use of the previous versions.

3.3 Mapping data flow exercises

Effective mapping of data flows – by its very nature – require the data owner to assess the individual circumstances that surround the specific, planned data processing activity under consideration together with the regulatory framework. For those who are inexperienced with data flow mapping to those who would like to maintain their skills, it can be useful to practise this approach.

The following three scenarios have been designed (by the report authors) in order to practise mapping data flows for different types of data. These three scenario-based data flow mapping exercises may be used by an individual or as part of a group-based exercise.[33]

3.3.1 Scenario A: Medical Research

Brief overview: You are part of a privately-funded medical research team who are focused on an examination of the (non-) effective treatment of *condition A*. Your research relies on large-scale data analysis of health data from a variety of sources, including private health clinics, pharmaceutical companies, survey companies and personal health-trackers.

You have been given access to *dataset x* to re-use as part of your research into the treatment of *condition A*.

***Dataset x* – some further information:**

- Patient-monitoring data, including the course of treatment of administered to 1278 patients with *condition A.*
- Individual-level data.
- Collected by medical staff working at *Hospital Z*, *Ward G* during 2010-2015.
- Direct identifiers are masked (e.g. no patient names appear in the dataset).
- The data provider is a data market place that claims *dataset x* is anonymised.

1. Sketch a data situation model that maps the flow of *Dataset x* from *Ward G* of *Hospital Z* to the medical research team.
2. Your medical research team aims to publish a research paper in an open access journal and, where possible, the underlying datasets. Add this activity to your data situation model sketch for *Dataset x*.

[33] Note that the ODI has also produced workshop material on mapping data ecosystems – see [36, pp. 9-15]. However, this ODI workshop guidance (and the wider report) does not share the same focus as this Data Pitch Deliverable 3.5 (i.e. data flow mapping for GDPR-compliance).

3.3.2 Scenario B: Pollution-Reduction Strategy

Brief overview: You are part of policy team tasked with the development of a strategy to reduce pollution in the middle of *City A*. Your analysis relies on data from a variety of sources, including traffic monitoring data, real-time sensor readings monitoring levels of pollution, and information from public transport providers.

A private bus firm – that operates within *City A* and the surrounding areas – has given you access to *dataset y* to re-use as part of your strategy development.

***Dataset y*– some further information:**

- Passenger-monitoring data taken from registered bus cards that details the individual journeys taken by each passenger during 2017 (where this data is available).
- The number of non-registered passengers who use the bus each day.
- Direct identifiers are masked (e.g. no passenger names appear in the dataset.)
- Age ranges are recorded.
- The private bus firm claims that *dataset y* is anonymised.

1. Sketch a data situation model that maps the flow of *Dataset y* from the private bus firm to the policy team.
2. Your policy team aims to publish this strategy on the City Council's website for consultation and, where possible, the underlying datasets. Add this activity to your data situation model sketch for *Dataset y*.

3.3.3 Scenario C: Predicting Future Product Trends

Brief overview: You are part of product strategy and marketing team tasked with the enrichment of your company's product portfolio through future trends predictions. Your analysis relies on data from a variety of sources, including internal customer data, survey data and information from competitors about new product releases.

A digital analytics company that analyses market trends has given you access to *dataset z* to re-use for your future trends predictions.

***Dataset z* – some further information:**

- Large-scale data taken from a social media platform.
- Click-rates for advertisements on the social media platform.
- Online tracking information (e.g. what websites have been visited by social platform users).
- Predicted preferences from social media data.
- The digital analytics company claims that *dataset z* is anonymised.

1. Sketch a data situation model that maps the flow of *Dataset z* from the social media platform to the product strategy and marketing team.
2. A member of your product and strategy team aims to use this analysis as part of a talk at an international corporate event that focuses on future products in your sector. Add this activity to your data situation model sketch for *Dataset z*.

3.4 Summary: mapping data flows

- Data flow mapping is one way in which those involved with data processing activities can demonstrate compliance with the GDPR.
- Data flow mapping can be used at enterprise-level and/or dataset-level.
- Data flow mapping has numerous benefits, including gap-identification and risk mitigation.
- The content of a data flow map is most crucial – not the format it takes.
- An effective data flow map will take into consideration both the technical and organisational aspects pertaining to a particular data situation.
- Access to robust provenance information is an advantage for mapping data flow activities.

4. Part C – The development of training tools

4.1 D.3.5: dissemination

As aforementioned, the key rationale for this toolkit update is to raise-awareness of the important role that data flow mapping can play in responsible data sharing and re-usage within open innovation programmes; in particular with GDPR-compliance. This aim is achieved via the guidance provided in (i) this report – and further value-added legal training materials (derived from Parts A and B of this report) in the form of (ii) a prototype e-learning tool on data protection and the basics of mapping data flows ("the prototype e-learning tool") and (iii) a workshop.

4.2 The prototype e-learning tool

4.2.1 Purpose

As previously stated, the purpose of the three paper-based legal decision-trees is to help communicate some of the key aspects of the GDPR in a simple way to data owners; by representing a key series of concepts and their outcomes. A decision was taken to make these decision-trees interactive through the development of the prototype e-learning tool (accessible via http://pz-wp-test.synote.io/ [last accessed 4 June 2018]) in order to enable tailored responses to a data owner's planned data processing activity. Note: screen-shots from this prototype e-learning tool are located in Annex 1 of this report.

4.2.1 Technical development

This section provides a brief overview of the technical development of the prototype tool, which was built via the following three stages:

Stage 1: Design.

- **An agile co-design process.** An agile co-design process was adopted to suit the ongoing development of the legal decision-trees.
- **Mock-up screen shots.** From the outset, mock-up screen shots of the tool were shared between the content designers and developer. These mock-up screen shots were used to ensure that the real requirements were addressed in the generation of potential applications.
- **Use of a content management system (CMS).** A CMS system was selected so as to provide the non-developer (i.e. one of the content designers) with a simple and straight-forward means for (later) customisation of the content and workflow. Therefore, it was important for the developer to evaluate several CMSs in order to select the most appropriate CMS that met the specified requirements.

Stage 2: Implementation.

- **WordPress.** WordPress was used and deployed in the implementation phase.
- **Three plug-ins.** The following three WordPress plugins were used during the tool development:
  - i) **Elementor Page Builder –** This was applied in order to create the basic webpages.
  - ii) **Chained Quiz** – This was installed and used in order to: (a) create the decision-tree questions; and (b) manage the decision-tree logic (i.e. where the next question depends on the answer to the previous question).
  - iii) **Popup Builder** – This was activated in order to control and manage pop-up windows (that are used to display further information about a given term within the tool).

Stage 3: Testing.

- **Three test processes:** The prototype e-learning tool was tested via the following three processes:
  - i) **Workflow test.** This test assessed whether the pages of the tool were continuously displayed in the correct order.
  - ii) **Logic test.** This test validated that the tool would lead users to different pages depending on their answers to specific-questions.

iii) **User test.** This test asked end users to utilise the tool and provide feedback. In accordance with this feedback, the developer would then apply any necessary amendments to the tool.

4.3 Legal training workshop (May 2018)

As part of the Data Pitch programme, a legal training workshop – on data protection and the basics of mapping data flows – was delivered to invited SMEs in May 2018. This workshop provided an opportunity to test the prototype e-learning tool and receive feedback from data owners (see Annex 3 for further information).

4.4 Summary

The guidance given on data protection and the basics of mapping data flows is provided via three modes of communication:

1. Paper-based – this deliverable report.
2. Workshop – the workshop materials in Annex 2 can be re-used.
3. Prototype e-learning tool – an interactive version of the three legal decision-trees.

There is scope to improve the current version of the prototype e-learning tool – see Annex 3 for further information.

5. Conclusion

5.1 Key points

The Legal and Privacy Toolkit v2 (D3.5) is a toolkit update that extends the data protection guidance provided in the first version of the toolkit. Ultimately, this report aims to show that anonymisation assessment is a crucial part of any planned data processing activity. It addresses the D3.5. objective outlined by the Grant Agreement by providing: (i) data owners with a basis of legal guidance about GDPR-compliance (explored in Part A) in order to take advantage of (ii) the data flow mapping approach (the basics of which are outlined in Part B) that can be used in the creation of data situation models to further support anonymisation assessment.

5.1.1 Data spectrum, context and purpose

The nature of data is changeable – and ultimately predicated on the specific context and purpose of a (planned) data processing activity. For instance, anonymised data can be de-identified – and the same dataset can be considered as non-personal and personal under different sets of circumstances. Therefore, just because a dataset was used for non-personal purposes in the past does not mean that it cannot be utilised for personal purposes in the future. In consequence, it is crucial for data owners to be cognisant of the significant impact the context and purpose of a planned data situation has on every aspect of an assessment of anonymisation practices.

5.1.2 Data flow mapping

While data flow mapping is not a panacea for GDPR-compliance, it is a useful tool to employ so that: gaps between the regulatory framework and how data are processed in practice are revealed; (potential) high-risk data processing activities are identified and risks can be mitigated; individuals are well-positioned to make good decisions about data processing; and, the areas that require further legal training and guidance are exposed.

After a data owner has mapped their data flows, the next step will be to create (i) a risk register, (ii) a catalogue of risks and (iii) associated control measures in order to demonstrate appropriate mitigation of risks.

5.1.3 Suggested further reading

- Article 29 Data Protection Working Party, “Opinion 04/2007 on the concept of personal data (01248/07/EN WP 136),” 20 June 2007. [Online]. Available: http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf. [Accessed 31 May 2018].
- --, “Opinion 05/2014 on Anonymisation Techniques (0829/14/EN; WP216),” 10 April 2014. [Online]. Available: http://www.pdpjournals.com/docs/88197.pdf. [Accessed 3 April 2018].
- --, “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 (17/EN WP 248),” 4 April 2017 (Adopted). [Online]. Available: http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236. [Accessed 27 May 2018]
- Conference of the Independent Data Protection Authorities of the Bund and the Länder, *The Standard Data Protection Mode*
- CNIL, Privacy Impact Assessment (PIA): Methodology (how to carry out a PIA) (edition of June, 2015)
- --, Privacy Impact Assessment (PIA): Tools (templates and knowldge bases) (edition of June, 2015)
- --, Privacy Impact Assessment (PIA): Methodology (edition of February, 2018) <https://www.cnil.fr/sites/default/files/atoms/files/cnil-pia-1-en-methodology.pdf> accessed 25 May 2018
- --, Privacy Impact Assessment (PIA): Templates (edition of February, 2018) <https://www.cnil.fr/sites/default/files/atoms/files/cnil-pia-2-en-templates.pdf> accessed 25 May 2018
- Information Commissioner's Office (ICO), “Anonymisation: managing data protection risk code of practice,” November 2012. [Online]. Available: https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf. [Accessed 3 April 2018].

- --, “Big data, artificial intelligence, machine learning and data protection (20170904; v2.2),” (in particular Chapter 3: Compliance Tools, pp. 58-61) 3 March 2017. [Online]. Available: https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf. [Accessed 5 April 2018].
- --, “Determining what is personal data (v1.1),” [Online]. Available: https://ico.org.uk/media/for-organisations/documents/1554/determining-what-is-personal-data.pdf.
- --, “Guide to the General Data Protection Regulation (GDPR)”, [Online]. Available at: https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/. [Accessed 4 June 2018]. Including: “Lawful basis for processing,” [Online]. Available: https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/#ib3. [Accessed 30 May 2018].
- --, “What is personal data? – A quick reference guide (V1.1),” [Online]. Available: https://ico.org.uk/media/for-organisations/documents/1549/determining_what_is_personal_data_quick_reference_guide.pdf. [Accessed 28 March 2018].
- J. Clark, “Legislative Comment - GDPR series: building a compliance programme,” Privacy & Data Protection, vol. 17, no. 3, pp. 7-9, 2017.
- J. Graves, “Data flow management: why and how,” Network Security, no. 1, pp. 5-6, 2017.
- LINDDUN Privacy Threat Modelling, [Online]. Available: https://linddun.org/index.php. [Accessed 31 May 2018].
- M. Elliot, E. Mackey , K. O’Hara and C. Tudor , “UK Anonymisation Network (UKAN): The Anonymisation Decision-Making Framework,” 2016. [Online]. Available: http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf. [Accessed 20 February 2018].
- M. Hansen, M. Jensen and M. Rost, "Protection Goals for Privacy Engineering," 2015 IEEE Security and Privacy Workshops (SPW), San Jose, CA, USA, 2015, pp. 159-166. doi:10.1109/SPW.2015.13
- N. Fulford and K. Oastler , “People, processes, technology - a how to guide to data mapping,” Privacy & Data Protection, vol. 16, no. 8, pp. 6-8, 2016.
- R. Hu, S. Stalla-Bourdillon, M. Yang, V. Schiavo and V. Sassone, “Bridging Policy, Regulation and Practice?A techno-legal Analysis of Three Types of Data in the GDPR,” in *Data Protection and Privacy*, Hart Publishing, 2017, [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3034261 [Accessed 4 June 2018].

5.2 Next steps

During the course of the programme, the toolkit is a living document. The final version of the toolkit is due to be submitted in December 2019 at the end of the programme. The final deliverable – D3.9. – will focus on the legal and privacy aspects of transnational, cross-sector data sharing in open innovation.

6. References

[1] M. Elliot, E. Mackey , K. O'Hara and C. Tudor , "UK Anonymisation Network (UKAN): The Anonymisation Decision-Making Framework," 2016. [Online]. Available: http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf. [Accessed 20 February 2018].

[2] Article 29 Data Protection Working Party, "Opinion 05/2014 on Anonymisation Techniques (0829/14/EN; WP216)," 10 April 2014. [Online]. Available: http://www.pdpjournals.com/docs/88197.pdf. [Accessed 3 April 2018].

[3] REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL 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 (GDPR)," Official Journal of the European Union, 27 April 2016. [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN. [Accessed 13 March 2018].

[4] S. Stalla-Bourdillon and A. Knight, "D3.1 – Legal and Privacy Toolkit v1.0," Data Pitch H2020-ICT-2016-1; Project Number: 732506 , 30 June 2017. [Online]. Available: http://www.datapitch.eu/wp-content/uploads/2017/06/PUBLIC-LEGAL-AND-PRIVACY-TOOLKIT-VERSION-1.0-DELIVERABLE-8.1-FINAL-30-JUNE-2017.pdf. [Accessed 22 February 2018].

[5] N. Fulford and K. Oastler , "People, processes, technology - a how to guide to data mapping," *Privacy & Data Protection,* vol. 16, no. 8, pp. 6-8, 2016.

[6] K. Knight, "IAPP Global Privacy Summit 2013 Presentation - Data Flow Mapping: The Good, the Bad, and the Ugly," 7 March 2013. [Online]. Available: https://iapp.org/media/presentations/13Summit/S13_Good_Bad_Ugly_PPT.pdf. [Accessed 31 May 2018].

[7] J. Clark, "Legislative Comment - GDPR series: building a compliance programme," *Privacy & Data Protection,* vol. 17, no. 3, pp. 7-9, 2017.

[8] IT Governance , "Data flow mapping under the EU GDPR," [Online]. Available: https://www.itgovernance.co.uk/gdpr-data-mapping. [Accessed 31 May 2018].

[9] S. Stalla-Bourdillon and A. Knight, "Anonymous Data v. Personal Data - A False Debate: An EU Perspective on Anonymization, Pseudonymization and Personal Data," *Wisconsin International Law Journal,* 2017.

[10] Information Commissioner's Office (ICO), "Determining what is personal data (v1.1)," [Online]. Available: https://ico.org.uk/media/for-organisations/documents/1554/determining-what-is-personal-data.pdf.

[11] Information Commissioner's Office (ICO) , "What is personal data? – A quick reference guide (V1.1)," [Online]. Available: https://ico.org.uk/media/for-organisations/documents/1549/determining_what_is_personal_data_quick_reference_guide.pdf. [Accessed 28 March 2018].

[12] Article 29 Data Protection Working Party, "Opinion 04/2007 on the concept of personal data (01248/07/EN WP 136)," 20 June 2007. [Online]. Available: http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf. [Accessed 31 May 2018].

[13] S. Wood, "ICO blog: Anonymisation – opportunities and risks," Information Commissioner's Office (ICO) Blog, 16 November 2012. [Online]. Available: https://iconewsblog.org.uk/2012/11/16/ico-blog-anonymisation-opportunities-and-risks/. [Accessed 16 March 2018].

[14] Information Commissioner's Office (ICO), "Anonymisation: managing data protection risk code of practice," November 2012. [Online]. Available: https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf. [Accessed 3 April 2018].

[15] Z. Alexin, “Does fair anonymization exist?,” *International Review of Law, Computers & Technology,* vol. 28, no. 1, pp. 21-44, 2013.

[16] M. Oswald , “Something Bad Might Happen: Lawyers, Anonymization, and Risk,” *XRDS: Crossroads, The ACM Magazine for Students - The Complexities of Privacy and Anonymity,* vol. 20, no. 1, pp. 22-26, 2013.

[17] Information and Privacy Commissioner, Ontario, Canada, “Looking Forward: De-identification Developments – New Tools, New Challenges,” May 2013. [Online]. Available: https://www.ipc.on.ca/wp-content/uploads/2013/05/pbd-de-identification_developments.pdf. [Accessed 13 March 2018].

[18] K. Brimsted, “Anonymisation - a process living on borrowed time?,” *Privacy & Data Protection,* vol. 14, no. 7, pp. 3-5, 2014.

[19] J. Clark, “Legislative Comment - GDPR series: anonymisation and pseudonymisation,” *Privacy & Data Protection,* vol. 18, no. 1, pp. 10-12, 2017.

[20] “Data Protection Bill [HL] 2017-19: Progress of the Bill,” [Online]. Available: https://services.parliament.uk/bills/2017-19/dataprotection.html. [Accessed 20 June 2018].

[21] “Data Protection Act 2018,” [Online]. Available: http://www.legislation.gov.uk/ukpga/2018/12/part/6/crossheading/offences-relating-to-personal-data/enacted. [Accessed 20 June 2018].

[22] S. Joyee De and D. Le Métayer, “Privacy Risk Analysis,” in *Synthesis Lectures on Information Security, Privacy, &Trust (eBook)*, Morgan & Claypool Publishers, 2016, pp. 1-117.

[23] Article 29 Data Protection Working Party, “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 (17/EN WP 248),” 4 April 2017 (Adopted). [Online]. Available: http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236. [Accessed 27 May 2018].

[24] Agencia Española de Protección de Datos (AEPD) , “GUIA PRÁCTICA DE Análisis de riesgos en los tratamientos de datos personales sujetos al RGPD,” [Online]. Available: https://iapp.org/media/pdf/resource_center/AnalisisDeRiesgosRGPD.pdf. [Accessed 15 March 2018].

[25] Information Commissioner's Office (ICO), “Examples of processing ‘likely to result in high risk’,” [Online]. Available: https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/data-protection-impact-assessments-dpias/examples-of-processing-likely-to-result-in-high-risk/. [Accessed 26 May 2018].

[26] R. Clark, “Quality Assurance for Security Applications of Big Data,” in *European Intelligence and Security Informatics Conference (EISIC'16)*, Uppsala, Sweden, 2016.

[27] R. D. Riley, P. C. Lambert and G. Abo-Zaid, “Meta-analysis of individual participant data: rationale, conduct, and reporting,” *The British Medical Journal,* vol. 340, p. c221, 2010.

[28] Norwegian Centre for Research Data (NSD), “Individual Level Data,” [Online]. Available: http://www.nsd.uib.no/nsd/english/individualdata.html. [Accessed 14 March 2018].

[29] Information Commissioner's Office (ICO) , “Guide to the General Data Protection Regulation (GDPR): Lawful basis for processing,” [Online]. Available: https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/#ib3. [Accessed 30 May 2018].

[30] Information Commissioner's Office (ICO), “ICO: Lawful Basis Interactive Guidance Tool,” [Online]. Available: https://ico.org.uk/for-organisations/resources-and-support/lawful-basis-interactive-guidance-tool/. [Accessed 30 May 2018].

[31] PricewaterhouseCoopers (PwC), “General Data Protection Regulation: Anonymisation and pseudonymisation,” 2017. [Online]. Available: https://www.pwc.com.cy/en/publications/assets/general-data-protection-regulation-anonymisation-and-pseudonymisation-january-2017.pdf. [Accessed 3 April 2018].

[32] European Union Agency for Network and Information Security (ENISA), “Handbook on Security of Personal Data Processing,” December 2017. [Online]. Available: https://www.enisa.europa.eu/publications/handbook-on-security-of-personal-data-processing. [Accessed 12 March 2018].

[33] Information Commissioner's Office (ICO), “Data protection impact assessments,” [Online]. Available: https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/. [Accessed 30 May 2018].

[34] Commission nationale de l'informatique et des libertés (CNIL), “The open source PIA software helps to carry out data protection impact assesement,” 29 January 2018. [Online]. Available: https://www.cnil.fr/en/open-source-pia-software-helps-carry-out-data-protection-impact-assesment. [Accessed 30 May 2018].

[35] Information Commissioner's Office (ICO), “How do we document our processing activities?,” [Online]. Available: https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/documentation/how-do-we-document-our-processing-activities/. [Accessed 31 May 2018].

[36] Open Data Institute (ODI), “Mapping data ecosystems,” 23 March 2018. [Online]. Available: https://docs.google.com/document/d/1vSqoHOYT5u6vrCHIebCS0rze0gWwXOspeEowWzwake8/edit. [Accessed 20 June 2018].

[37] J. Graves, “Data flow management: why and how,” *Network Security,* no. 1, pp. 5-6, 2017.

[38] LINDDUN Privacy Threat Modelling, [Online]. Available: https://linddun.org/index.php. [Accessed 31 May 2018].

[39] IT Governance, “Data Flow Mapping Tool,” [Online]. Available: https://www.itgovernance.co.uk/shop/Product/data-flow-mapping-tool. [Accessed 31 May 2018].

[40] B. Treacy, “Expert comment,” *Privacy & Data Protection,* vol. 13, no. 2, p. 2, 2012.

[41] B. Lubarsky, “Re-Identification of “Anonymized” Data,” *Georgetown Law Technology Review,* vol. 202, no. 1, 2017.

[42] The Organisation for Economic Co-operation and Development (OECD), “Glossary of Statistical Terms: Aggregation,” 10 June 2013 (last updated). [Online]. Available: https://stats.oecd.org/glossary/detail.asp?ID=68. [Accessed 14 March 2018].

[43] M. Rouse, “Definition: data aggregation,” TechTarget, September 2005 (last updated). [Online]. Available: http://searchsqlserver.techtarget.com/definition/data-aggregation. [Accessed 14 March 2018].

[44] Agencia Española de Protección de Datos (AEPD), [Online]. Available: https://www.agpd.es/portalwebAGPD/index-iden-idphp.php. [Accessed 15 March 2018].

[45] C. S. Dempwolf, J. Auer and M. D’Ippolito , “Innovation Accelerators: Defining Characteristics Among Startup Assistance Organizations (Small Business Administration, Office of Advocacy under contract number SBAHQ-13-M-0197),” October 2014. [Online]. Available: https://www.sba.gov/sites/default/files/rs425-Innovation-Accelerators-Report-FINAL.pdf. [Accessed 22 March 2018].

[46] Open Data Institute (ODI), “The Data Spectrum: The Data Spectrum helps you understand the language of data,” [Online]. Available: https://theodi.org/about-the-odi/the-data-spectrum/. [Accessed 22 March 2018].

[47] Data Pitch - Innovation Programme, “About Data Pitch,” [Online]. Available: https://datapitch.eu/about-us/start-up/. [Accessed 22 March 2018].

[48] R. Thomas , “Risk, accountability, and binding corporate codes: a "smarter" approach to data protection,” *Privacy & Data Protection,* vol. 13, no. 7, pp. 3-6, 2013.

[49] The International Association of Privacy Professionals (iapp) and OneTrust: Privacy Management Software, “IAPP-OneTrust Research: Bridging ISO 27001 to GDPR,” 27

March 2018. [Online]. Available: https://iapp.org/media/pdf/resource_center/IAPP-OneTrust-Bridging-ISO-GDPR-final.pdf. [Accessed 27 March 2018].

[50] M. Hintze and K. El Emam, “Privacy Analytics - White Paper: Comparing the Benefits of Pseudonymization and Anonymization Under the GDPR,” 17 August 2017. [Online]. Available: https://iapp.org/media/pdf/resource_center/PA_WP2-Anonymous-pseudonymous-comparison.pdf. [Accessed 3 April 2018].

[51] C. Wiper, “Information Commissioner's Office (ICO) blog: Seven things you should know about the ICO’s big data report,” 28 July 2014. [Online]. Available: https://iconewsblog.org.uk/2014/07/28/seven-things-you-should-know-about-the-icos-big-data-report/. [Accessed 3 April 2018].

[52] Data Pitch, “Challenges: first call for the 2017-2018 programme,” [Online]. Available: https://datapitch.eu/challenges/. [Accessed 10 April 2018].

[53] Data Pitch, “Challenges overview: first call for the 2017-2018 programme,” [Online]. Available: https://drive.google.com/file/d/0B9IIZV_CjqLceGVudW1LcWN4M0k/view. [Accessed 10 April 2018].

[54] W. Kerber, “Editorial - Governance of data: exclusive property v access,” *International Review of Intellectual Property and Competition Law,* vol. 47, no. 7, pp. 759-762, 2016.

[55] G. Thomas, “Assigning Data Ownership,” The Data Governance Institute, 28 September 2013. [Online]. Available: http://www.datagovernance.com/assigning-data-ownership/. [Accessed 10 April 2018].

[56] G. Thomas, “Working with Data Stewards,” The Data Governance Institute, 28 September 2013. [Online]. Available: http://www.datagovernance.com/working-with-data-stewards/. [Accessed 10 April 2018].

[57] T. Hoeren, “Big data and the ownership in data: recent developments in Europe,” *European Intellectual Property Review ,* vol. 36, no. 12, pp. 751-754 , 2014.

[58] Oxford English Dictionaries (OED) Online , “Define: anonymise,” [Online]. Available: https://en.oxforddictionaries.com/definition/anonymize. [Accessed 11 April 2018].

[59] A. Hern, “New law could criminalise uncovering personal data abuses, advocate warns,” The Guardian Online, 14 August 2017. [Online]. Available: https://www.theguardian.com/technology/2017/aug/14/data-protection-bill-criminalise-privacy-research-advocate-warns. [Accessed 11 April 2018].

[60] C. M. O’Keefe, S. Otorepec, M. Elliot, E. Mackey and K. O’Hara, “The De-Identification DecisionMaking Framework (CSIRO Reports EP173122 and EP175702),” 18 September 2017. [Online]. Available: https://publications.csiro.au/rpr/download?pid=csiro:EP173122&dsid=DS2. [Accessed 12 April 2018].

[61] Australian Government: , “De-Identification Decision-Making Framework: Office of the Australian Information Commissioner,” [Online]. Available: https://www.oaic.gov.au/agencies-and-organisations/guides/de-identification-decision-making-framework. [Accessed 12 April 2018].

[62] M. Phillips, E. S. Dove and B. M. Knoppers, “Criminal Prohibition of Wrongful Re-identification: Legal Solution or Minefield for Big Data?,” *Journal of Bioethical Inquiry,* vol. 14, no. 4, p. 527–539, 2017.

[63] European Data Protection Supervisor (EDPS): The EU's independent data protection authority , “Glossary: "Article 29 Working Party",” [Online]. Available: https://edps.europa.eu/data-protection/data-protection/glossary/a_en. [Accessed 13 April 2018].

[64] “DIRECTIVE 95/46/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the protection of individuals with regard to the processing of personal data and on the free movement of such data,” 24 October 1995. [Online]. Available: https://edps.europa.eu/sites/edp/files/publication/dir_1995_46_en.pdf. [Accessed 13 April 2018].

[65] European Commission Website, “Article 29 Working Party Newsroom,” [Online]. Available: http://ec.europa.eu/newsroom/article29/news.cfm?item_type=1358. [Accessed 13 April 2018].

[66] OUT-LAW.COM, “ICO: Anonymised data doesn't HAVE to guarantee your privacy,” The Register , 26 November 2012. [Online]. Available: https://www.theregister.co.uk/2012/11/26/anonymising_data_does_not_guarantee_privacy/. [Accessed 13 April 2018].

[67] P. Ralph and S. Ng, “Will there be a new data protection offence for the UK beyond GDPR?,” PricewaterhouseCoopers (PwC): Data protection and privacy gloabl insights, 8 September 2017. [Online]. Available: http://pwc.blogs.com/data_protection/2017/09/will-there-be-a-new-data-protection-offence-for-the-uk-beyond-gdpr.html. [Accessed 13 April 2018].

[68] R. Chirgwin, “UK Data Protection Bill tweaked to protect security researchers: Re-identification of data will not be a crime, as long as you warn the authorities,” The Register, 10 January 2018. [Online]. Available: https://www.theregister.co.uk/2018/01/10/uk_data_protection_bill_tweaked_to_protect_security_researchers/. [Accessed 13 April 2018].

[69] European Union Agency for Fundamental Rights, European Court of Human Rights, and Council of Europe, “Handbook on European data protection law,” June 2014. [Online]. Available: http://fra.europa.eu/en/publication/2014/handbook-european-data-protection-law. [Accessed 13 April 2018].

[70] European Commission (EC) , “Open Innovation Resources: Policy initiatives, funding schemes and support services related to open innovation.,” [Online]. Available: https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-innovation-resources_en. [Accessed 26 May 2018].

[71] J. Polonetsky , O. Tene and K. Finch, “Shades of Gray: Seeing the Full Spectrum of Practical Data De-identification,” *Santa Clara Law Review,* vol. 56, no. 3, 2016.

[72] P. M. Schwartz and D. J. Solove, “The PII Problem: Privacy and a New Concept of Personally Identifiable Information,” *N.Y.U. L.Q. Rev.,* vol. 86, no. 1814, 2011.

[73] The UK Data Service, “Census microdata guide: "Samples of individual person-level records drawn from the 1991, 2001 and 2011 Censuses",” [Online]. Available: https://census.ukdataservice.ac.uk/use-data/guides/microdata. [Accessed 30 May 2018].

[74] UK Data Service, “Census aggregate data guide,” [Online]. Available: https://census.ukdataservice.ac.uk/use-data/guides/aggregate-data. [Accessed 30 May 2018].

[75] R. Hu, S. Stalla-Bourdillon, M. Yang, V. Schiavo and V. Sassone, “Bridging Policy, Regulation and Practice?A techno-legal Analysis of Three Types of Data in the GDPR,” in *Data Protection and Privacy*, Hart Publishing, 2017.

[76] UK Data Service, “Manage Data: Document Your Data - "Make data clear to understand and easy to use",” [Online]. Available: https://www.ukdataservice.ac.uk/manage-data/document. [Accessed 4 June 2018].

[77] L. Moreau, “The Foundations for Provenance on the Web,” *Foundations and Trends in Web Science ,* vol. 2, no. 2–3, pp. 99-241, 2010.

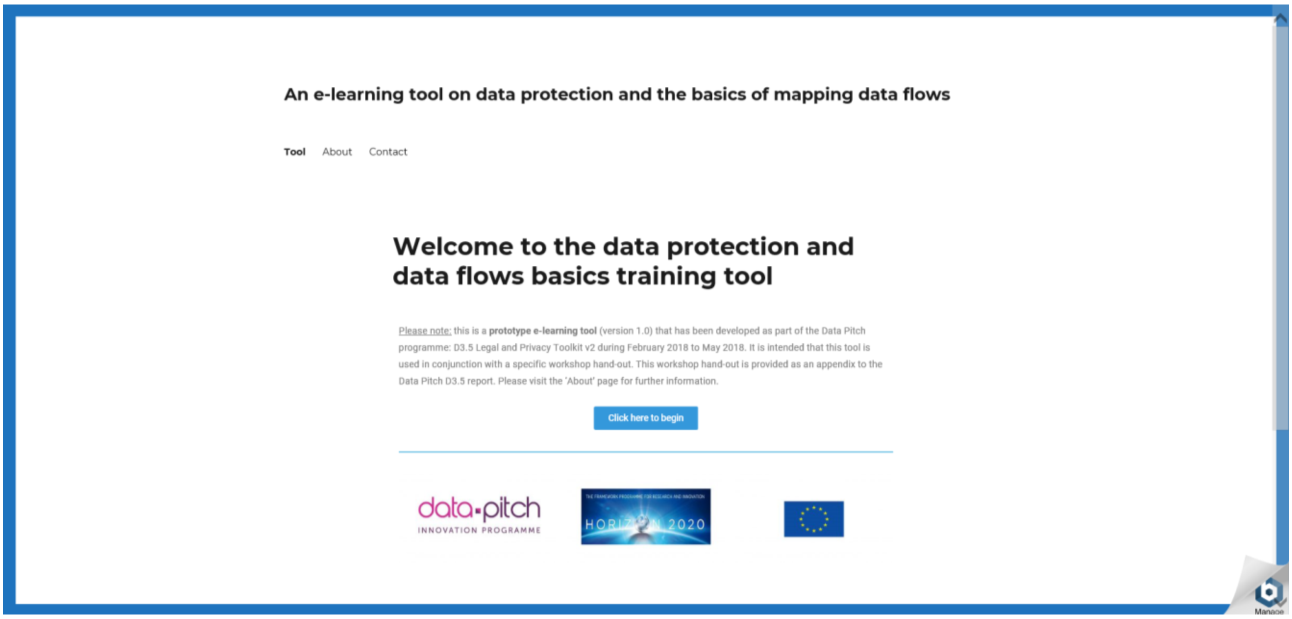
[78] Information Commissioner's Office (ICO), [Online]. Available: https://ico.org.uk/. [Accessed 20 6 2018].

[79] L. Floridi and M. Taddeo, “What is data ethics?,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences,* vol. 374, no. 2083, 2016.

[80] McDermott Will & Emery, “Article 29 working party opinion on the definition of consent: an unambiguous view of the future,” Lexology, 28 September 2011. [Online]. Available: https://www.lexology.com/library/detail.aspx?g=01d43c3a-2a36-4d3d-b0d6-cfdb1d3be067. [Accessed 13 April 2018].

Annex 1: Prototype e-learning tool screen-shots

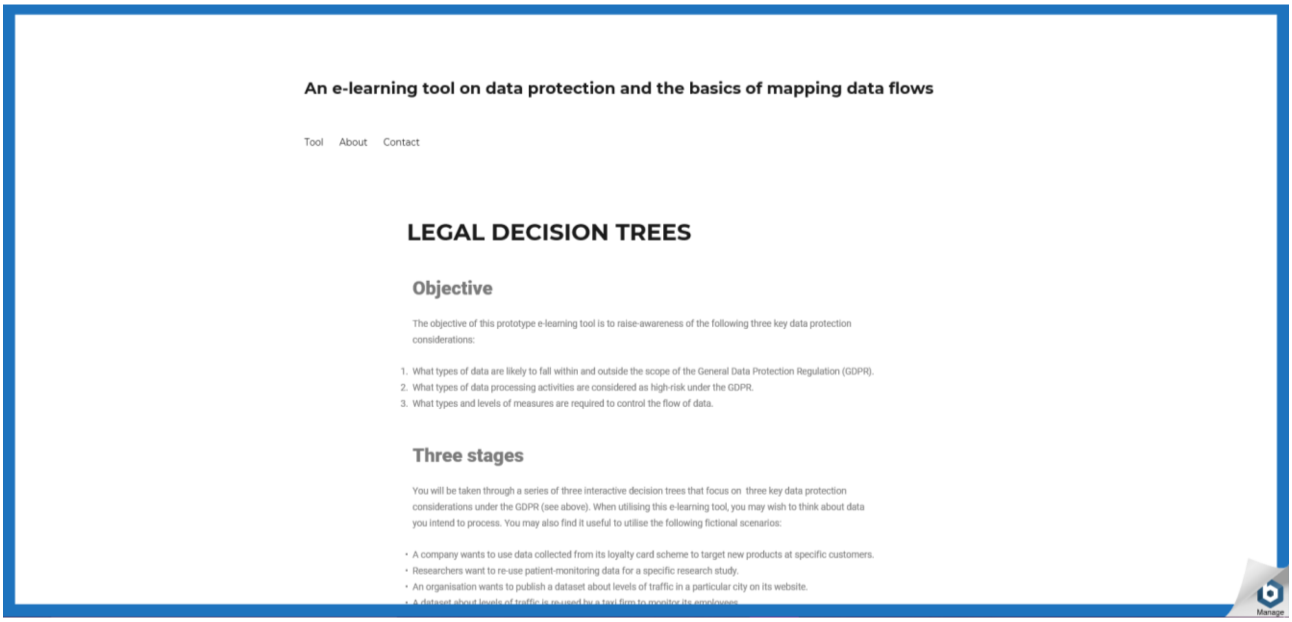
Figure 5 Prototype e-learning tool: screen-shots of introductory pages (in sequence)



An e-learning tool on data protection and the basics of mapping data flows

Welcome to the data protection and data flows basics training tool

Please note: this is a **prototype e-learning tool** (version 1.0) that has been developed as part of the Data Pitch programme: D3.5 Legal and Privacy Toolkit v2 during February 2018 to May 2018. It is intended that this tool is used in conjunction with a specific workshop hand-out. This workshop hand-out is provided as an appendix to the Data Pitch D3.5 report. Please visit the 'About' page for further information.



An e-learning tool on data protection and the basics of mapping data flows

LEGAL DECISION TREES

Objective

The objective of this prototype e-learning tool is to raise-awareness of the following three key data protection considerations:

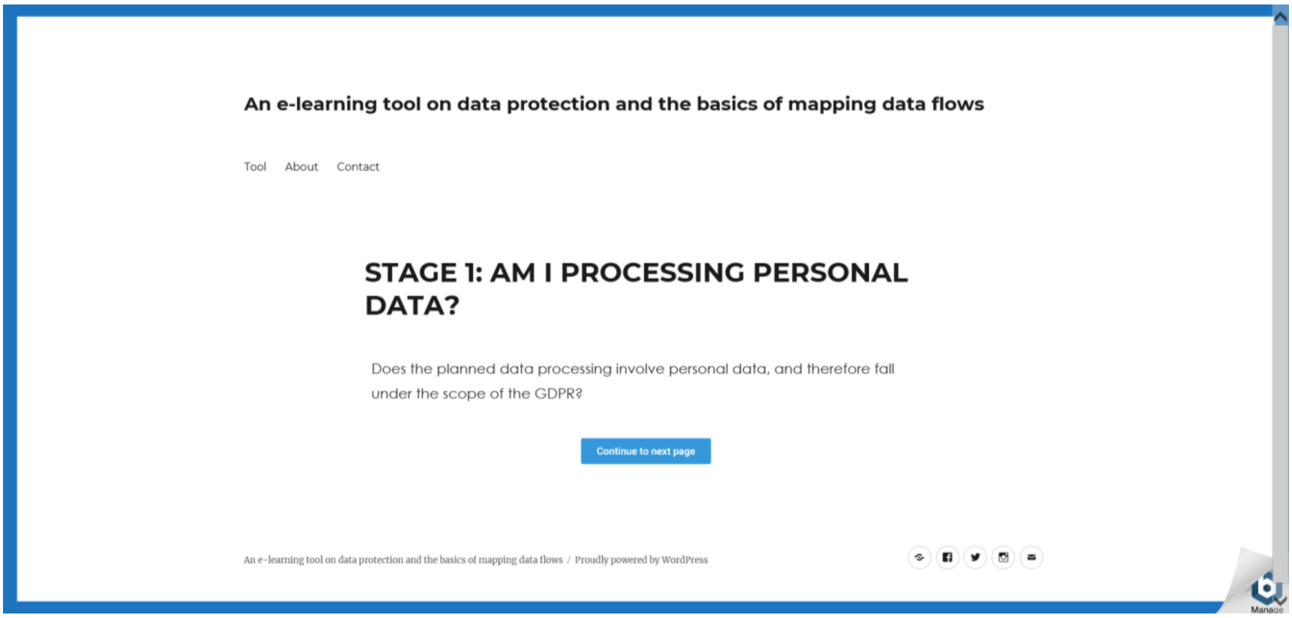
1. What types of data are likely to fall within and outside the scope of the General Data Protection Regulation (GDPR).
2. What types of data processing activities are considered as high-risk under the GDPR.
3. What types and levels of measures are required to control the flow of data.

Three stages

You will be taken through a series of three interactive decision trees that focus on three key data protection considerations under the GDPR (see above). When utilising this e-learning tool, you may wish to think about data you intend to process. You may also find it useful to utilise the following fictional scenarios:

- A company wants to use data collected from its loyalty card scheme to target new products at specific customers.
- Researchers want to re-use patient-monitoring data for a specific research study.
- An organisation wants to publish a dataset about levels of traffic in a particular city on its website.
- A dataset about levels of traffic is re-used by a taxi firm to monitor its employees

Figure 6 Prototype e-learning tool: screen-shots of Legal Decision Tree 1 (in sequence – when all answers to questions are 'No')

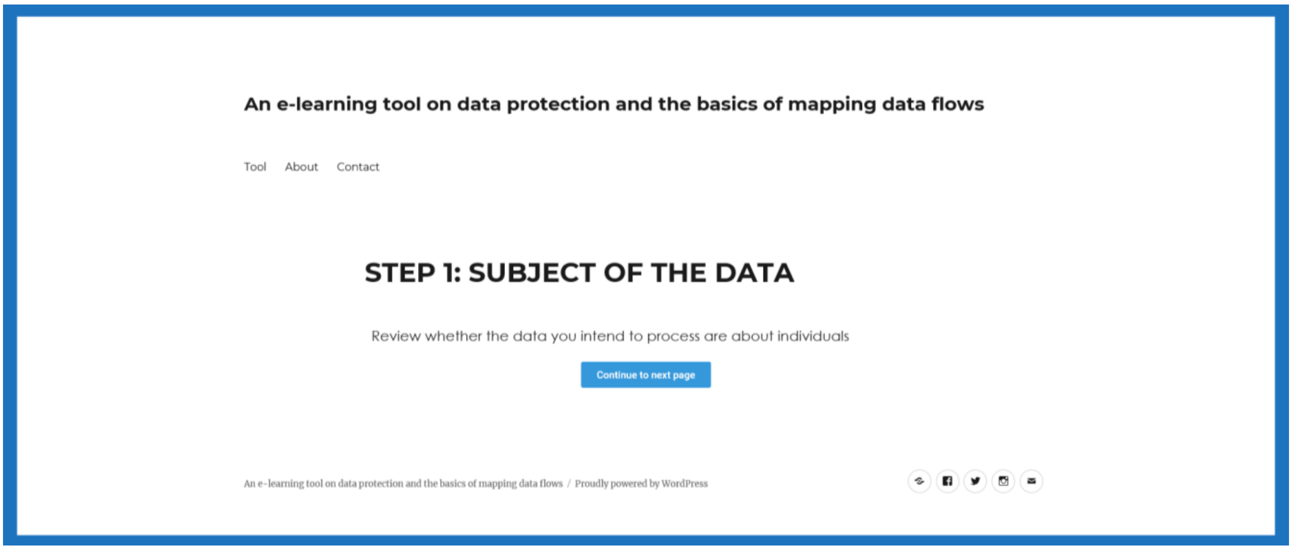


An e-learning tool on data protection and the basics of mapping data flows

STAGE 1: AM I PROCESSING PERSONAL DATA?

Does the planned data processing involve personal data, and therefore fall under the scope of the GDPR?

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress

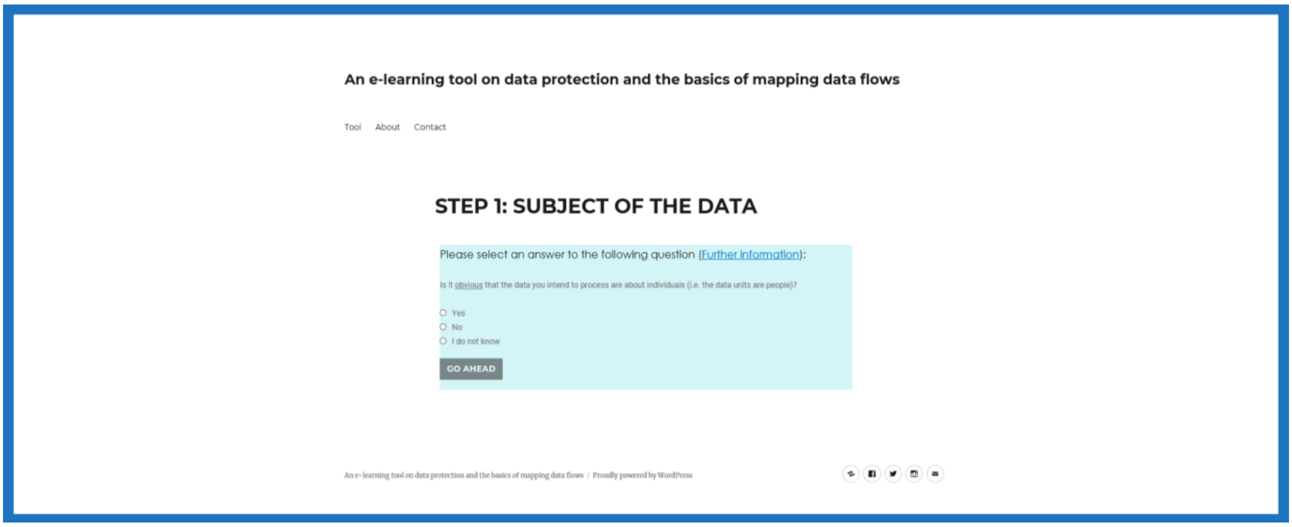


An e-learning tool on data protection and the basics of mapping data flows

STEP 1: SUBJECT OF THE DATA

Review whether the data you intend to process are about individuals

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress



An e-learning tool on data protection and the basics of mapping data flows

Tool About Contact

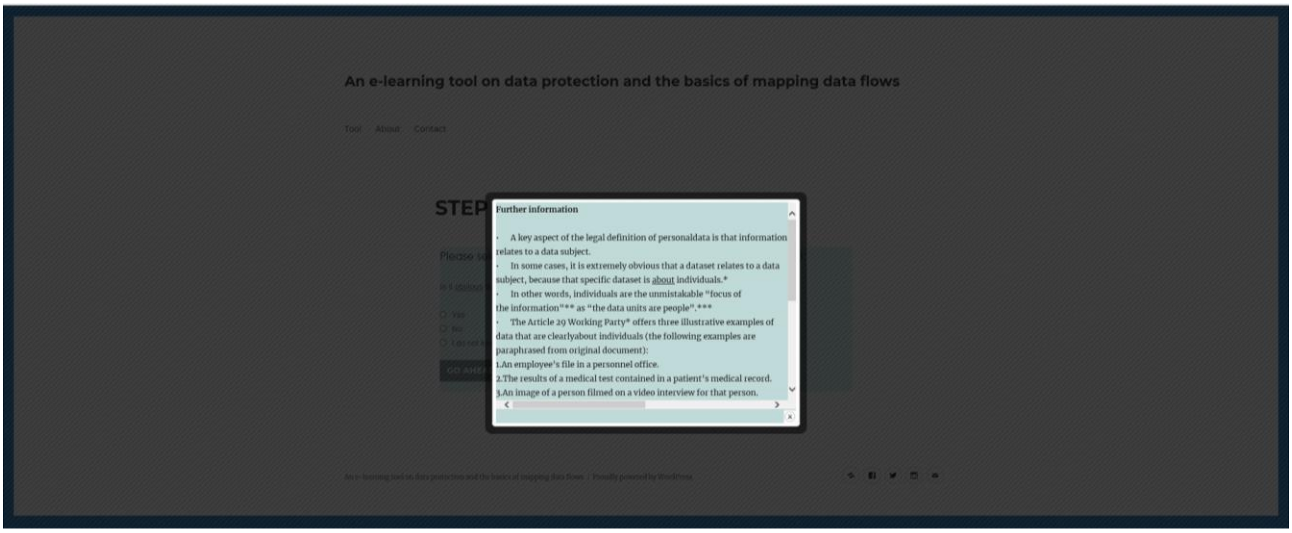
STEP 1: SUBJECT OF THE DATA

Please select an answer to the following question (Further information):

Is it obvious that the data you intend to process are about individuals (i.e. the data units are people)?

○ Yes
○ No
○ I do not know

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress

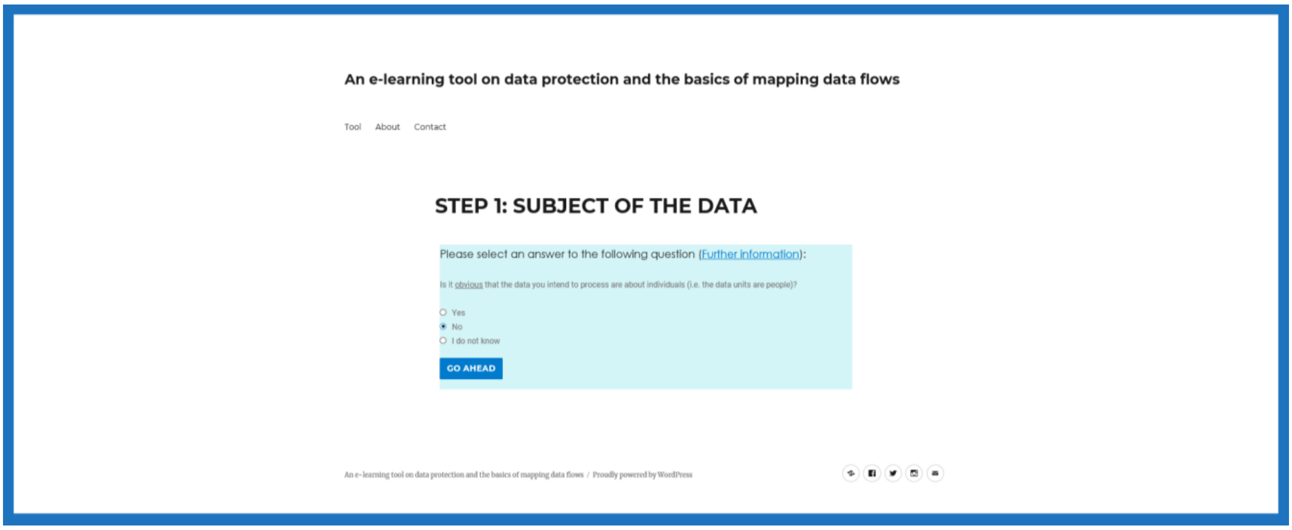


An e-learning tool on data protection and the basics of mapping data flows

Further information

- A key aspect of the legal definition of personaldata is that information relates to a data subject.
- In some cases, it is extremely obvious that a dataset relates to a data subject, because that specific dataset is about individuals.*
- In other words, individuals are the unmistakable "focus of the information"** as "the data units are people".***
- The Article 29 Working Party* offers three illustrative examples of data that are clearlyabout individuals (the following examples are paraphrased from original document):

1.An employee's file in a personnel office.
2.The results of a medical test contained in a patient's medical record.
3.An image of a person filmed on a video interview for that person.



An e-learning tool on data protection and the basics of mapping data flows

Tool About Contact

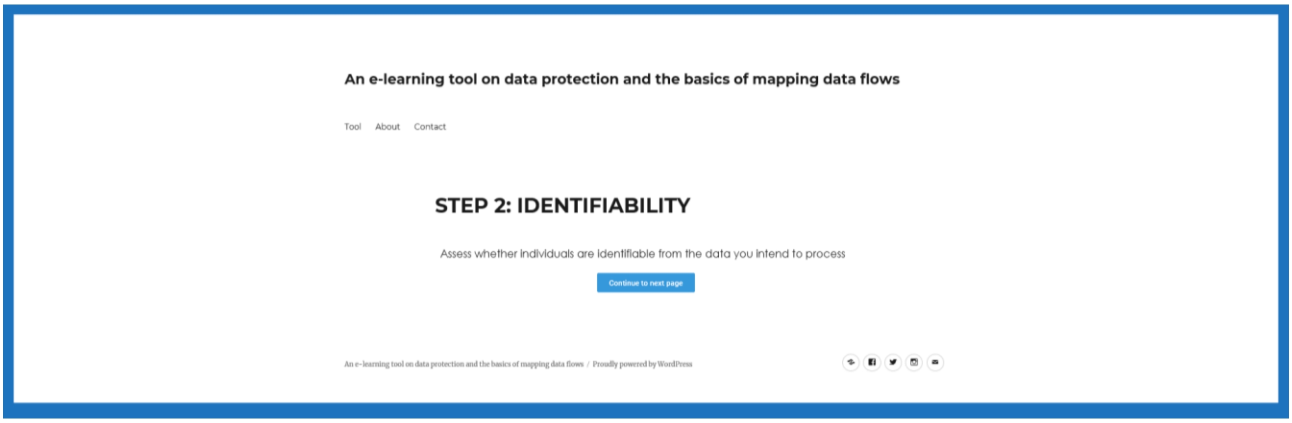
STEP 1: SUBJECT OF THE DATA

Please select an answer to the following question (Further information):

Is it obvious that the data you intend to process are about individuals (i.e. the data units are people)?

○ Yes
● No
○ I do not know

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress

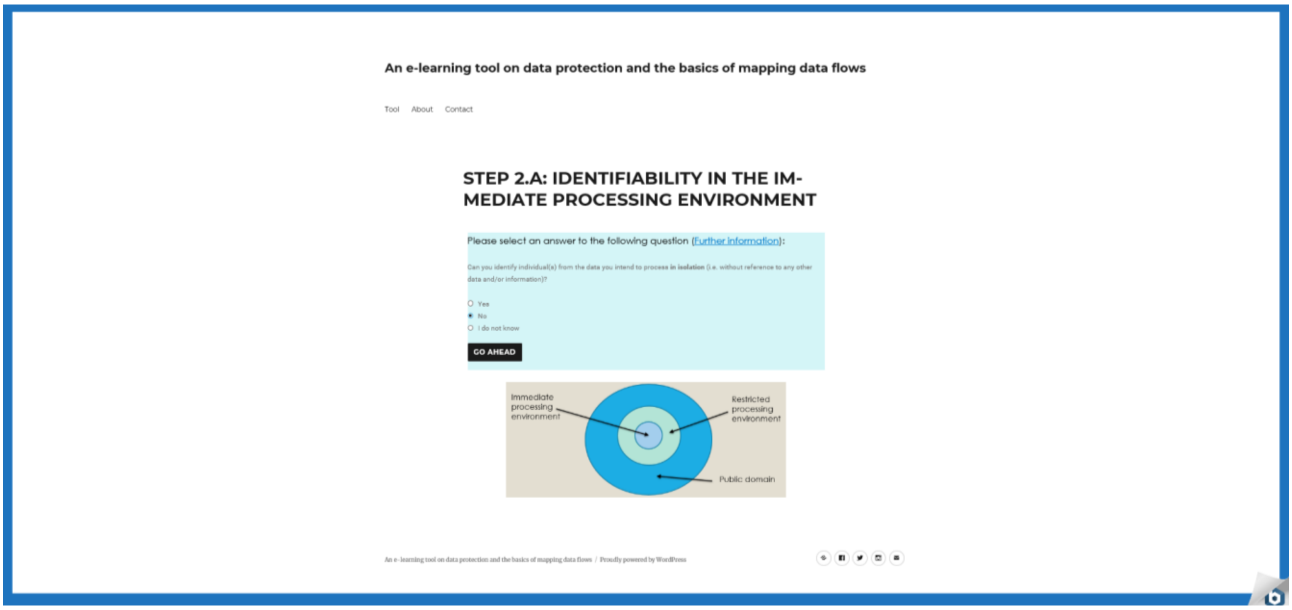


An e-learning tool on data protection and the basics of mapping data flows

STEP 2: IDENTIFIABILITY

Assess whether individuals are identifiable from the data you intend to process

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress



An e-learning tool on data protection and the basics of mapping data flows

STEP 2.A: IDENTIFIABILITY IN THE IMMEDIATE PROCESSING ENVIRONMENT

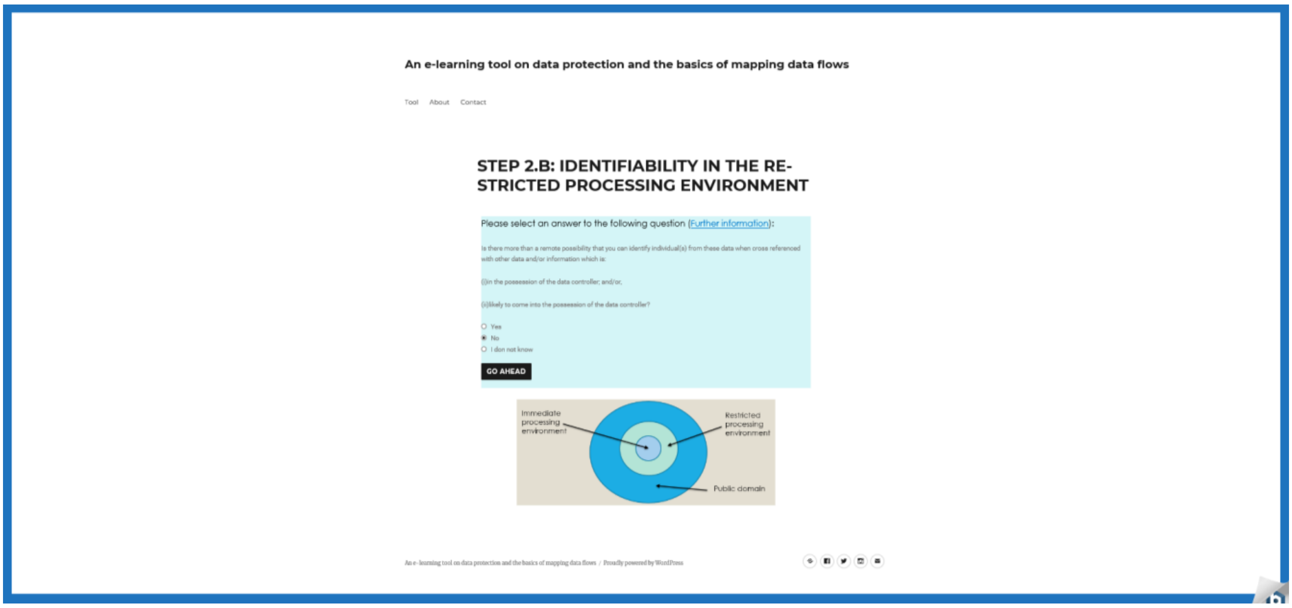
Please select an answer to the following question (Further information):

Can you identify individual(s) from the data you intend to process **in isolation** (i.e. without reference to any other data and/or information)?

○ Yes
● No
○ I do not know

GO AHEAD
Immediate processing environment
Restricted processing environment
Public domain

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress



An e-learning tool on data protection and the basics of mapping data flows

STEP 2.B: IDENTIFIABILITY IN THE RESTRICTED PROCESSING ENVIRONMENT

Please select an answer to the following question (Further information):

Is there more than a remote possibility that you can identify individual(s) from these data when cross referenced with other data and/or information which is:

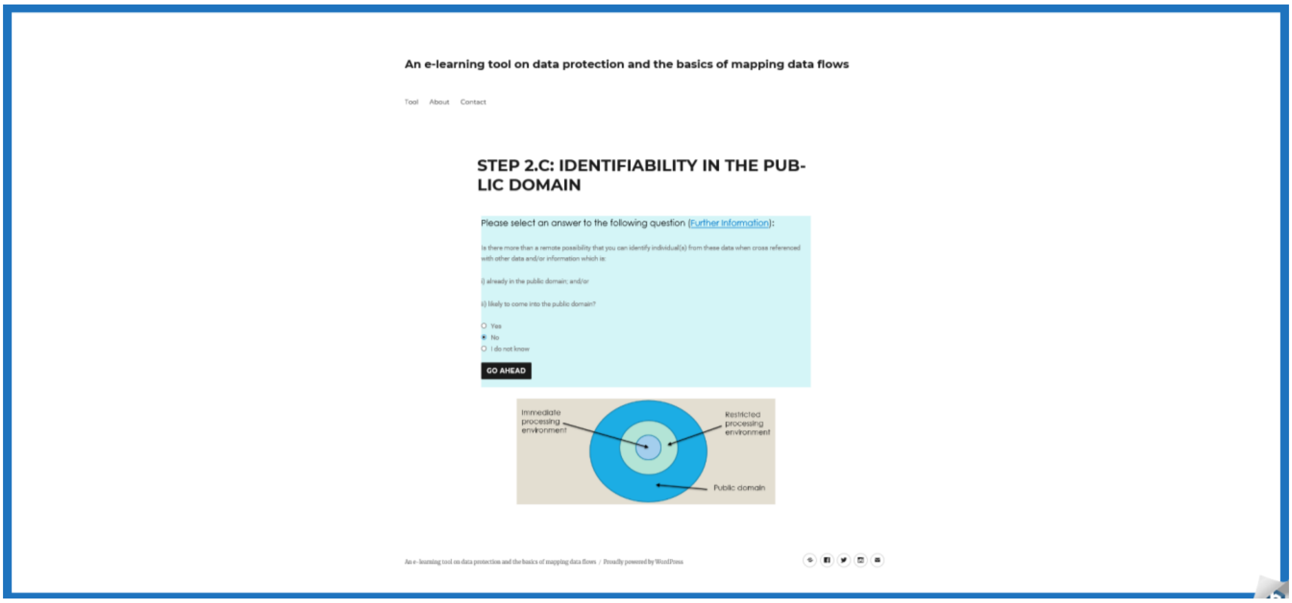
(i)in the possession of the data controller; and/or,

(ii)likely to come into the possession of the data controller?

○ Yes
● No
○ I don not know

GO AHEAD
Immediate processing environment
Restricted processing environment
Public domain

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress



An e-learning tool on data protection and the basics of mapping data flows

Tool About Contact

STEP 2.C: IDENTIFIABILITY IN THE PUBLIC DOMAIN

Please select an answer to the following question (Further Information):

Is there more than a remote possibility that you can identify individual(s) from these data when cross referenced with other data and/or information which is:

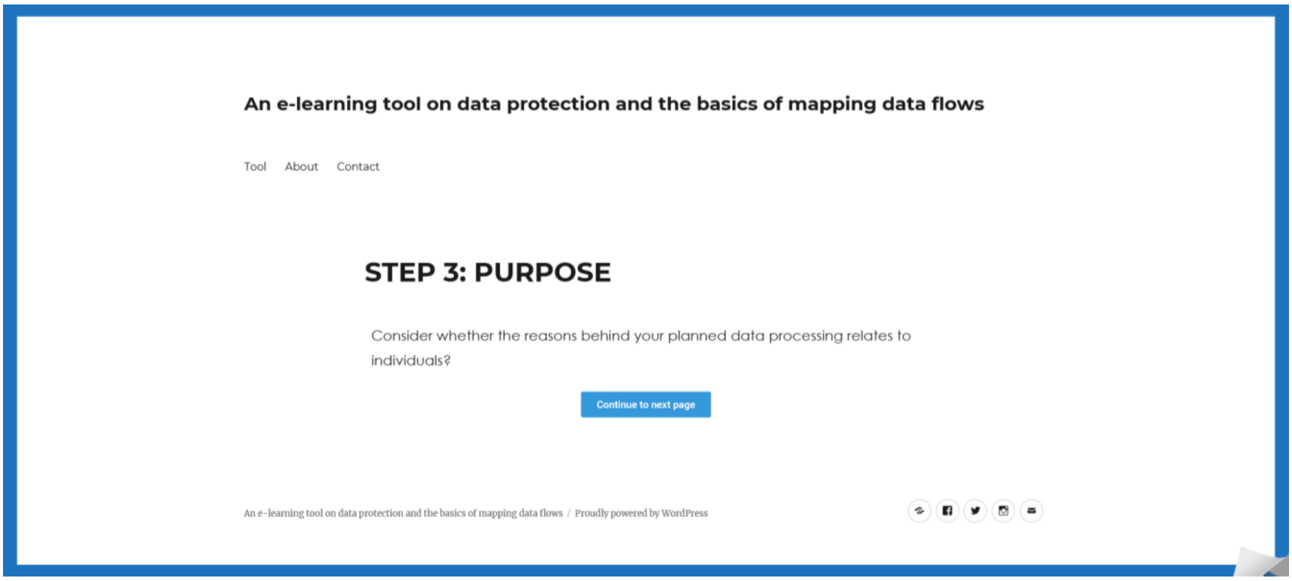
i) already in the public domain; and/or

ii) likely to come into the public domain?

○ Yes
● No
○ I do not know

GO AHEAD
Immediate processing environment
Restricted processing environment
Public domain

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress



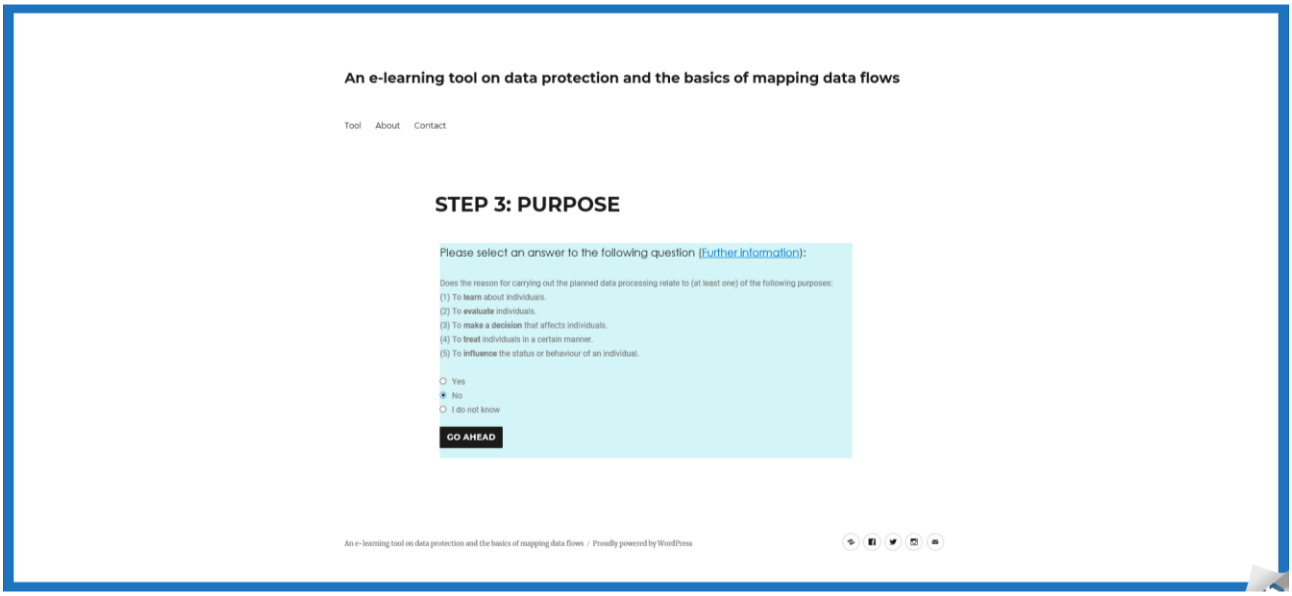
An e-learning tool on data protection and the basics of mapping data flows

Tool About Contact

STEP 3: PURPOSE

Consider whether the reasons behind your planned data processing relates to individuals?

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress


An e-learning tool on data protection and the basics of mapping data flows
Tool
About
Contact

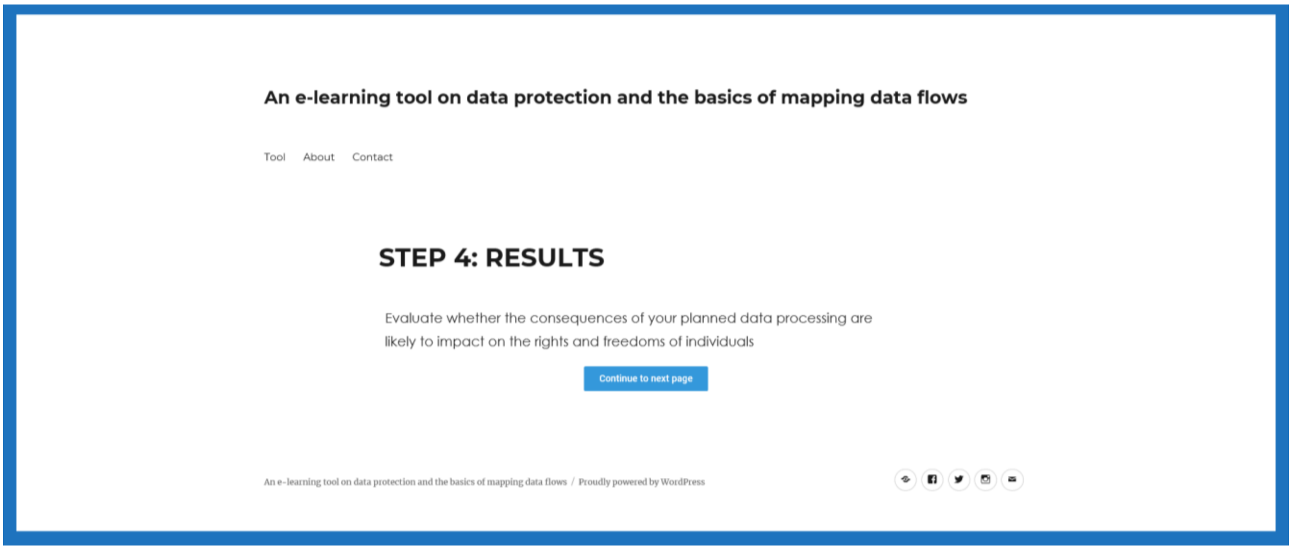
STEP 3: PURPOSE

Please select an answer to the following question (Further information):

Does the reason for carrying out the planned data processing relate to (at least one) of the following purposes:
(1) To **learn** about individuals.
(2) To **evaluate** individuals.
(3) To **make a decision** that affects individuals.
(4) To **treat** individuals in a certain manner.
(5) To **influence** the status or behaviour of an individual.

○ Yes
● No
○ I do not know

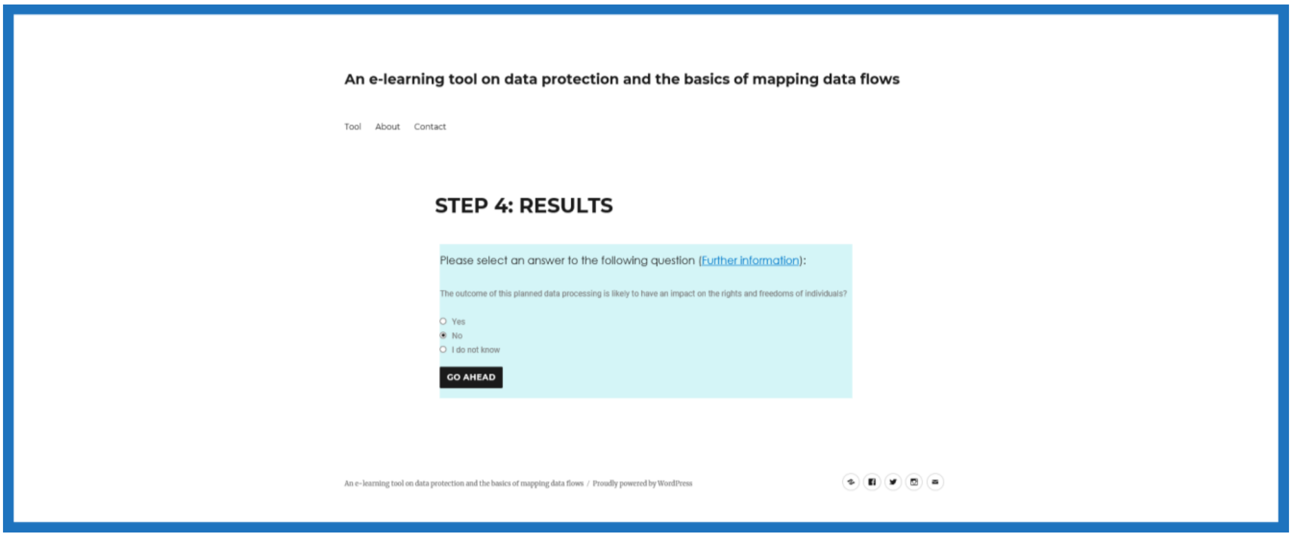
GO AHEAD
An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress


An e-learning tool on data protection and the basics of mapping data flows
Tool
About
Contact

STEP 4: RESULTS

Evaluate whether the consequences of your planned data processing are likely to impact on the rights and freedoms of individuals

Continue to next page
An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress


An e-learning tool on data protection and the basics of mapping data flows
Tool
About
Contact

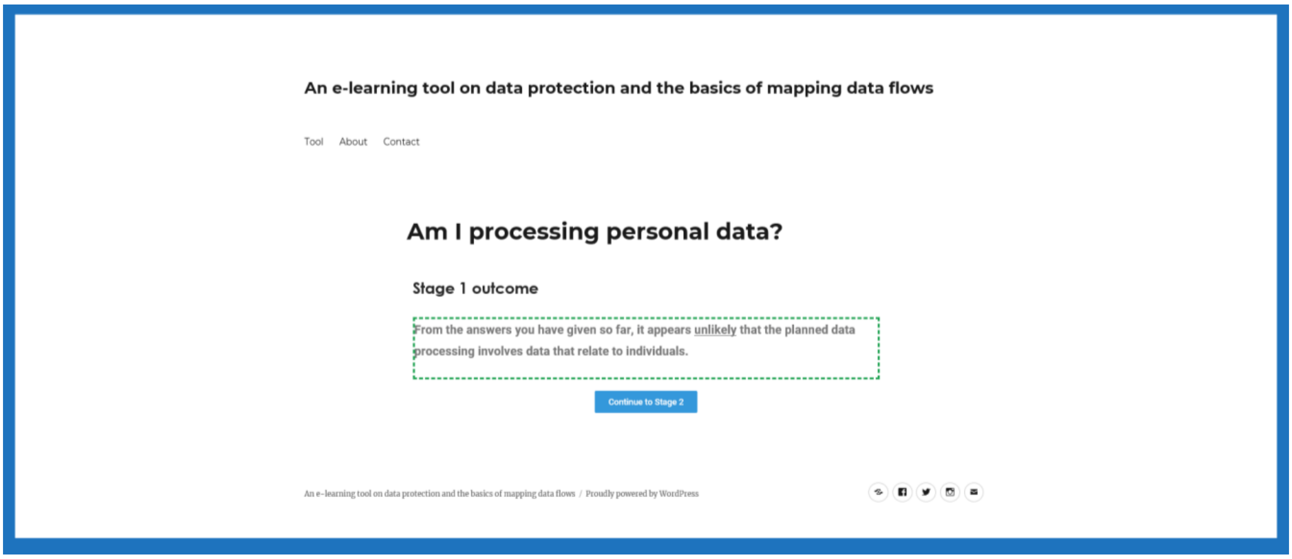
STEP 4: RESULTS

Please select an answer to the following question (Further information):

The outcome of this planned data processing is likely to have an impact on the rights and freedoms of individuals?

○ Yes
● No
○ I do not know

GO AHEAD
An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress



An e-learning tool on data protection and the basics of mapping data flows

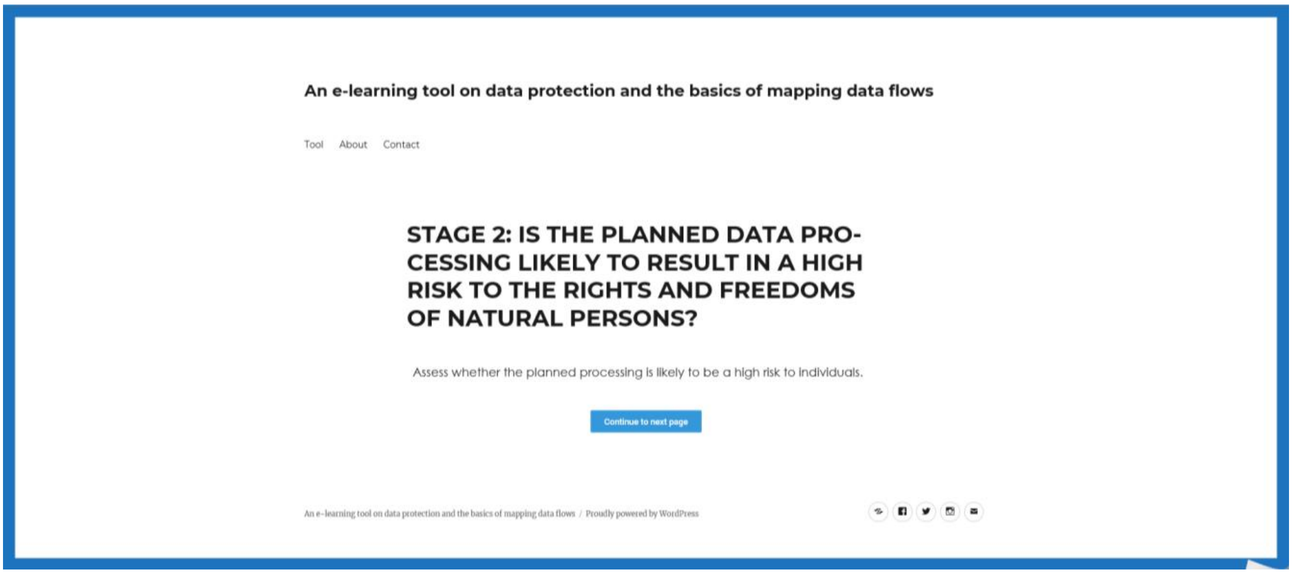
Am I processing personal data?

Stage 1 outcome

From the answers you have given so far, it appears unlikely that the planned data processing involves data that relate to individuals.

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress

Figure 7 Prototype e-learning tool: screen-shots of Legal Decision-Tree 2 (in sequence – when all answers to questions are 'No' – note: sequences shortened)

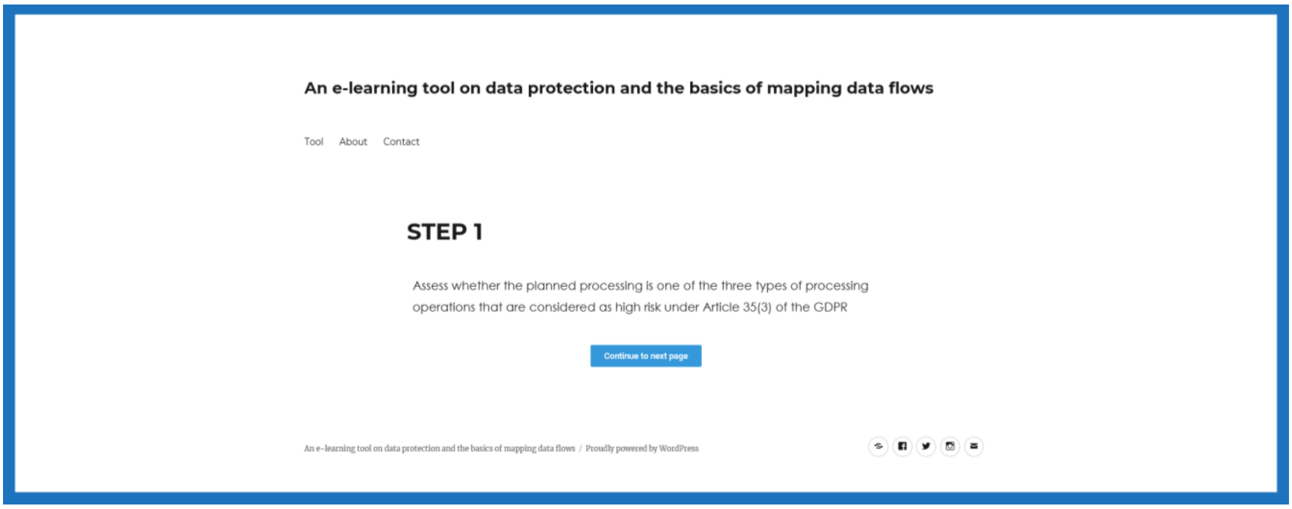


An e-learning tool on data protection and the basics of mapping data flows

STAGE 2: IS THE PLANNED DATA PROCESSING LIKELY TO RESULT IN A HIGH RISK TO THE RIGHTS AND FREEDOMS OF NATURAL PERSONS?

Assess whether the planned processing is likely to be a high risk to individuals.

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress

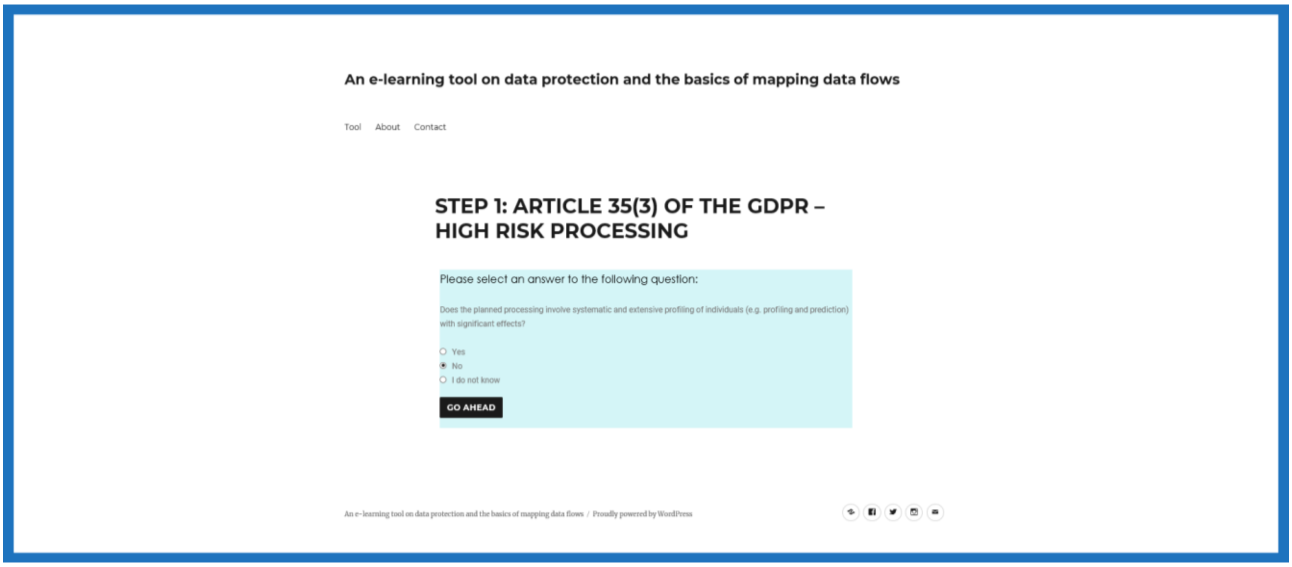


An e-learning tool on data protection and the basics of mapping data flows

STEP 1

Assess whether the planned processing is one of the three types of processing operations that are considered as high risk under Article 35(3) of the GDPR

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress



An e-learning tool on data protection and the basics of mapping data flows

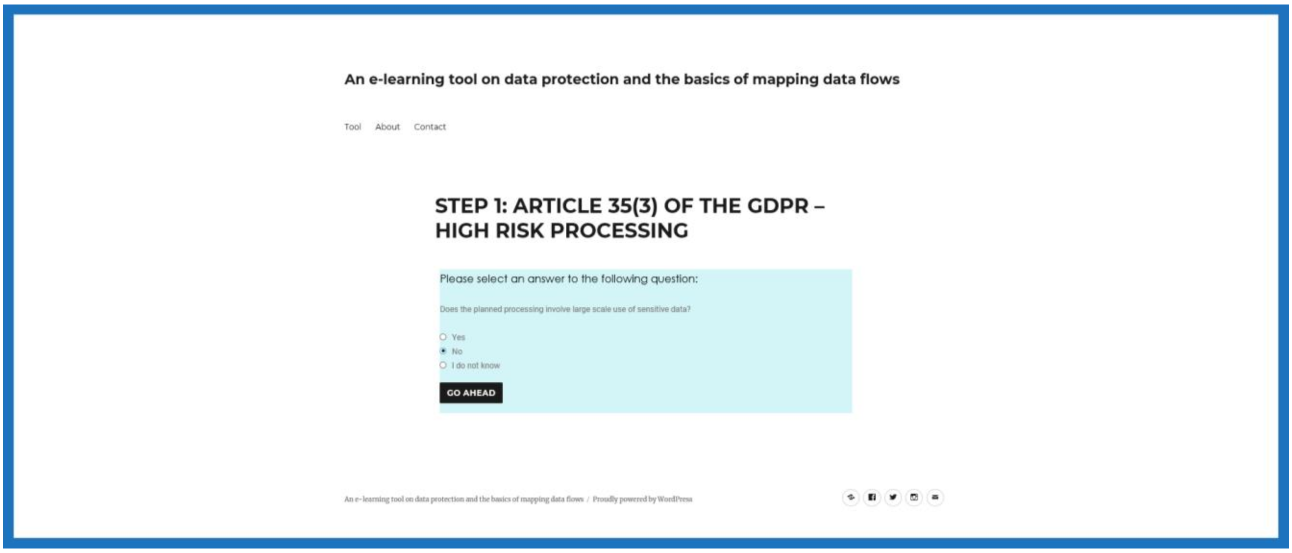
STEP 1: ARTICLE 35(3) OF THE GDPR – HIGH RISK PROCESSING

Please select an answer to the following question:

Does the planned processing involve systematic and extensive profiling of individuals (e.g. profiling and prediction) with significant effects?

○ Yes
● No
○ I do not know

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress



An e-learning tool on data protection and the basics of mapping data flows

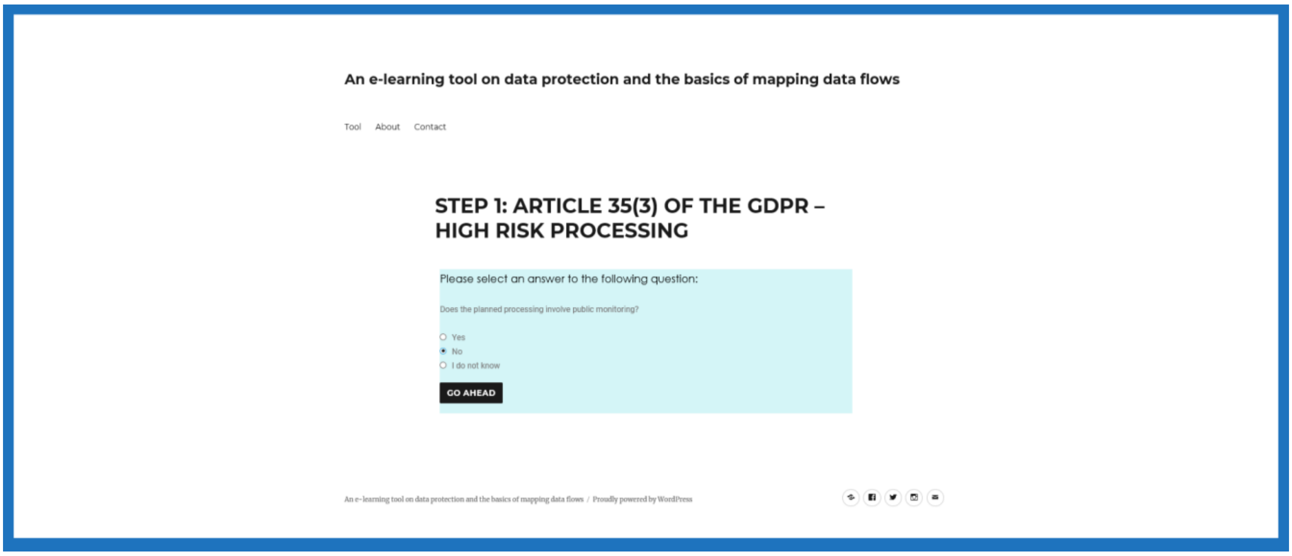
STEP 1: ARTICLE 35(3) OF THE GDPR – HIGH RISK PROCESSING

Please select an answer to the following question:

Does the planned processing involve large scale use of sensitive data?

○ Yes
● No
○ I do not know

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress



An e-learning tool on data protection and the basics of mapping data flows

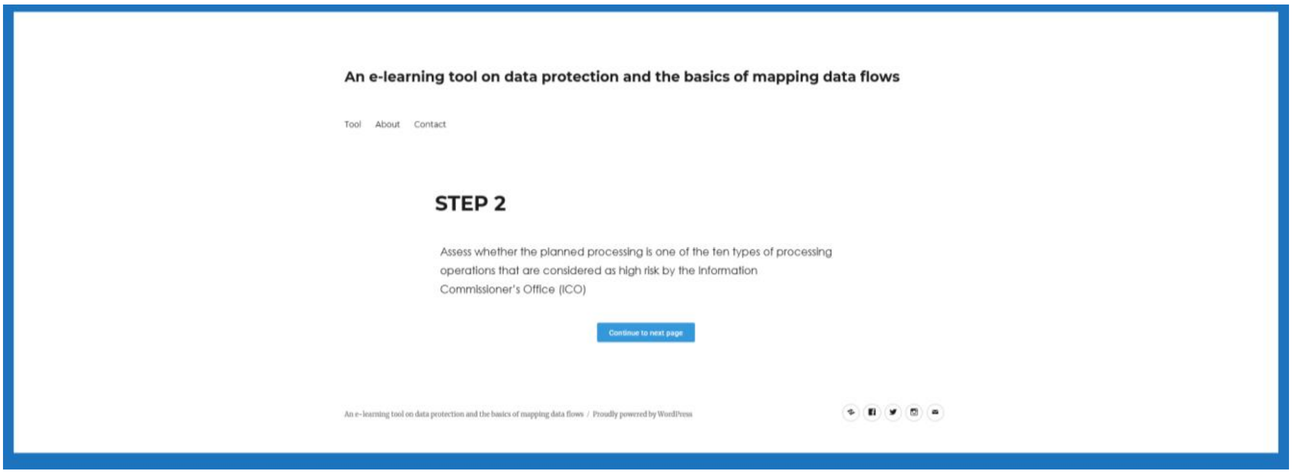
STEP 1: ARTICLE 35(3) OF THE GDPR – HIGH RISK PROCESSING

Please select an answer to the following question:

Does the planned processing involve public monitoring?

○ Yes
● No
○ I do not know

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress

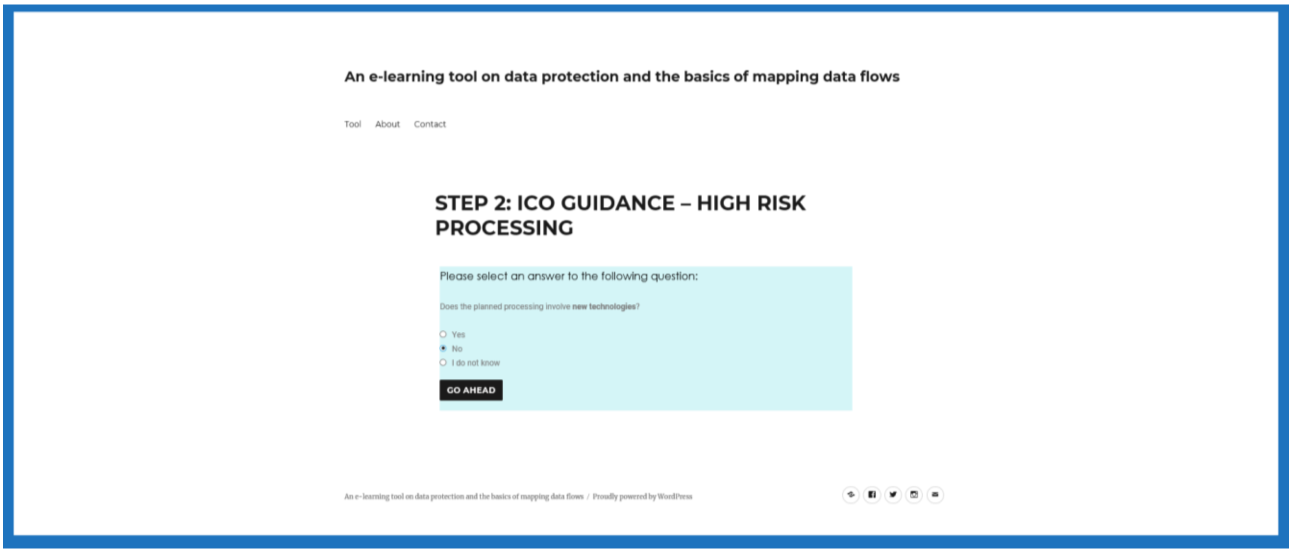


An e-learning tool on data protection and the basics of mapping data flows

STEP 2

Assess whether the planned processing is one of the ten types of processing operations that are considered as high risk by the Information Commissioner's Office (ICO)

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress



An e-learning tool on data protection and the basics of mapping data flows

STEP 2: ICO GUIDANCE – HIGH RISK PROCESSING

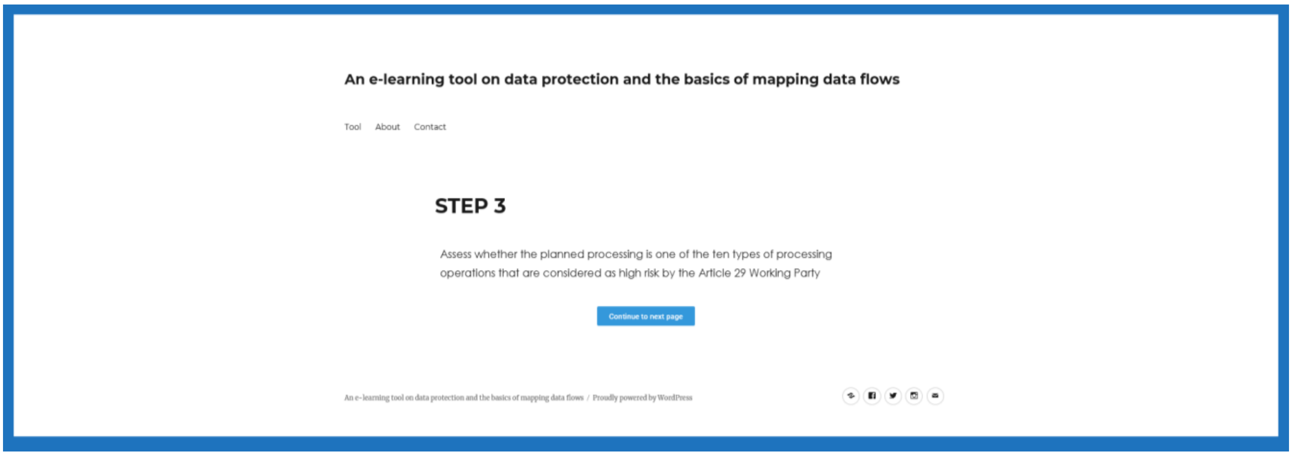
Please select an answer to the following question:

Does the planned processing involve **new technologies**?

○ Yes
● No
○ I do not know

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress

[Note: sequence shortened here. User answers 'No' to remaining nine questions on ICO's ten point criteria. Then user moves to Step 3 below.]

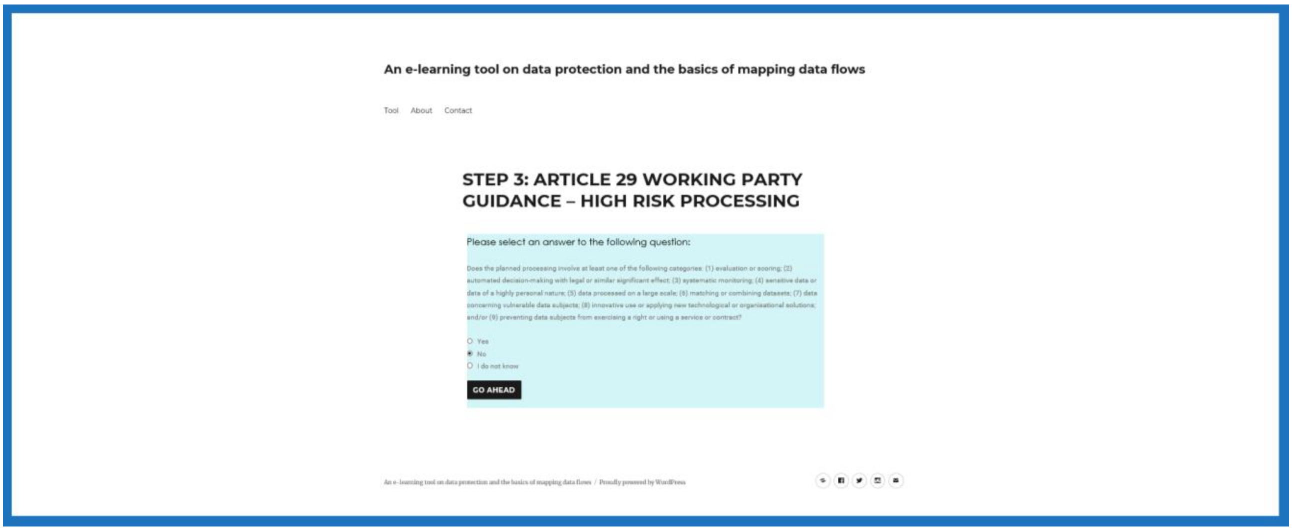


An e-learning tool on data protection and the basics of mapping data flows

STEP 3

Assess whether the planned processing is one of the ten types of processing operations that are considered as high risk by the Article 29 Working Party

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress



An e-learning tool on data protection and the basics of mapping data flows

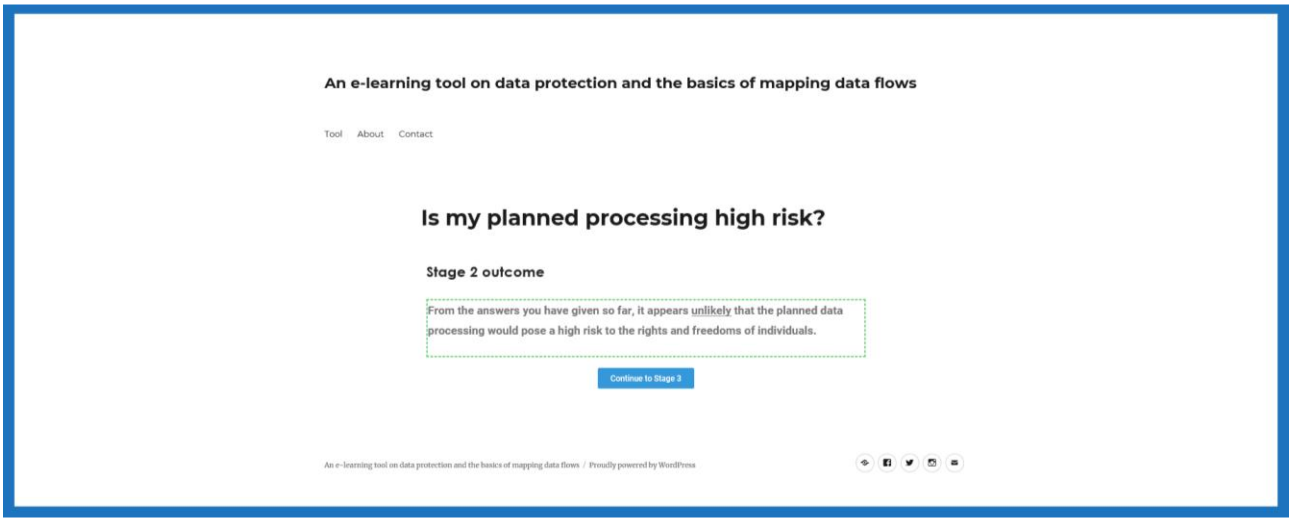
STEP 3: ARTICLE 29 WORKING PARTY GUIDANCE – HIGH RISK PROCESSING

Please select an answer to the following question:

Does the planned processing involve at least one of the following categories: (1) evaluation or scoring; (2) automated decision-making with legal or similar significant effect; (3) systematic monitoring; (4) sensitive data or data of a highly personal nature; (5) data processed on a large scale; (6) matching or combining datasets; (7) data concerning vulnerable data subjects; (8) innovative use or applying new technological or organisational solutions; and/or (9) preventing data subjects from exercising a right or using a service or contract?

○ Yes
● No
○ I do not know

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress



An e-learning tool on data protection and the basics of mapping data flows

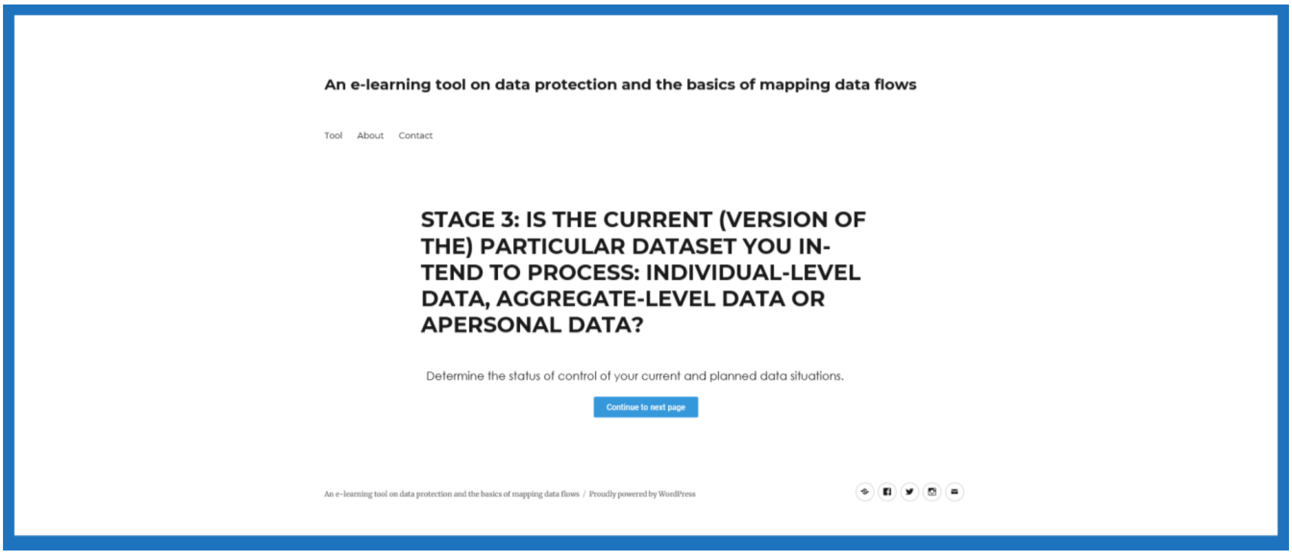
Is my planned processing high risk?

Stage 2 outcome

From the answers you have given so far, it appears unlikely that the planned data processing would pose a high risk to the rights and freedoms of individuals.

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress

Figure 8 Prototype e-learning tool: screen-shots of Legal Decision-Tree 3 (in sequence – focus on Category A: Individual-Level Data example)

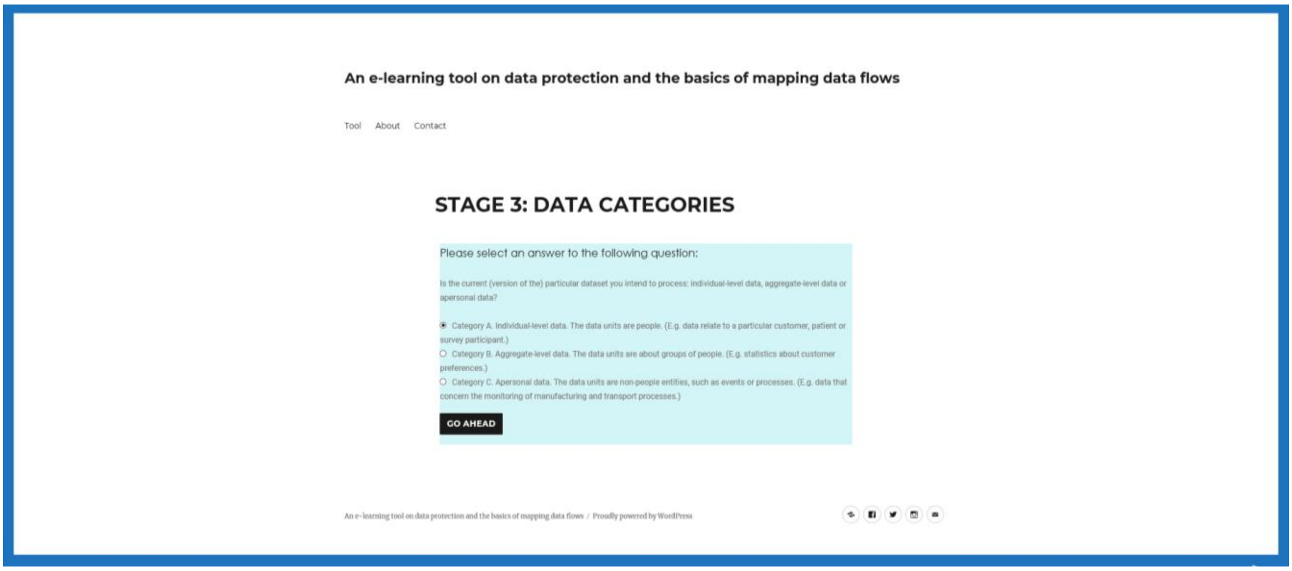


An e-learning tool on data protection and the basics of mapping data flows
Tool
About
Contact

STAGE 3: IS THE CURRENT (VERSION OF THE) PARTICULAR DATASET YOU INTEND TO PROCESS: INDIVIDUAL-LEVEL DATA, AGGREGATE-LEVEL DATA OR APERSONAL DATA?

Determine the status of control of your current and planned data situations.

Continue to next page
An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress



An e-learning tool on data protection and the basics of mapping data flows
Tool
About
Contact

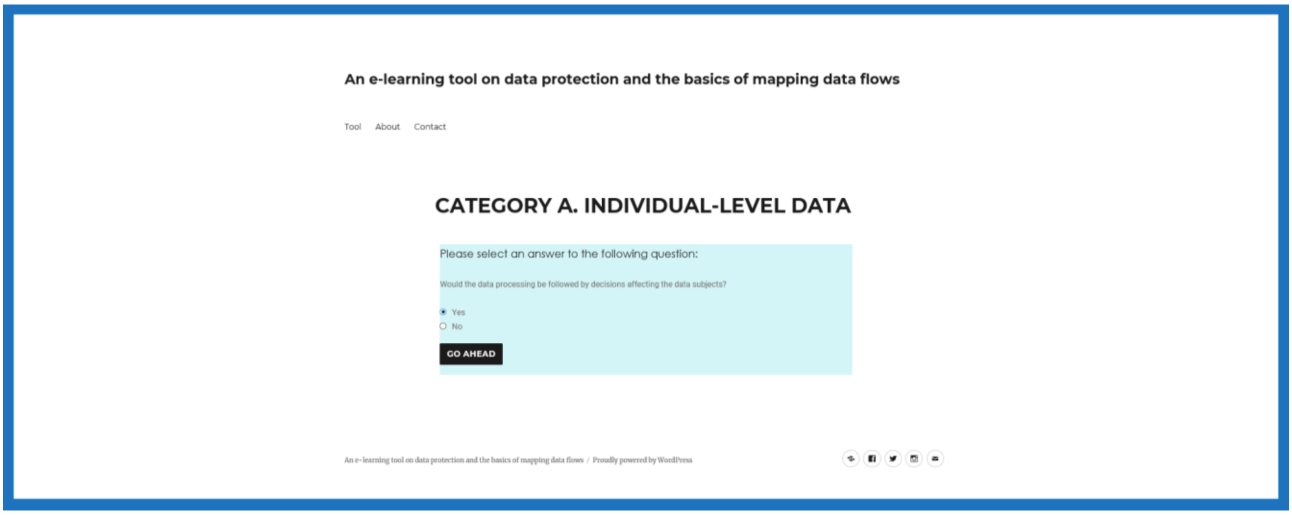
STAGE 3: DATA CATEGORIES

Please select an answer to the following question:

Is the current (version of the) particular dataset you intend to process: individual-level data, aggregate-level data or apersonal data?

● Category A. Individual-level data. The data units are people. (E.g. data relate to a particular customer, patient or survey participant.)
○ Category B. Aggregate-level data. The data units are about groups of people. (E.g. statistics about customer preferences.)
○ Category C. Apersonal data. The data units are non-people entities, such as events or processes. (E.g. data that concern the monitoring of manufacturing and transport processes.)

GO AHEAD
An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress



An e-learning tool on data protection and the basics of mapping data flows

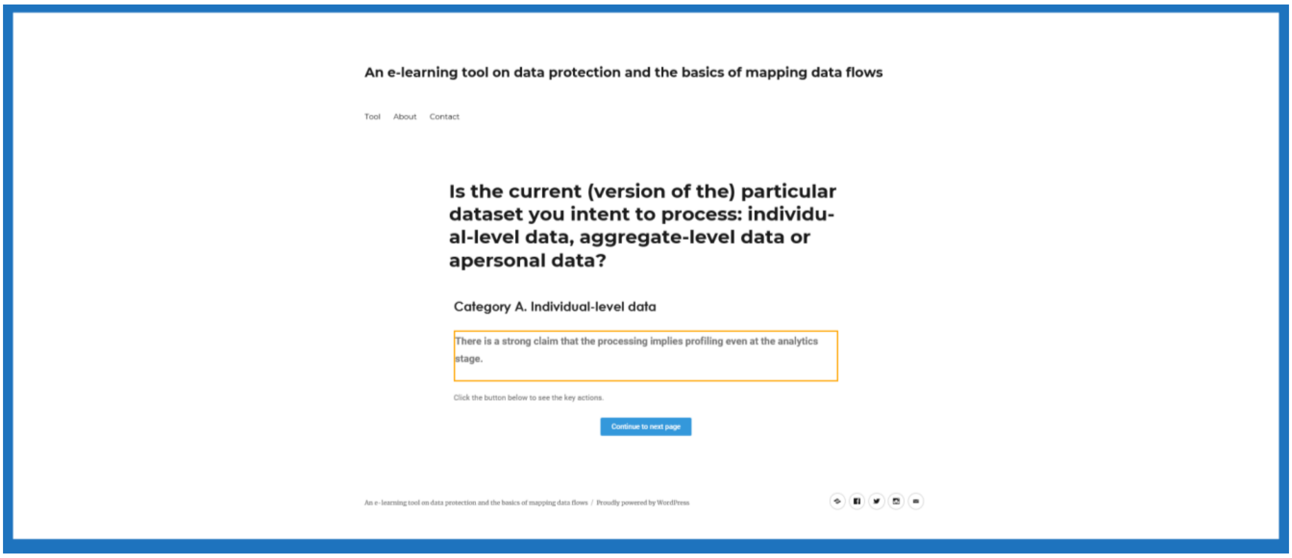
CATEGORY A. INDIVIDUAL-LEVEL DATA

Please select an answer to the following question:

Would the data processing be followed by decisions affecting the data subjects?

● Yes
○ No

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress



An e-learning tool on data protection and the basics of mapping data flows

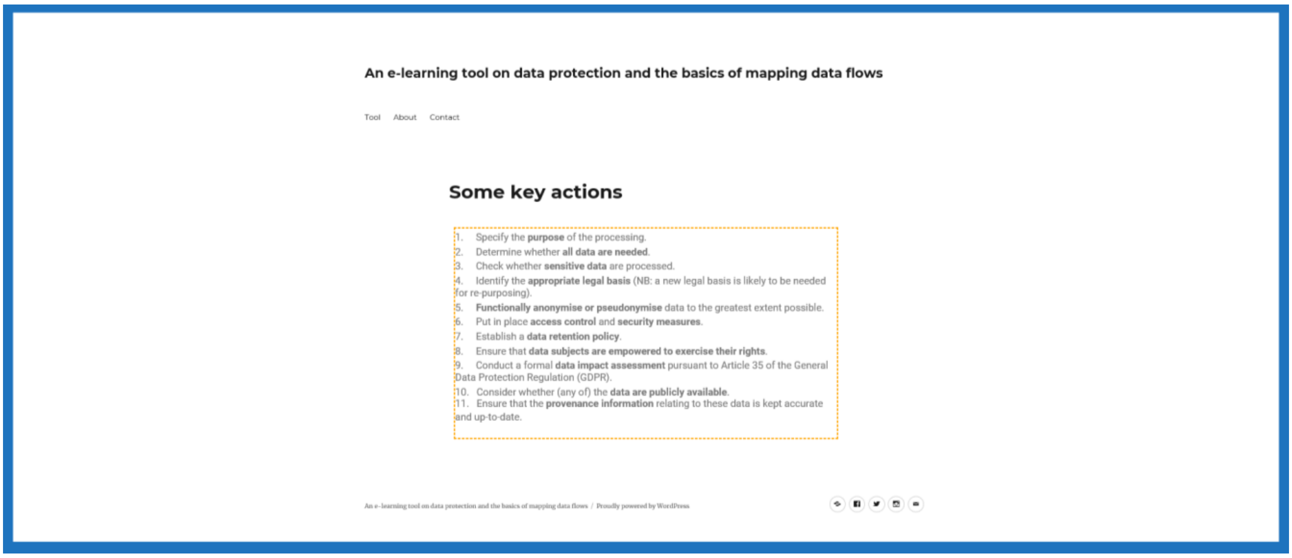
Is the current (version of the) particular dataset you intent to process: individual-level data, aggregate-level data or apersonal data?

Category A. Individual-level data

There is a strong claim that the processing implies profiling even at the analytics stage.

Click the button below to see the key actions.

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress



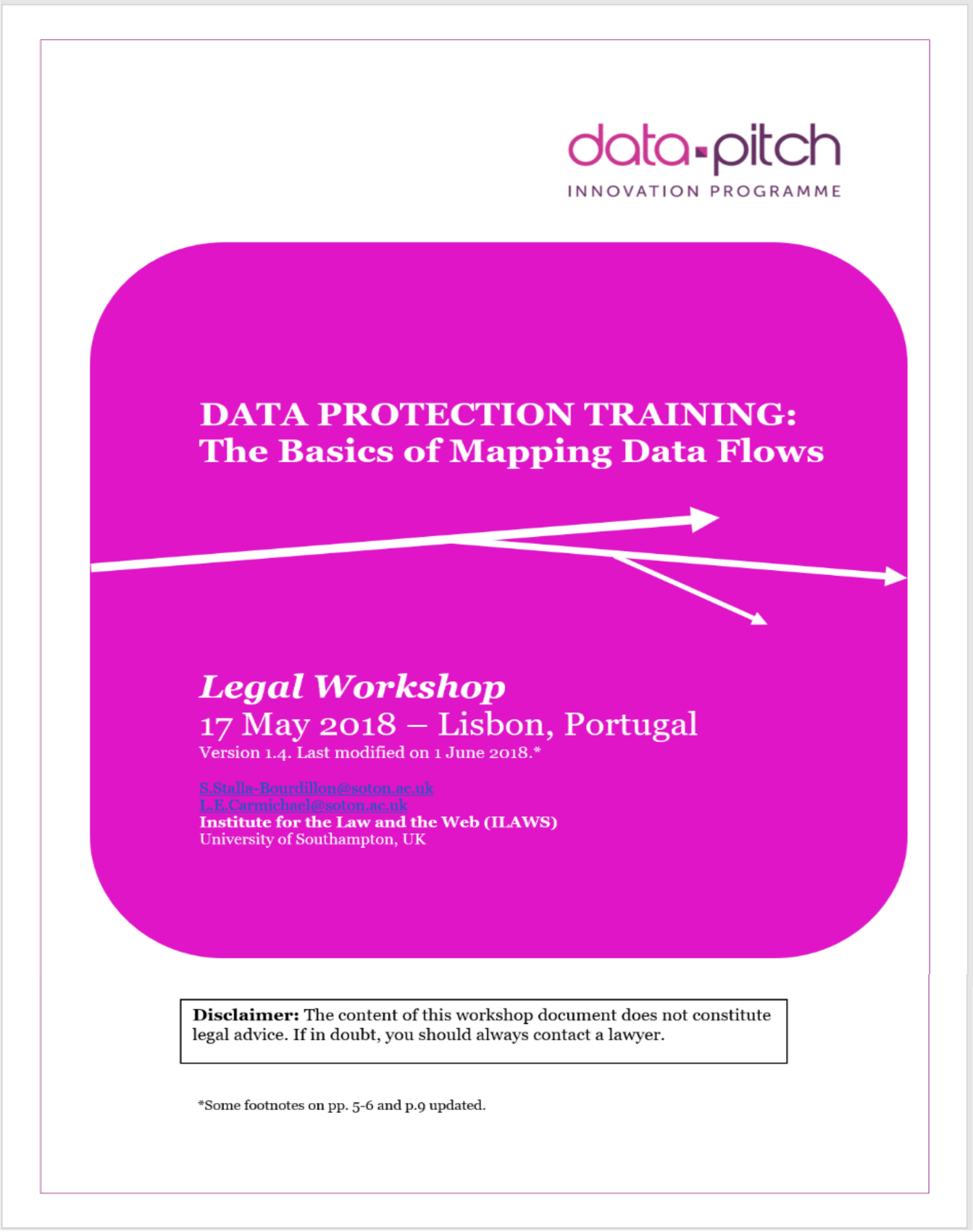
An e-learning tool on data protection and the basics of mapping data flows

Some key actions

1. Specify the **purpose** of the processing.
2. Determine whether **all data are needed**.
3. Check whether **sensitive data** are processed.
4. Identify the **appropriate legal basis** (NB: a new legal basis is likely to be needed for re-purposing).
5. **Functionally anonymise or pseudonymise** data to the greatest extent possible.
6. Put in place **access control** and **security measures**.
7. Establish a **data retention policy**.
8. Ensure that **data subjects are empowered to exercise their rights**.
9. Conduct a formal **data impact assessment** pursuant to Article 35 of the General Data Protection Regulation (GDPR).
10. Consider whether (any of) the **data are publicly available**.
11. Ensure that the **provenance information** relating to these data is kept accurate and up-to-date.

An e-learning tool on data protection and the basics of mapping data flows / Proudly powered by WordPress

Annex 2: Legal workshop document



data·pitch
INNOVATION PROGRAMME

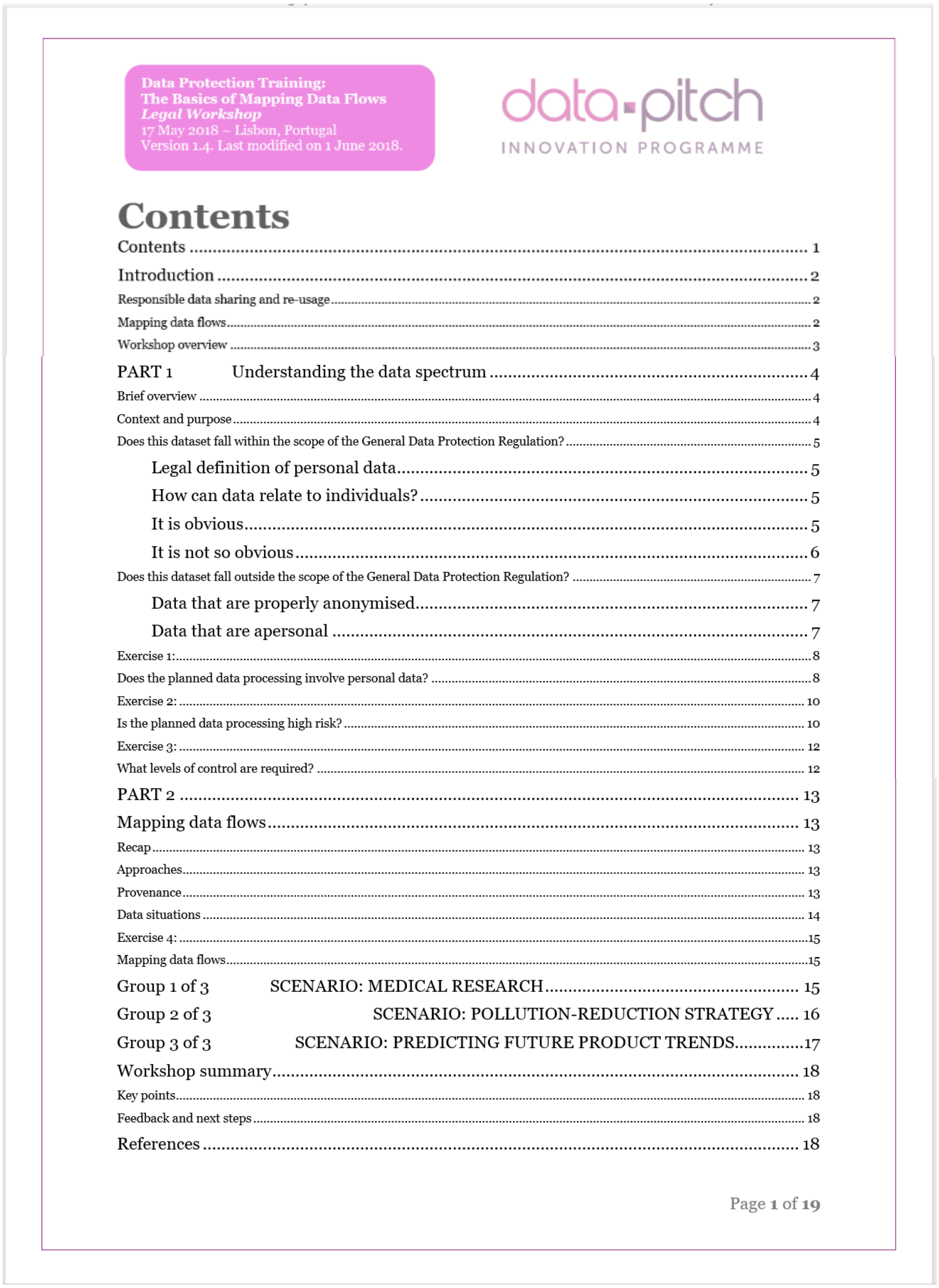
DATA PROTECTION TRAINING:
The Basics of Mapping Data Flows

Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.*

S.Stalla-Bourdillon@soton.ac.uk
L.E.Carmichael@soton.ac.uk
Institute for the Law and the Web (ILAWS)
University of Southampton, UK

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

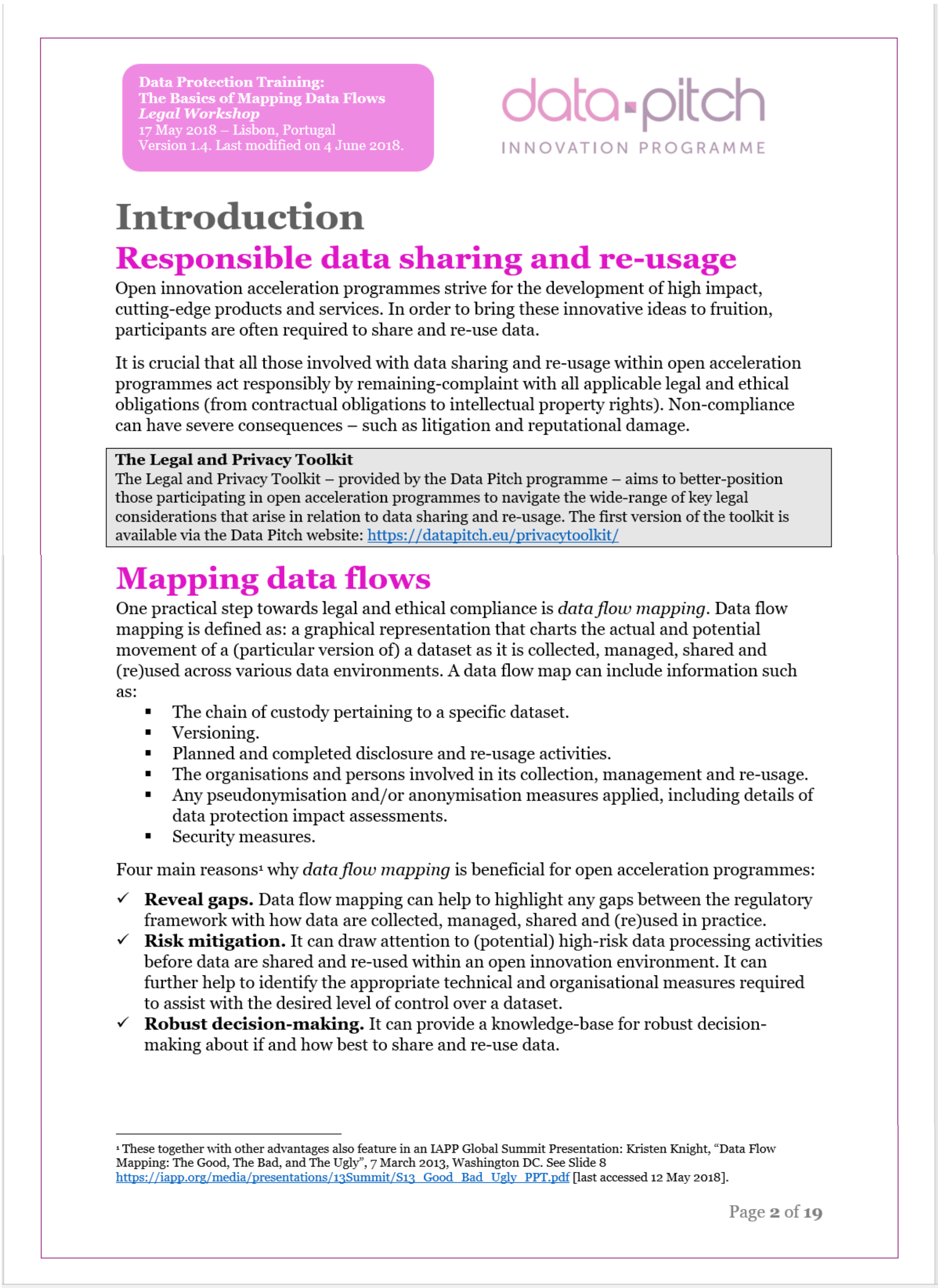
*Some footnotes on pp. 5-6 and p.9 updated.



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

Contents

Page 1 of 19



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 4 June 2018.

data-pitch
INNOVATION PROGRAMME

Introduction

Responsible data sharing and re-usage

Open innovation acceleration programmes strive for the development of high impact, cutting-edge products and services. In order to bring these innovative ideas to fruition, participants are often required to share and re-use data.

It is crucial that all those involved with data sharing and re-usage within open acceleration programmes act responsibly by remaining-complaint with all applicable legal and ethical obligations (from contractual obligations to intellectual property rights). Non-compliance can have severe consequences – such as litigation and reputational damage.

The Legal and Privacy Toolkit
The Legal and Privacy Toolkit – provided by the Data Pitch programme – aims to better-position those participating in open acceleration programmes to navigate the wide-range of key legal considerations that arise in relation to data sharing and re-usage. The first version of the toolkit is available via the Data Pitch website: https://datapitch.eu/privacytoolkit/

Mapping data flows

One practical step towards legal and ethical compliance is *data flow mapping*. Data flow mapping is defined as: a graphical representation that charts the actual and potential movement of a (particular version of) a dataset as it is collected, managed, shared and (re)used across various data environments. A data flow map can include information such as:

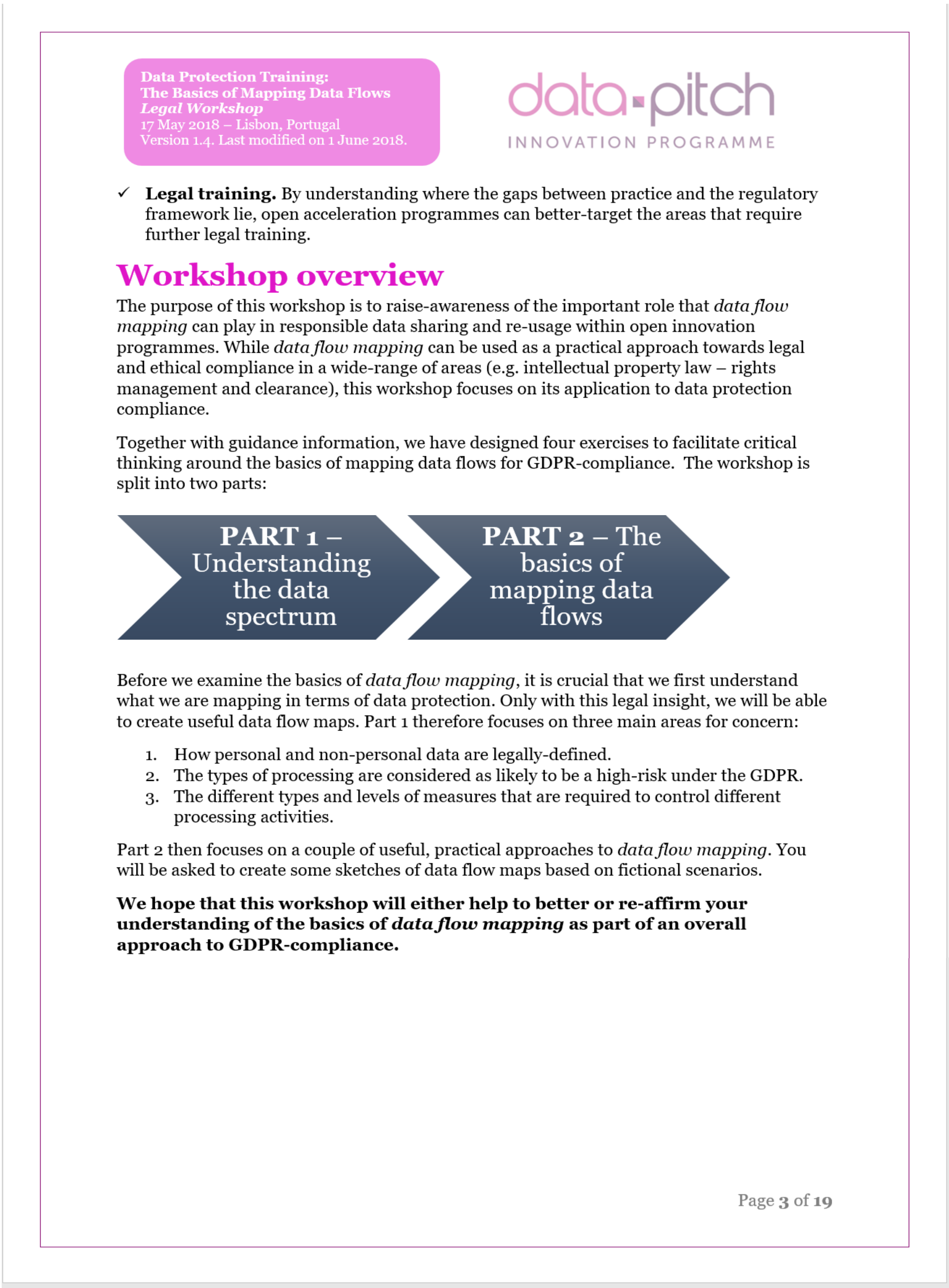
- The chain of custody pertaining to a specific dataset.
- Versioning.
- Planned and completed disclosure and re-usage activities.
- The organisations and persons involved in its collection, management and re-usage.
- Any pseudonymisation and/or anonymisation measures applied, including details of data protection impact assessments.
- Security measures.

Four main reasons[1] why *data flow mapping* is beneficial for open acceleration programmes:

✓ **Reveal gaps.** Data flow mapping can help to highlight any gaps between the regulatory framework with how data are collected, managed, shared and (re)used in practice.
✓ **Risk mitigation.** It can draw attention to (potential) high-risk data processing activities before data are shared and re-used within an open innovation environment. It can further help to identify the appropriate technical and organisational measures required to assist with the desired level of control over a dataset.
✓ **Robust decision-making.** It can provide a knowledge-base for robust decision-making about if and how best to share and re-use data.

[1] These together with other advantages also feature in an IAPP Global Summit Presentation: Kristen Knight, "Data Flow Mapping: The Good, The Bad, and The Ugly", 7 March 2013, Washington DC. See Slide 8 https://iapp.org/media/presentations/13Summit/S13_Good_Bad_Ugly_PPT.pdf [last accessed 12 May 2018].

Page **2** of **19**



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

data·pitch
INNOVATION PROGRAMME

✓ **Legal training.** By understanding where the gaps between practice and the regulatory framework lie, open acceleration programmes can better-target the areas that require further legal training.

Workshop overview

The purpose of this workshop is to raise-awareness of the important role that *data flow mapping* can play in responsible data sharing and re-usage within open innovation programmes. While *data flow mapping* can be used as a practical approach towards legal and ethical compliance in a wide-range of areas (e.g. intellectual property law – rights management and clearance), this workshop focuses on its application to data protection compliance.

Together with guidance information, we have designed four exercises to facilitate critical thinking around the basics of mapping data flows for GDPR-compliance. The workshop is split into two parts:

PART 1 – Understanding the data spectrum

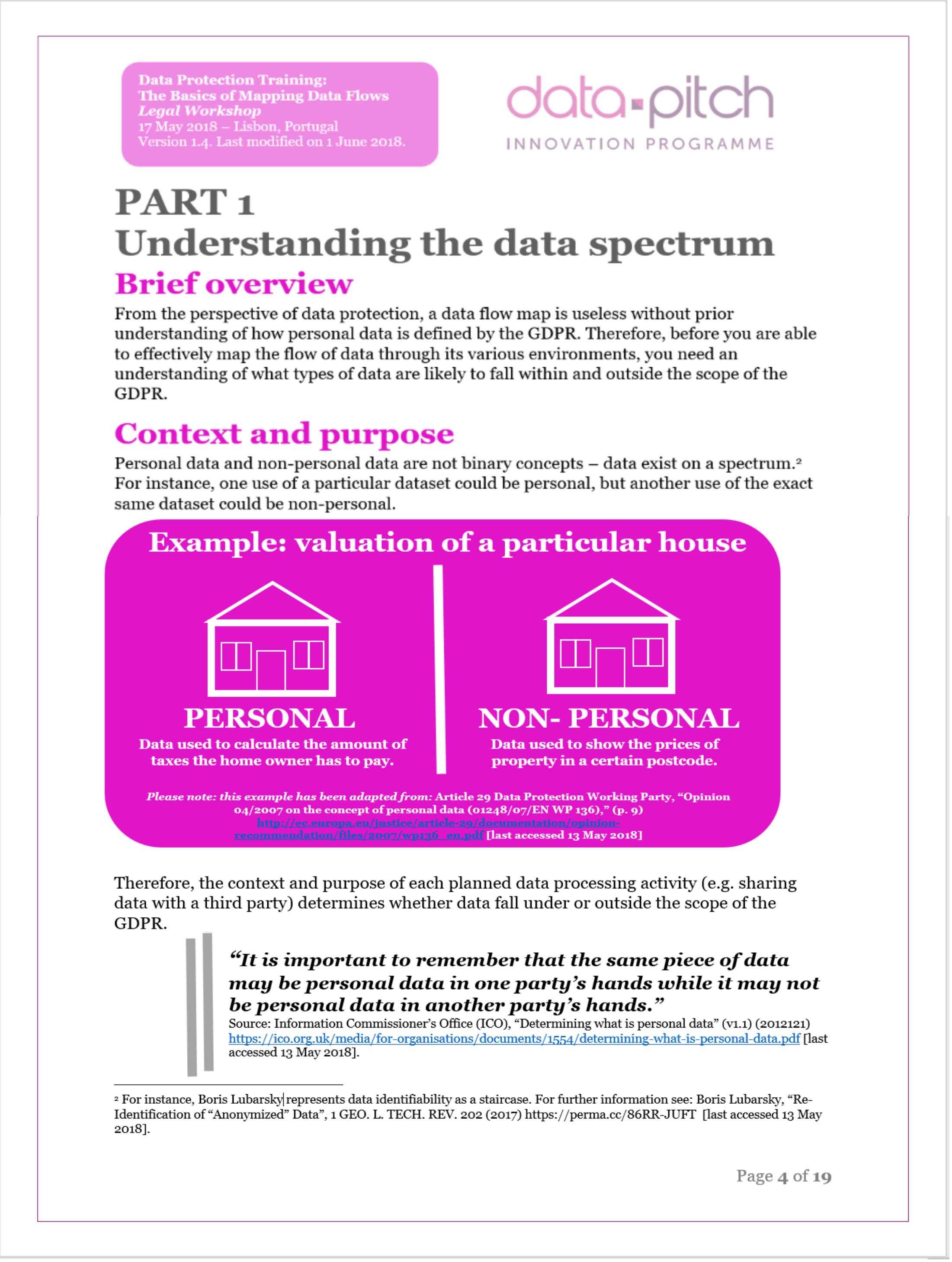
PART 2 – The basics of mapping data flows

Before we examine the basics of *data flow mapping*, it is crucial that we first understand what we are mapping in terms of data protection. Only with this legal insight, we will be able to create useful data flow maps. Part 1 therefore focuses on three main areas for concern:

1. How personal and non-personal data are legally-defined.
2. The types of processing are considered as likely to be a high-risk under the GDPR.
3. The different types and levels of measures that are required to control different processing activities.

Part 2 then focuses on a couple of useful, practical approaches to *data flow mapping*. You will be asked to create some sketches of data flow maps based on fictional scenarios.

We hope that this workshop will either help to better or re-affirm your understanding of the basics of *data flow mapping* as part of an overall approach to GDPR-compliance.



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

data·pitch
INNOVATION PROGRAMME

PART 1
Understanding the data spectrum

Brief overview

From the perspective of data protection, a data flow map is useless without prior understanding of how personal data is defined by the GDPR. Therefore, before you are able to effectively map the flow of data through its various environments, you need an understanding of what types of data are likely to fall within and outside the scope of the GDPR.

Context and purpose

Personal data and non-personal data are not binary concepts – data exist on a spectrum.[2] For instance, one use of a particular dataset could be personal, but another use of the exact same dataset could be non-personal.

Example: valuation of a particular house

PERSONAL
Data used to calculate the amount of taxes the home owner has to pay.

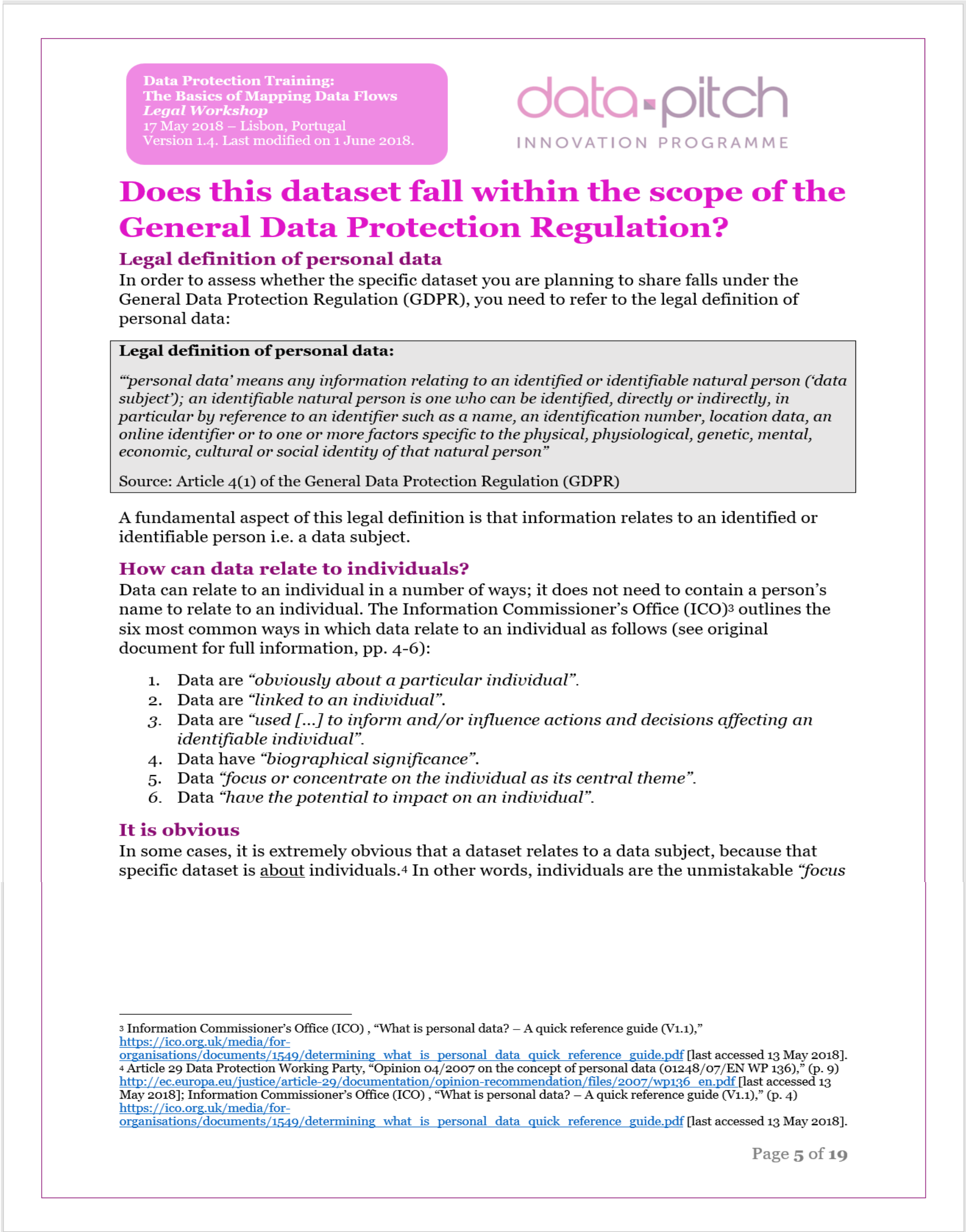
NON- PERSONAL
Data used to show the prices of property in a certain postcode.

Please note: this example has been adapted from: **Article 29 Data Protection Working Party, “Opinion 04/2007 on the concept of personal data (01248/07/EN WP 136),” (p. 9) http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf [last accessed 13 May 2018]**

Therefore, the context and purpose of each planned data processing activity (e.g. sharing data with a third party) determines whether data fall under or outside the scope of the GDPR.

> ***“It is important to remember that the same piece of data may be personal data in one party’s hands while it may not be personal data in another party’s hands.”***
> Source: Information Commissioner’s Office (ICO), “Determining what is personal data” (v1.1) (2012121) https://ico.org.uk/media/for-organisations/documents/1554/determining-what-is-personal-data.pdf [last accessed 13 May 2018].

[2] For instance, Boris Lubarsky represents data identifiability as a staircase. For further information see: Boris Lubarsky, “Re-Identification of “Anonymized” Data”, 1 GEO. L. TECH. REV. 202 (2017) https://perma.cc/86RR-JUFT [last accessed 13 May 2018].



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

data·pitch
INNOVATION PROGRAMME

Does this dataset fall within the scope of the General Data Protection Regulation?

Legal definition of personal data

In order to assess whether the specific dataset you are planning to share falls under the General Data Protection Regulation (GDPR), you need to refer to the legal definition of personal data:

> **Legal definition of personal data:**
>
> *"personal data' means 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"*
>
> Source: Article 4(1) of the General Data Protection Regulation (GDPR)

A fundamental aspect of this legal definition is that information relates to an identified or identifiable person i.e. a data subject.

How can data relate to individuals?

Data can relate to an individual in a number of ways; it does not need to contain a person's name to relate to an individual. The Information Commissioner's Office (ICO)[3] outlines the six most common ways in which data relate to an individual as follows (see original document for full information, pp. 4-6):

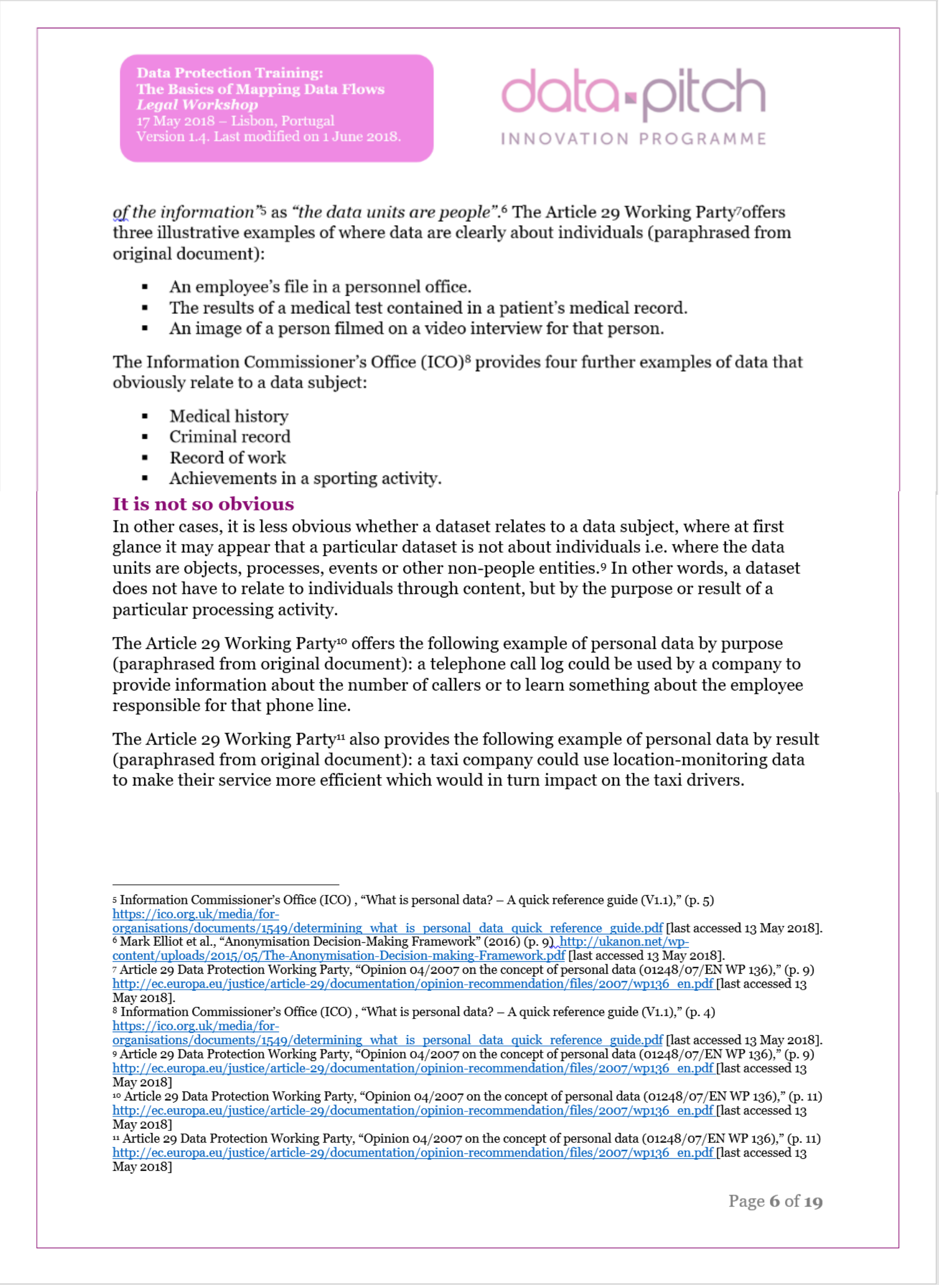
1. Data are *"obviously about a particular individual"*.
2. Data are *"linked to an individual"*.
3. Data are *"used [...] to inform and/or influence actions and decisions affecting an identifiable individual"*.
4. Data have *"biographical significance"*.
5. Data *"focus or concentrate on the individual as its central theme"*.
6. Data *"have the potential to impact on an individual"*.

It is obvious

In some cases, it is extremely obvious that a dataset relates to a data subject, because that specific dataset is about individuals.[4] In other words, individuals are the unmistakable *"focus*

[3] Information Commissioner's Office (ICO) , "What is personal data? – A quick reference guide (V1.1)," https://ico.org.uk/media/for-organisations/documents/1549/determining_what_is_personal_data_quick_reference_guide.pdf [last accessed 13 May 2018].

[4] Article 29 Data Protection Working Party, "Opinion 04/2007 on the concept of personal data (01248/07/EN WP 136)," (p. 9) http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf [last accessed 13 May 2018]; Information Commissioner's Office (ICO) , "What is personal data? – A quick reference guide (V1.1)," (p. 4) https://ico.org.uk/media/for-organisations/documents/1549/determining_what_is_personal_data_quick_reference_guide.pdf [last accessed 13 May 2018].



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

data·pitch
INNOVATION PROGRAMME

of the information"[5] as *"the data units are people"*.[6] The Article 29 Working Party[7]offers three illustrative examples of where data are clearly about individuals (paraphrased from original document):

- An employee's file in a personnel office.
- The results of a medical test contained in a patient's medical record.
- An image of a person filmed on a video interview for that person.

The Information Commissioner's Office (ICO)[8] provides four further examples of data that obviously relate to a data subject:

- Medical history
- Criminal record
- Record of work
- Achievements in a sporting activity.

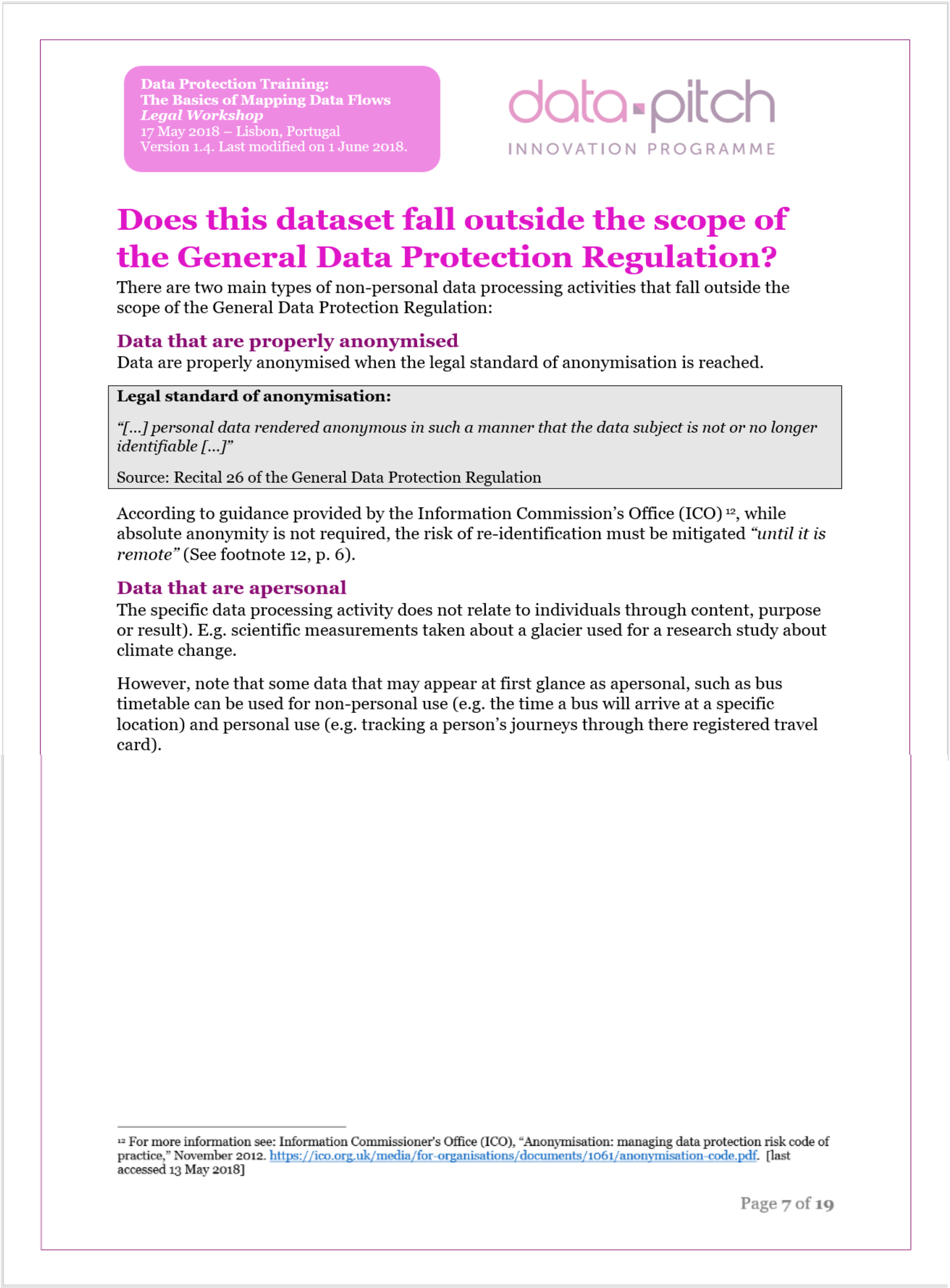
It is not so obvious

In other cases, it is less obvious whether a dataset relates to a data subject, where at first glance it may appear that a particular dataset is not about individuals i.e. where the data units are objects, processes, events or other non-people entities.[9] In other words, a dataset does not have to relate to individuals through content, but by the purpose or result of a particular processing activity.

The Article 29 Working Party[10] offers the following example of personal data by purpose (paraphrased from original document): a telephone call log could be used by a company to provide information about the number of callers or to learn something about the employee responsible for that phone line.

The Article 29 Working Party[11] also provides the following example of personal data by result (paraphrased from original document): a taxi company could use location-monitoring data to make their service more efficient which would in turn impact on the taxi drivers.

[5] Information Commissioner's Office (ICO) , "What is personal data? – A quick reference guide (V1.1)," (p. 5) https://ico.org.uk/media/for-organisations/documents/1549/determining_what_is_personal_data_quick_reference_guide.pdf [last accessed 13 May 2018].
[6] Mark Elliot et al., "Anonymisation Decision-Making Framework" (2016) (p. 9) http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf [last accessed 13 May 2018].
[7] Article 29 Data Protection Working Party, "Opinion 04/2007 on the concept of personal data (01248/07/EN WP 136)," (p. 9) http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf [last accessed 13 May 2018].
[8] Information Commissioner's Office (ICO) , "What is personal data? – A quick reference guide (V1.1)," (p. 4) https://ico.org.uk/media/for-organisations/documents/1549/determining_what_is_personal_data_quick_reference_guide.pdf [last accessed 13 May 2018].
[9] Article 29 Data Protection Working Party, "Opinion 04/2007 on the concept of personal data (01248/07/EN WP 136)," (p. 9) http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf [last accessed 13 May 2018]
[10] Article 29 Data Protection Working Party, "Opinion 04/2007 on the concept of personal data (01248/07/EN WP 136)," (p. 11) http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf [last accessed 13 May 2018]
[11] Article 29 Data Protection Working Party, "Opinion 04/2007 on the concept of personal data (01248/07/EN WP 136)," (p. 11) http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf [last accessed 13 May 2018]



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

data·pitch
INNOVATION PROGRAMME

Does this dataset fall outside the scope of the General Data Protection Regulation?

There are two main types of non-personal data processing activities that fall outside the scope of the General Data Protection Regulation:

Data that are properly anonymised

Data are properly anonymised when the legal standard of anonymisation is reached.

> **Legal standard of anonymisation:**
>
> *"[…] personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable […]"*
>
> Source: Recital 26 of the General Data Protection Regulation

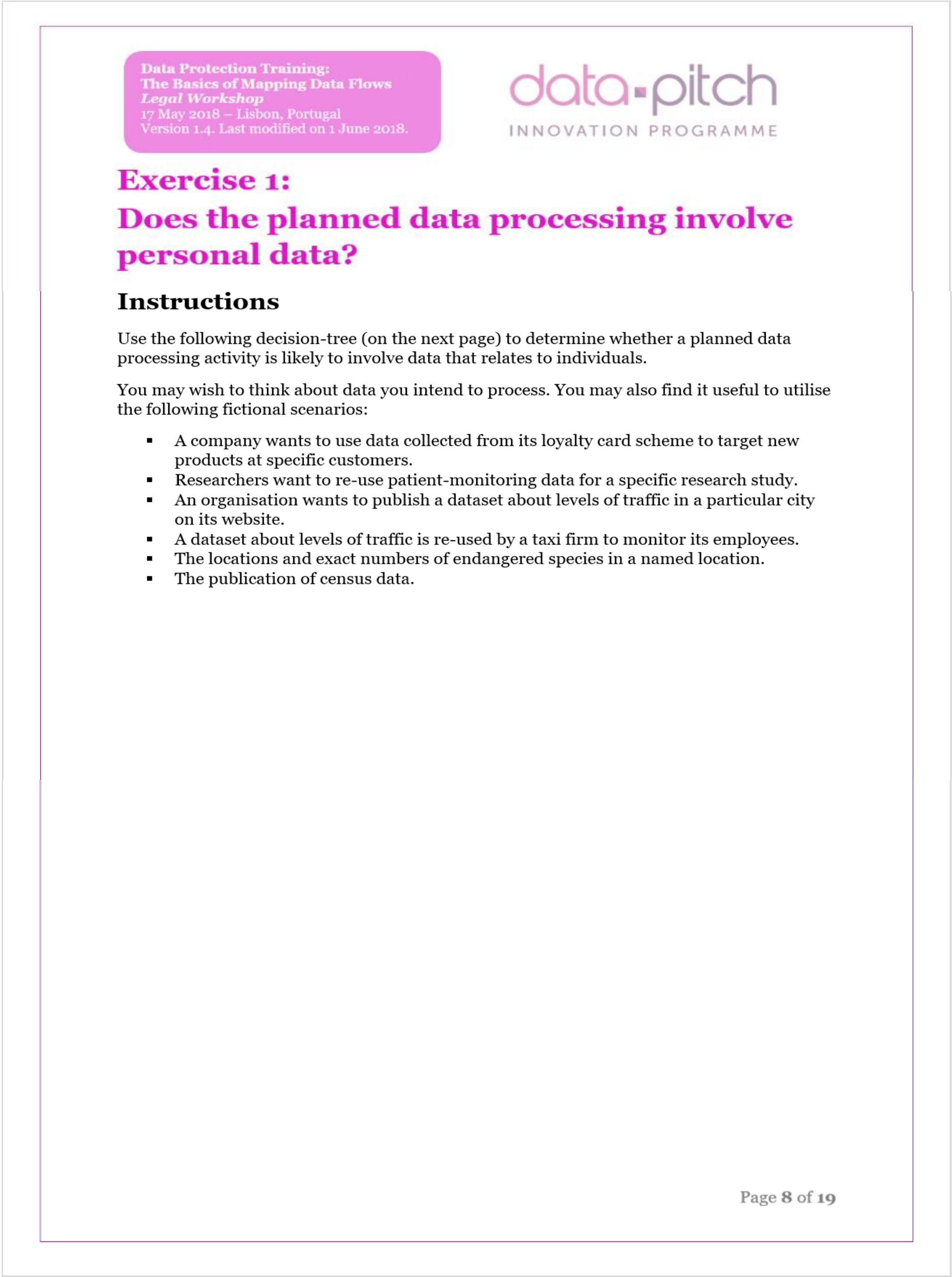
According to guidance provided by the Information Commissioner's Office (ICO) [12], while absolute anonymity is not required, the risk of re-identification must be mitigated *"until it is remote"* (See footnote 12, p. 6).

Data that are apersonal

The specific data processing activity does not relate to individuals through content, purpose or result). E.g. scientific measurements taken about a glacier used for a research study about climate change.

However, note that some data that may appear at first glance as apersonal, such as bus timetable can be used for non-personal use (e.g. the time a bus will arrive at a specific location) and personal use (e.g. tracking a person's journeys through there registered travel card).

[12] For more information see: Information Commissioner's Office (ICO), "Anonymisation: managing data protection risk code of practice," November 2012. https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf. [last accessed 13 May 2018]



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

data·pitch
INNOVATION PROGRAMME

Exercise 1:
Does the planned data processing involve personal data?

Instructions

Use the following decision-tree (on the next page) to determine whether a planned data processing activity is likely to involve data that relates to individuals.

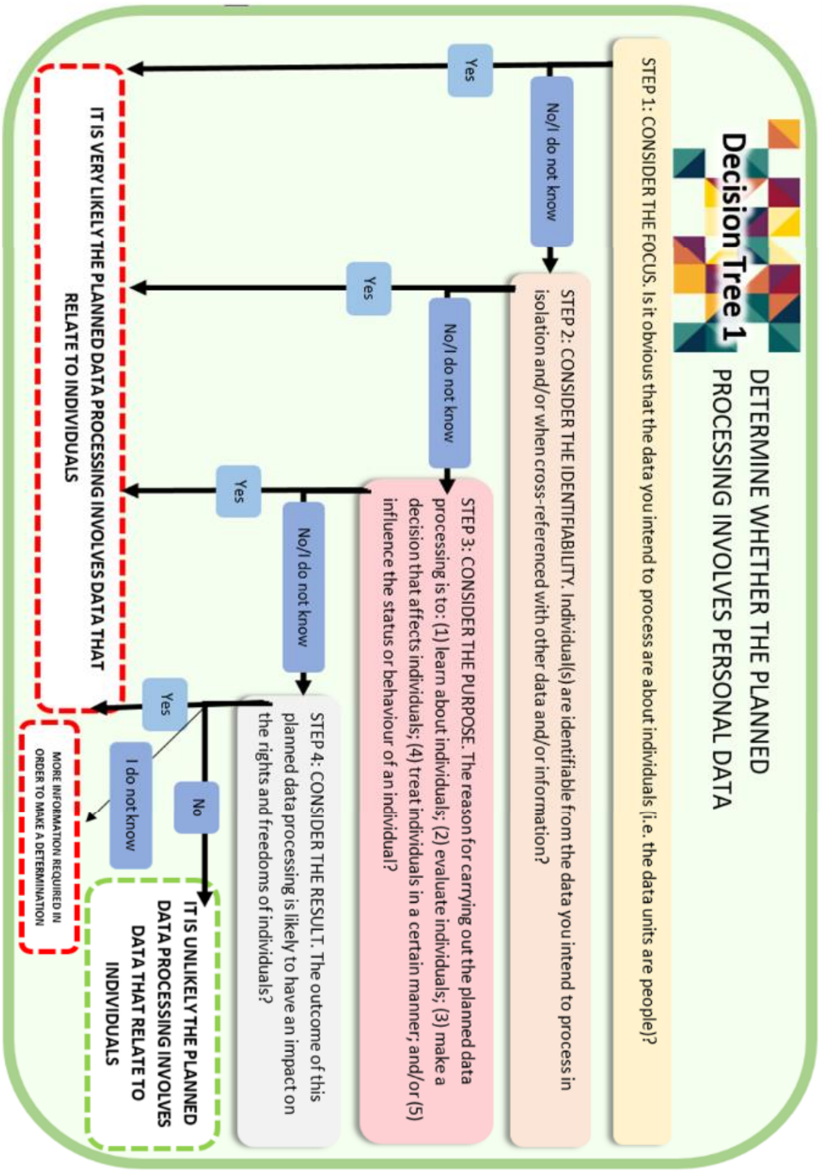
You may wish to think about data you intend to process. You may also find it useful to utilise the following fictional scenarios:

- A company wants to use data collected from its loyalty card scheme to target new products at specific customers.
- Researchers want to re-use patient-monitoring data for a specific research study.
- An organisation wants to publish a dataset about levels of traffic in a particular city on its website.
- A dataset about levels of traffic is re-used by a taxi firm to monitor its employees.
- The locations and exact numbers of endangered species in a named location.
- The publication of census data.

Page **8** of **19**

Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

data·pitch
INNOVATION PROGRAMME



Decision Tree 1

DETERMINE WHETHER THE PLANNED PROCESSING INVOLVES PERSONAL DATA

STEP 1: CONSIDER THE FOCUS. Is it obvious that the data you intend to process are about individuals (i.e. the data units are people)?

Yes

No/I do not know

STEP 2: CONSIDER THE IDENTIFIABILITY. Individual(s) are identifiable from the data you intend to process in isolation and/or when cross-referenced with other data and/or information?

Yes

No/I do not know

STEP 3: CONSIDER THE PURPOSE. The reason for carrying out the planned data processing is to: (1) learn about individuals; (2) evaluate individuals; (3) make a decision that affects individuals; (4) treat individuals in a certain manner; and/or (5) influence the status or behaviour of an individual?

Yes

No/I do not know

STEP 4: CONSIDER THE RESULT. The outcome of this planned data processing is likely to have an impact on the rights and freedoms of individuals?

Yes

No

I do not know

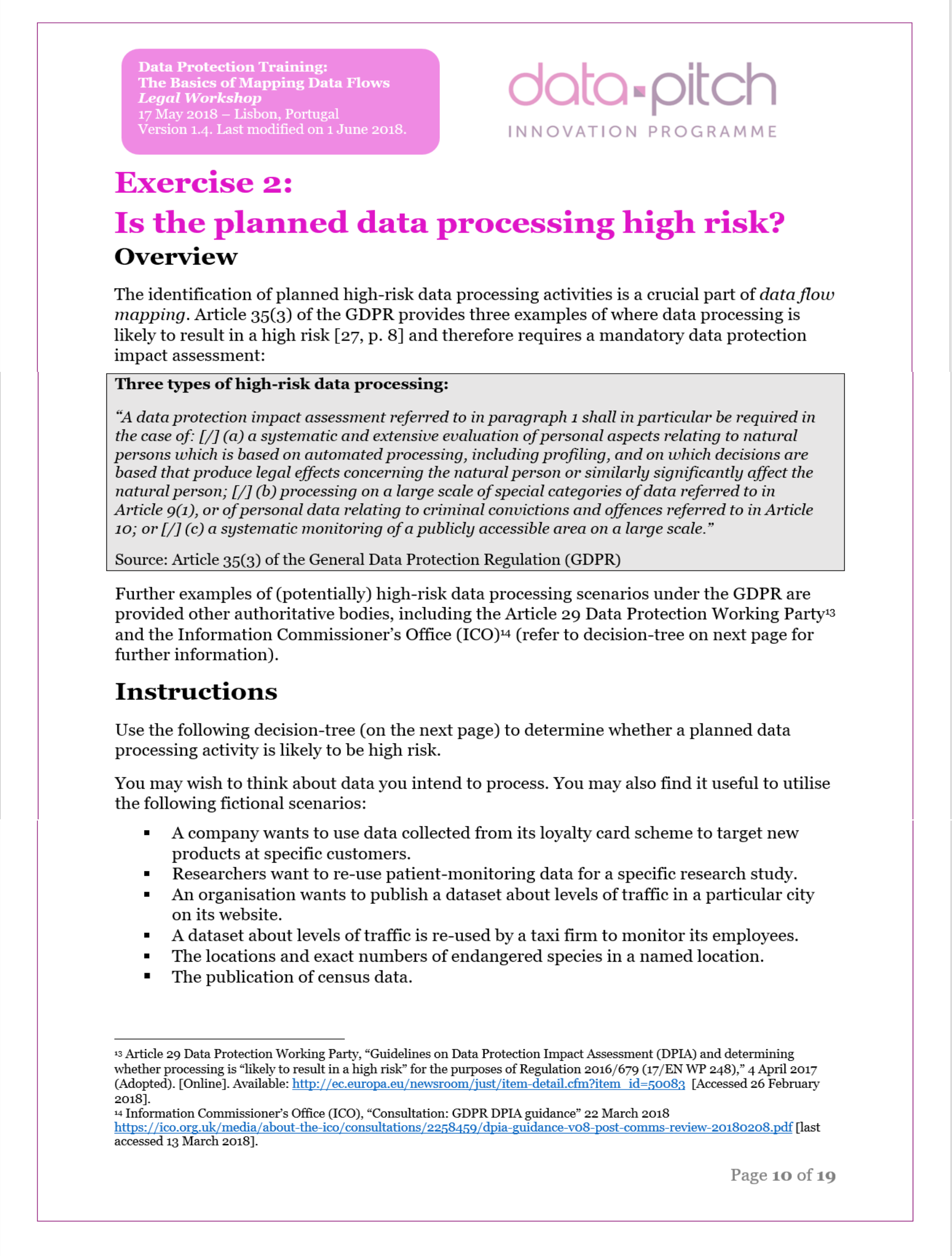
IT IS VERY LIKELY THE PLANNED DATA PROCESSING INVOLVES DATA THAT RELATE TO INDIVIDUALS

IT IS UNLIKELY THE PLANNED DATA PROCESSING INVOLVES DATA THAT RELATE TO INDIVIDUALS

MORE INFORMATION REQUIRED IN ORDER TO MAKE A DETERMINATION

NOTE – this decision-tree has been derived by the Toolkit authors from: (i) the Article 29 Data Protection Working Party Opinion (04/2007) on the concept of personal data (p. 10-11) Steps 1, 3 and 4; (ii) the UKAN Anonymisation Decision-Making Framework (p. 8) – Steps 1 and 2; (iii) Information Commissioner's Office (ICO), "What is personal data? – A quick reference guide (V1.1; and, (iv) S. Wood, "ICO blog: Anonymisation – opportunities and risks," Information Commissioner's Office (ICO) Blog, 16 November 2012. [Online]. Available: https://iconewsblog.org.uk/2012/11/16/ico-blog-anonymisation-opportunities-and-risks/ [last accessed 16 March 2018] – step 2.

Page **9** of **19**



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

data·pitch
INNOVATION PROGRAMME

Exercise 2:
Is the planned data processing high risk?

Overview

The identification of planned high-risk data processing activities is a crucial part of *data flow mapping*. Article 35(3) of the GDPR provides three examples of where data processing is likely to result in a high risk [27, p. 8] and therefore requires a mandatory data protection impact assessment:

> **Three types of high-risk data processing:**
>
> *"A data protection impact assessment referred to in paragraph 1 shall in particular be required in the case of: [/] (a) 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; [/] (b) processing on a large scale of special categories of data referred to in Article 9(1), or of personal data relating to criminal convictions and offences referred to in Article 10; or [/] (c) a systematic monitoring of a publicly accessible area on a large scale."*
>
> Source: Article 35(3) of the General Data Protection Regulation (GDPR)

Further examples of (potentially) high-risk data processing scenarios under the GDPR are provided other authoritative bodies, including the Article 29 Data Protection Working Party[13] and the Information Commissioner's Office (ICO)[14] (refer to decision-tree on next page for further information).

Instructions

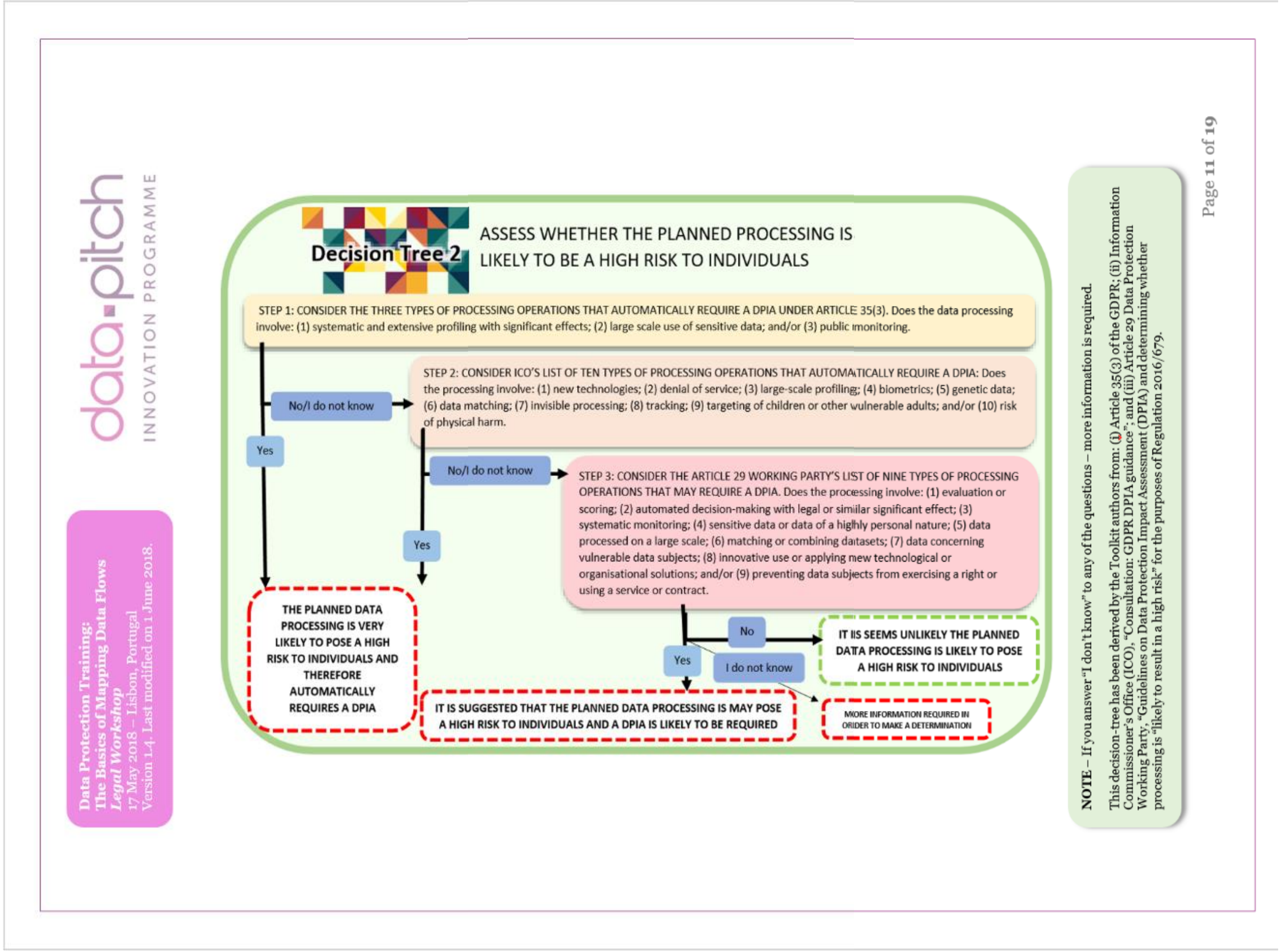
Use the following decision-tree (on the next page) to determine whether a planned data processing activity is likely to be high risk.

You may wish to think about data you intend to process. You may also find it useful to utilise the following fictional scenarios:

- A company wants to use data collected from its loyalty card scheme to target new products at specific customers.
- Researchers want to re-use patient-monitoring data for a specific research study.
- An organisation wants to publish a dataset about levels of traffic in a particular city on its website.
- A dataset about levels of traffic is re-used by a taxi firm to monitor its employees.
- The locations and exact numbers of endangered species in a named location.
- The publication of census data.

[13] Article 29 Data Protection Working Party, "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 (17/EN WP 248)," 4 April 2017 (Adopted). [Online]. Available: http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=50083 [Accessed 26 February 2018].

[14] Information Commissioner's Office (ICO), "Consultation: GDPR DPIA guidance" 22 March 2018 https://ico.org.uk/media/about-the-ico/consultations/2258459/dpia-guidance-v08-post-comms-review-20180208.pdf [last accessed 13 March 2018].



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

Decision Tree 2

ASSESS WHETHER THE PLANNED PROCESSING IS LIKELY TO BE A HIGH RISK TO INDIVIDUALS

STEP 1: CONSIDER THE THREE TYPES OF PROCESSING OPERATIONS THAT AUTOMATICALLY REQUIRE A DPIA UNDER ARTICLE 35(3). Does the data processing involve: (1) systematic and extensive profiling with significant effects; (2) large scale use of sensitive data; and/or (3) public monitoring.

STEP 2: CONSIDER ICO'S LIST OF TEN TYPES OF PROCESSING OPERATIONS THAT AUTOMATICALLY REQUIRE A DPIA: Does the processing involve: (1) new technologies; (2) denial of service; (3) large-scale profiling; (4) biometrics; (5) genetic data; (6) data matching; (7) invisible processing; (8) tracking; (9) targeting of children or other vulnerable adults; and/or (10) risk of physical harm.

STEP 3: CONSIDER THE ARTICLE 29 WORKING PARTY'S LIST OF NINE TYPES OF PROCESSING OPERATIONS THAT MAY REQUIRE A DPIA. Does the processing involve: (1) evaluation or scoring; (2) automated decision-making with legal or similar significant effect; (3) systematic monitoring; (4) sensitive data or data of a highly personal nature; (5) data processed on a large scale; (6) matching or combining datasets; (7) data concerning vulnerable data subjects; (8) innovative use or applying new technological or organisational solutions; and/or (9) preventing data subjects from exercising a right or using a service or contract.

THE PLANNED DATA PROCESSING IS VERY LIKELY TO POSE A HIGH RISK TO INDIVIDUALS AND THEREFORE AUTOMATICALLY REQUIRES A DPIA

IT IS SUGGESTED THAT THE PLANNED DATA PROCESSING IS MAY POSE A HIGH RISK TO INDIVIDUALS AND A DPIA IS LIKELY TO BE REQUIRED

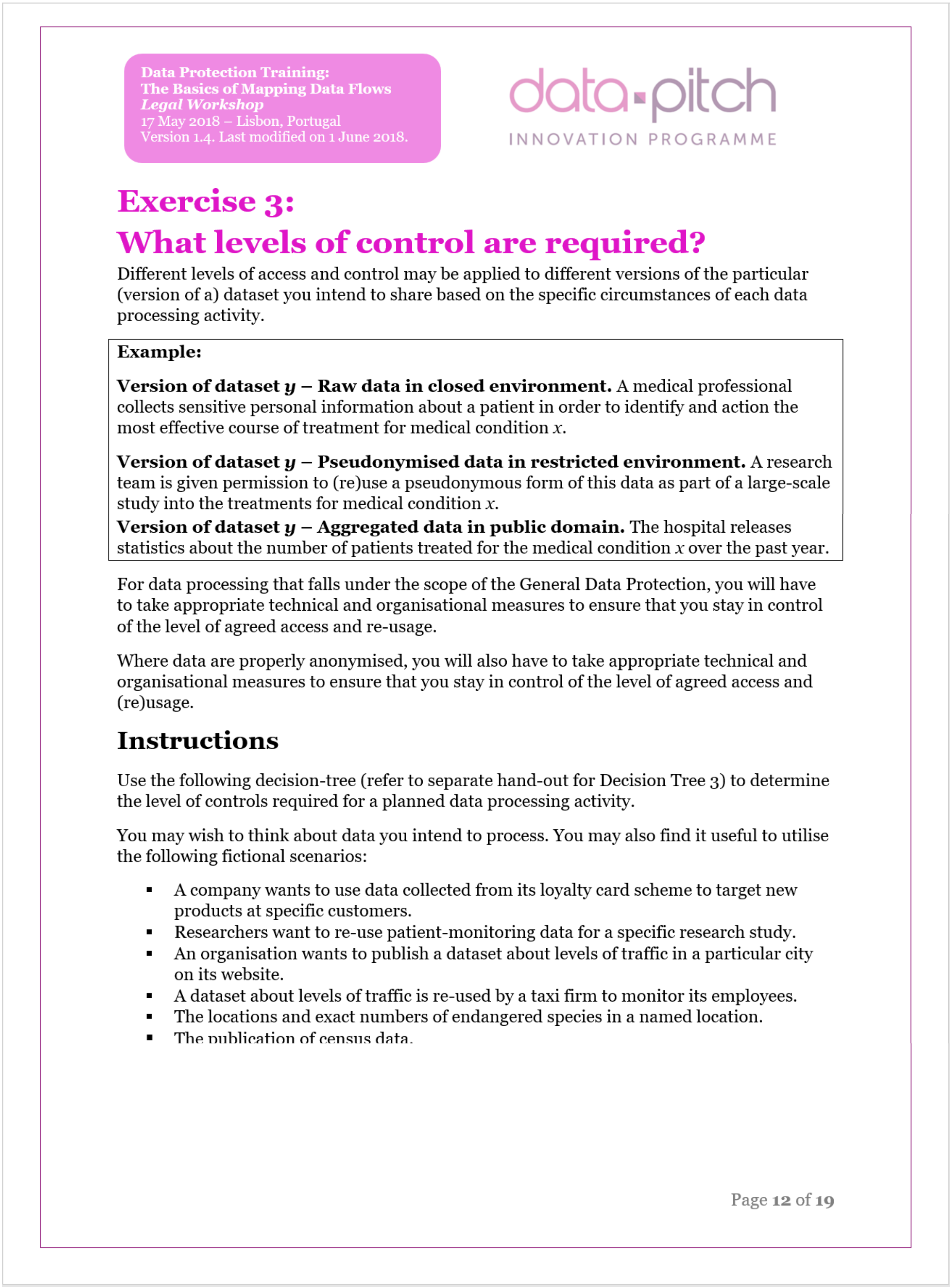
IT IS SEEMS UNLIKELY THE PLANNED DATA PROCESSING IS LIKELY TO POSE A HIGH RISK TO INDIVIDUALS

MORE INFORMATION REQUIRED IN ORDER TO MAKE A DETERMINATION

NOTE – If you answer "I don't know" to any of the questions – more information is required.

This decision-tree has been derived by the Toolkit authors from: (i) Article 35(3) of the GDPR; (ii) Information Commissioner's Office (ICO), "Consultation: GDPR DPIA guidance"; and (iii) Article 29 Data Protection Working Party, "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.

Page **11** of **19**



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

Exercise 3:
What levels of control are required?

Different levels of access and control may be applied to different versions of the particular (version of a) dataset you intend to share based on the specific circumstances of each data processing activity.

Example:

Version of dataset *y* – Raw data in closed environment. A medical professional collects sensitive personal information about a patient in order to identify and action the most effective course of treatment for medical condition *x*.

Version of dataset *y* – Pseudonymised data in restricted environment. A research team is given permission to (re)use a pseudonymous form of this data as part of a large-scale study into the treatments for medical condition *x*.

Version of dataset *y* – Aggregated data in public domain. The hospital releases statistics about the number of patients treated for the medical condition *x* over the past year.

For data processing that falls under the scope of the General Data Protection, you will have to take appropriate technical and organisational measures to ensure that you stay in control of the level of agreed access and re-usage.

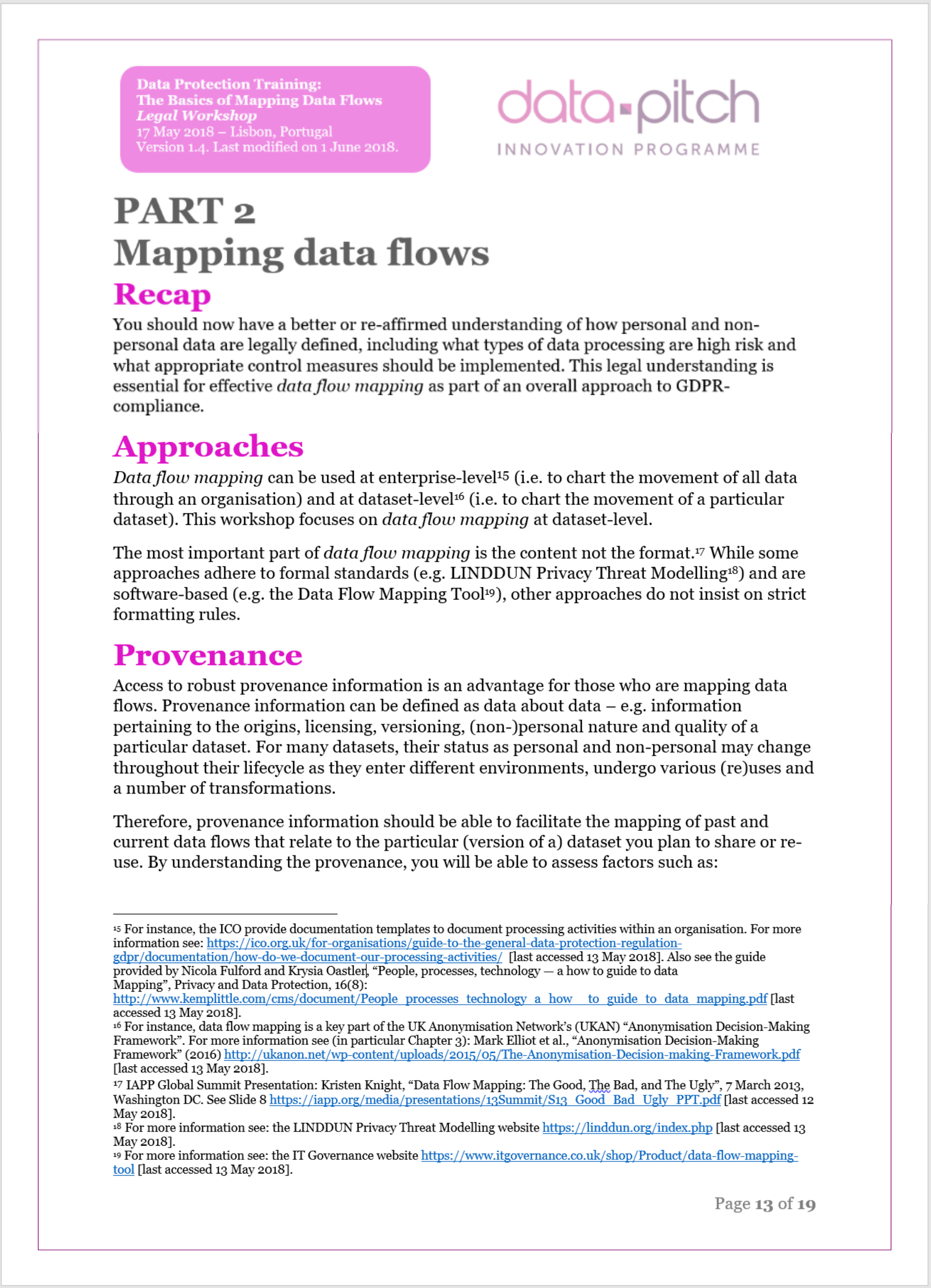
Where data are properly anonymised, you will also have to take appropriate technical and organisational measures to ensure that you stay in control of the level of agreed access and (re)usage.

Instructions

Use the following decision-tree (refer to separate hand-out for Decision Tree 3) to determine the level of controls required for a planned data processing activity.

You may wish to think about data you intend to process. You may also find it useful to utilise the following fictional scenarios:

- A company wants to use data collected from its loyalty card scheme to target new products at specific customers.
- Researchers want to re-use patient-monitoring data for a specific research study.
- An organisation wants to publish a dataset about levels of traffic in a particular city on its website.
- A dataset about levels of traffic is re-used by a taxi firm to monitor its employees.
- The locations and exact numbers of endangered species in a named location.
- The publication of census data.



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

data·pitch
INNOVATION PROGRAMME

PART 2
Mapping data flows

Recap

You should now have a better or re-affirmed understanding of how personal and non-personal data are legally defined, including what types of data processing are high risk and what appropriate control measures should be implemented. This legal understanding is essential for effective *data flow mapping* as part of an overall approach to GDPR-compliance.

Approaches

Data flow mapping can be used at enterprise-level[15] (i.e. to chart the movement of all data through an organisation) and at dataset-level[16] (i.e. to chart the movement of a particular dataset). This workshop focuses on *data flow mapping* at dataset-level.

The most important part of *data flow mapping* is the content not the format.[17] While some approaches adhere to formal standards (e.g. LINDDUN Privacy Threat Modelling[18]) and are software-based (e.g. the Data Flow Mapping Tool[19]), other approaches do not insist on strict formatting rules.

Provenance

Access to robust provenance information is an advantage for those who are mapping data flows. Provenance information can be defined as data about data – e.g. information pertaining to the origins, licensing, versioning, (non-)personal nature and quality of a particular dataset. For many datasets, their status as personal and non-personal may change throughout their lifecycle as they enter different environments, undergo various (re)uses and a number of transformations.

Therefore, provenance information should be able to facilitate the mapping of past and current data flows that relate to the particular (version of a) dataset you plan to share or re-use. By understanding the provenance, you will be able to assess factors such as:

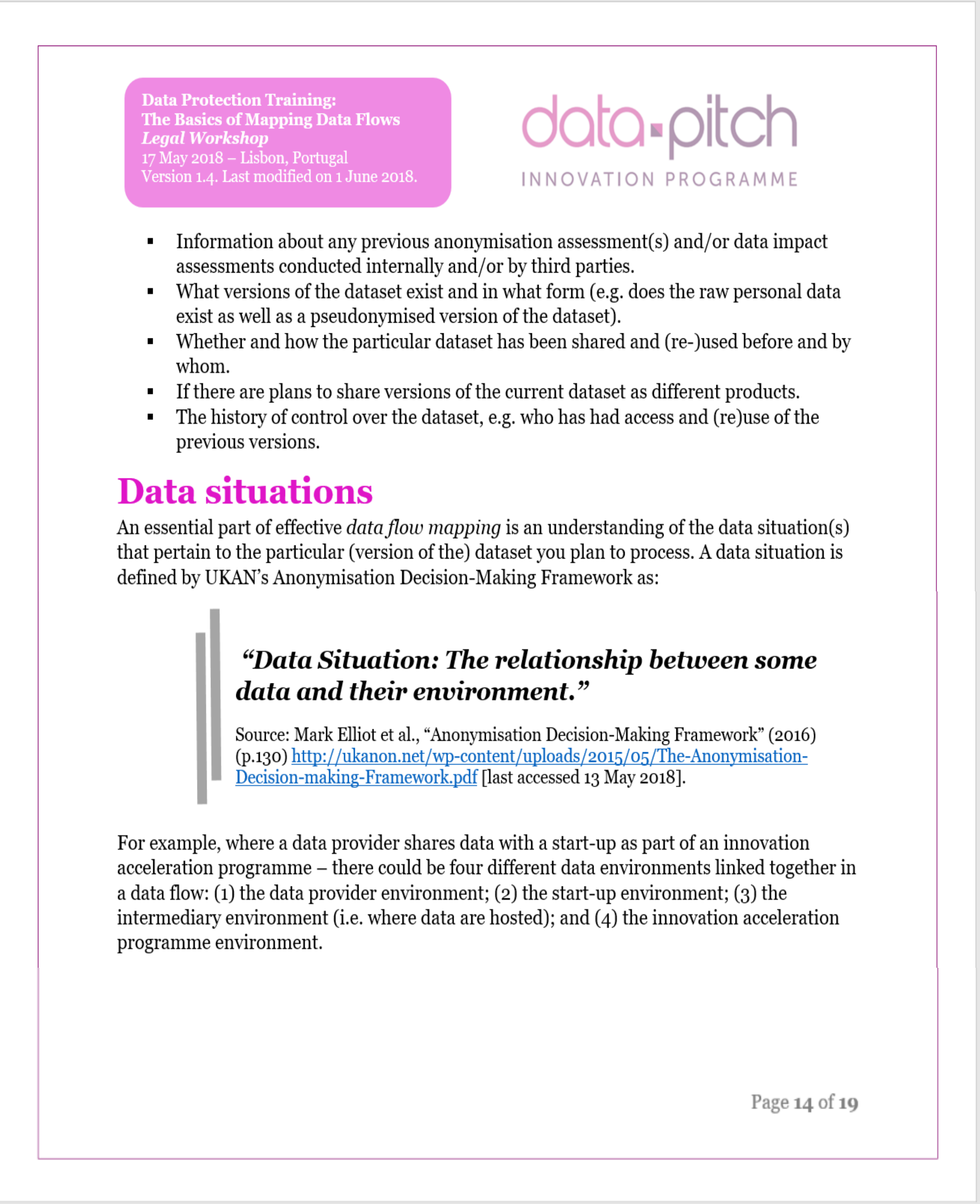
[15] For instance, the ICO provide documentation templates to document processing activities within an organisation. For more information see: https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/documentation/how-do-we-document-our-processing-activities/ [last accessed 13 May 2018]. Also see the guide provided by Nicola Fulford and Krysia Oastler, "People, processes, technology — a how to guide to data Mapping", Privacy and Data Protection, 16(8): http://www.kemplittle.com/cms/document/People_processes_technology_a_how__to_guide_to_data_mapping.pdf [last accessed 13 May 2018].

[16] For instance, data flow mapping is a key part of the UK Anonymisation Network's (UKAN) "Anonymisation Decision-Making Framework". For more information see (in particular Chapter 3): Mark Elliot et al., "Anonymisation Decision-Making Framework" (2016) http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf [last accessed 13 May 2018].

[17] IAPP Global Summit Presentation: Kristen Knight, "Data Flow Mapping: The Good, The Bad, and The Ugly", 7 March 2013, Washington DC. See Slide 8 https://iapp.org/media/presentations/13Summit/S13_Good_Bad_Ugly_PPT.pdf [last accessed 12 May 2018].

[18] For more information see: the LINDDUN Privacy Threat Modelling website https://linddun.org/index.php [last accessed 13 May 2018].

[19] For more information see: the IT Governance website https://www.itgovernance.co.uk/shop/Product/data-flow-mapping-tool [last accessed 13 May 2018].



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

data·pitch
INNOVATION PROGRAMME

- Information about any previous anonymisation assessment(s) and/or data impact assessments conducted internally and/or by third parties.
- What versions of the dataset exist and in what form (e.g. does the raw personal data exist as well as a pseudonymised version of the dataset).
- Whether and how the particular dataset has been shared and (re-)used before and by whom.
- If there are plans to share versions of the current dataset as different products.
- The history of control over the dataset, e.g. who has had access and (re)use of the previous versions.

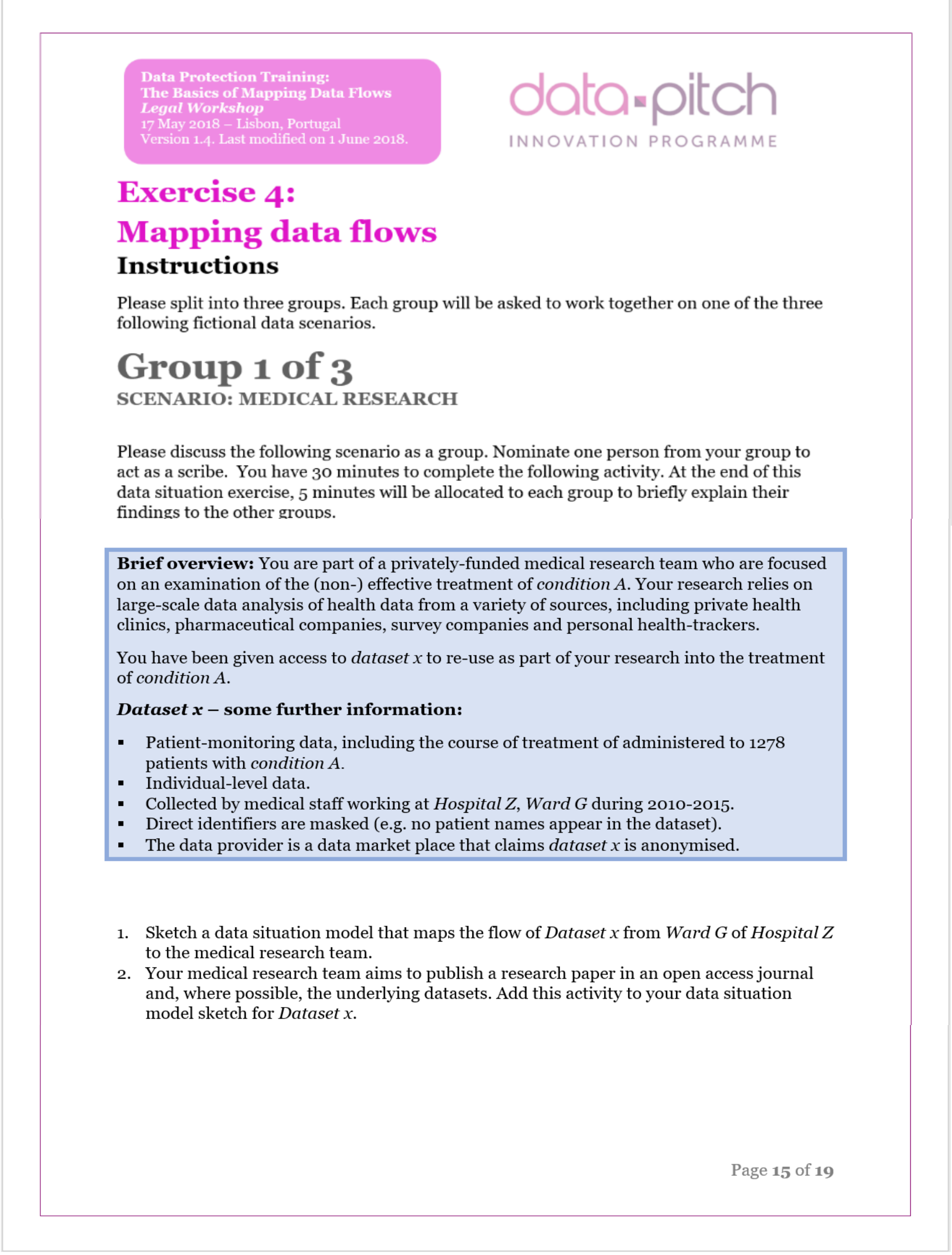
Data situations

An essential part of effective *data flow mapping* is an understanding of the data situation(s) that pertain to the particular (version of the) dataset you plan to process. A data situation is defined by UKAN's Anonymisation Decision-Making Framework as:

> ***"Data Situation: The relationship between some data and their environment."***
>
> Source: Mark Elliot et al., "Anonymisation Decision-Making Framework" (2016) (p.130) http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf [last accessed 13 May 2018].

For example, where a data provider shares data with a start-up as part of an innovation acceleration programme – there could be four different data environments linked together in a data flow: (1) the data provider environment; (2) the start-up environment; (3) the intermediary environment (i.e. where data are hosted); and (4) the innovation acceleration programme environment.

Page **14** of **19**



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

data·pitch
INNOVATION PROGRAMME

Exercise 4:
Mapping data flows

Instructions

Please split into three groups. Each group will be asked to work together on one of the three following fictional data scenarios.

Group 1 of 3

SCENARIO: MEDICAL RESEARCH

Please discuss the following scenario as a group. Nominate one person from your group to act as a scribe. You have 30 minutes to complete the following activity. At the end of this data situation exercise, 5 minutes will be allocated to each group to briefly explain their findings to the other groups.

Brief overview: You are part of a privately-funded medical research team who are focused on an examination of the (non-) effective treatment of *condition A*. Your research relies on large-scale data analysis of health data from a variety of sources, including private health clinics, pharmaceutical companies, survey companies and personal health-trackers.

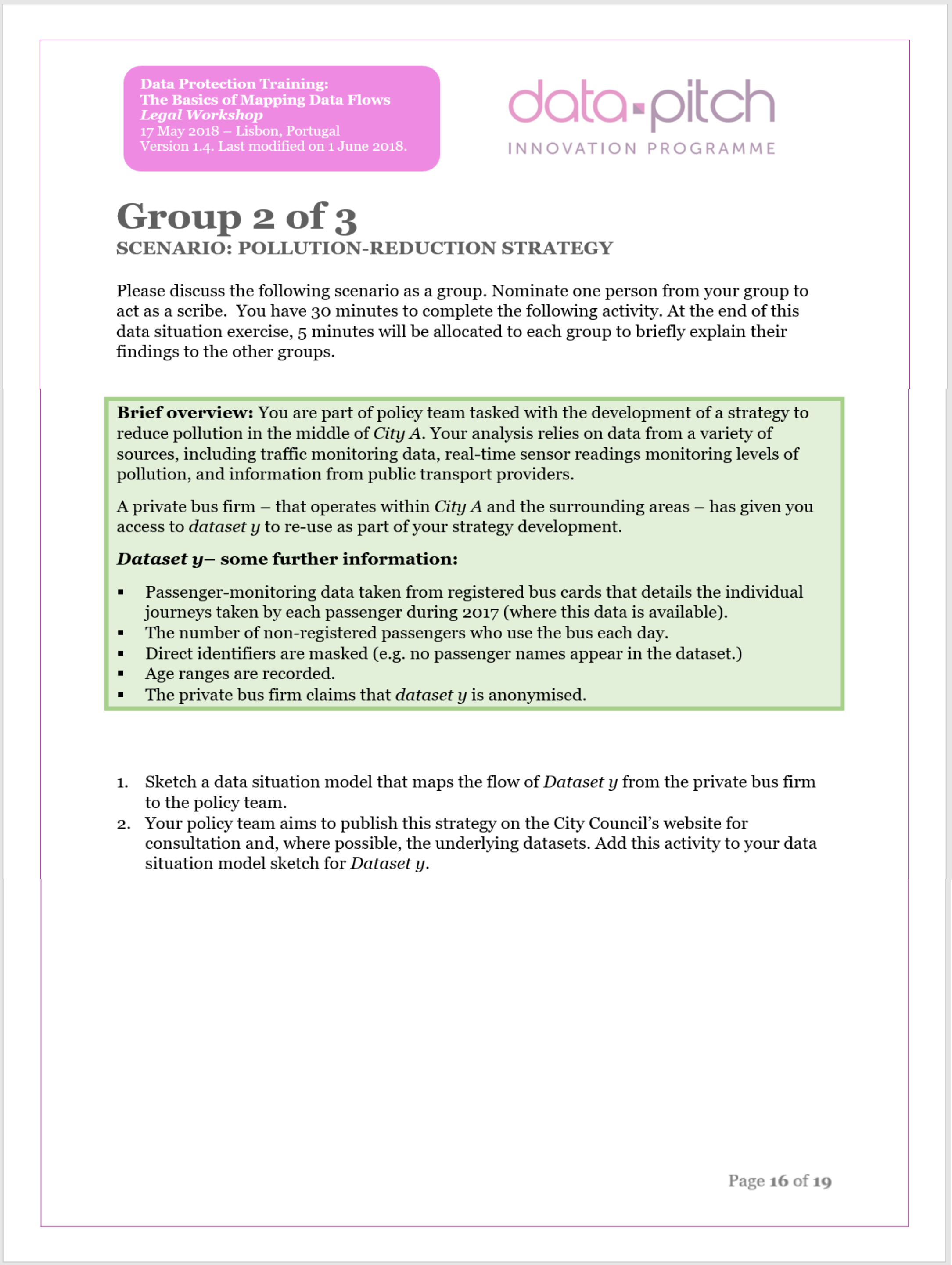
You have been given access to *dataset x* to re-use as part of your research into the treatment of *condition A*.

***Dataset x* – some further information:**

- Patient-monitoring data, including the course of treatment of administered to 1278 patients with *condition A*.
- Individual-level data.
- Collected by medical staff working at *Hospital Z*, *Ward G* during 2010-2015.
- Direct identifiers are masked (e.g. no patient names appear in the dataset).
- The data provider is a data market place that claims *dataset x* is anonymised.

1. Sketch a data situation model that maps the flow of *Dataset x* from *Ward G* of *Hospital Z* to the medical research team.
2. Your medical research team aims to publish a research paper in an open access journal and, where possible, the underlying datasets. Add this activity to your data situation model sketch for *Dataset x*.

Page 15 of 19



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

Group 2 of 3

SCENARIO: POLLUTION-REDUCTION STRATEGY

Please discuss the following scenario as a group. Nominate one person from your group to act as a scribe. You have 30 minutes to complete the following activity. At the end of this data situation exercise, 5 minutes will be allocated to each group to briefly explain their findings to the other groups.

Brief overview: You are part of policy team tasked with the development of a strategy to reduce pollution in the middle of *City A*. Your analysis relies on data from a variety of sources, including traffic monitoring data, real-time sensor readings monitoring levels of pollution, and information from public transport providers.

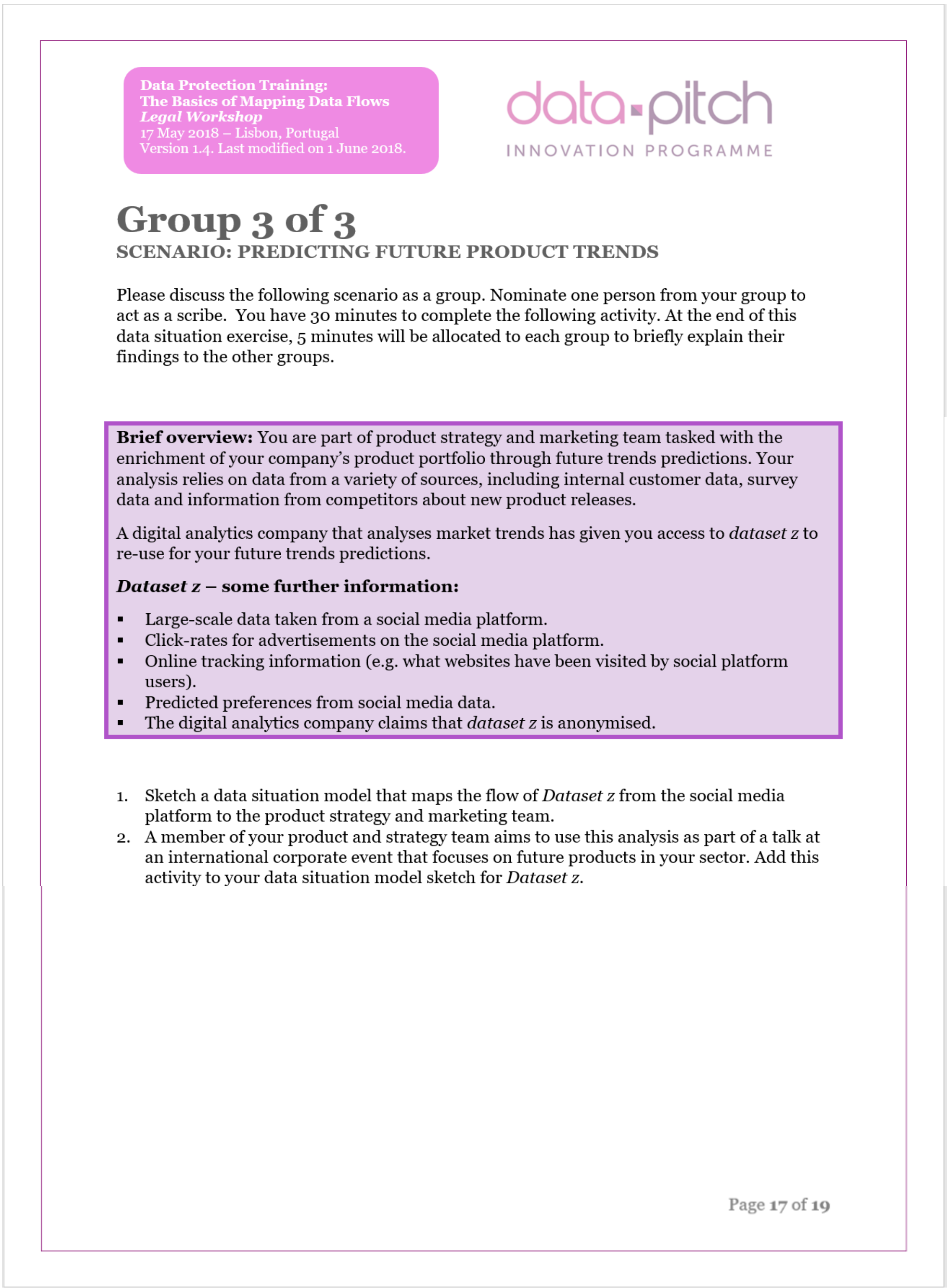
A private bus firm – that operates within *City A* and the surrounding areas – has given you access to *dataset y* to re-use as part of your strategy development.

***Dataset y*– some further information:**

- Passenger-monitoring data taken from registered bus cards that details the individual journeys taken by each passenger during 2017 (where this data is available).
- The number of non-registered passengers who use the bus each day.
- Direct identifiers are masked (e.g. no passenger names appear in the dataset.)
- Age ranges are recorded.
- The private bus firm claims that *dataset y* is anonymised.

1. Sketch a data situation model that maps the flow of *Dataset y* from the private bus firm to the policy team.
2. Your policy team aims to publish this strategy on the City Council's website for consultation and, where possible, the underlying datasets. Add this activity to your data situation model sketch for *Dataset y*.

Page **16** of **19**



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

data·pitch
INNOVATION PROGRAMME

Group 3 of 3

SCENARIO: PREDICTING FUTURE PRODUCT TRENDS

Please discuss the following scenario as a group. Nominate one person from your group to act as a scribe. You have 30 minutes to complete the following activity. At the end of this data situation exercise, 5 minutes will be allocated to each group to briefly explain their findings to the other groups.

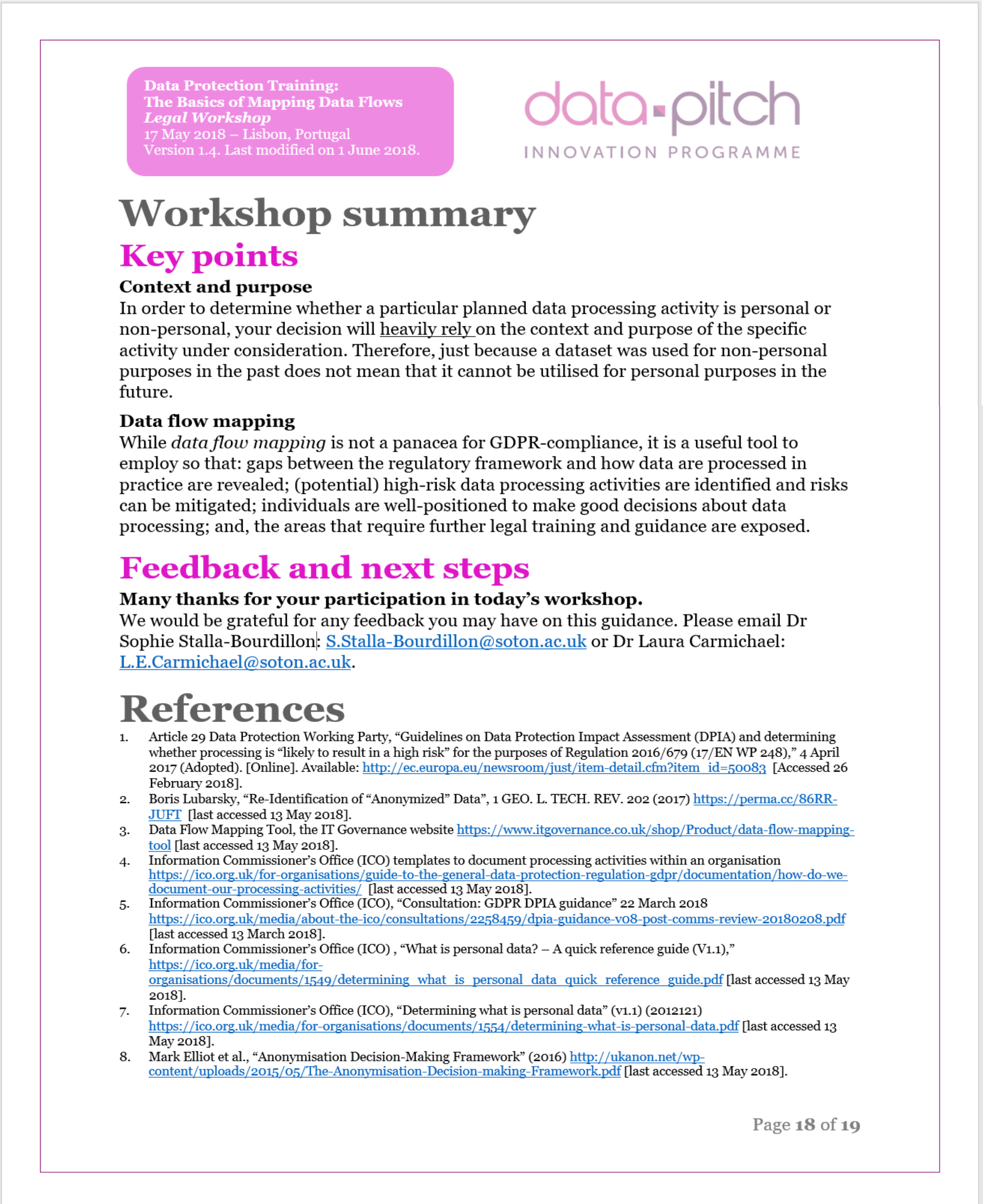
Brief overview: You are part of product strategy and marketing team tasked with the enrichment of your company's product portfolio through future trends predictions. Your analysis relies on data from a variety of sources, including internal customer data, survey data and information from competitors about new product releases.

A digital analytics company that analyses market trends has given you access to *dataset z* to re-use for your future trends predictions.

***Dataset z* – some further information:**

- Large-scale data taken from a social media platform.
- Click-rates for advertisements on the social media platform.
- Online tracking information (e.g. what websites have been visited by social platform users).
- Predicted preferences from social media data.
- The digital analytics company claims that *dataset z* is anonymised.

1. Sketch a data situation model that maps the flow of *Dataset z* from the social media platform to the product strategy and marketing team.
2. A member of your product and strategy team aims to use this analysis as part of a talk at an international corporate event that focuses on future products in your sector. Add this activity to your data situation model sketch for *Dataset z*.



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

Workshop summary

Key points

Context and purpose

In order to determine whether a particular planned data processing activity is personal or non-personal, your decision will heavily rely on the context and purpose of the specific activity under consideration. Therefore, just because a dataset was used for non-personal purposes in the past does not mean that it cannot be utilised for personal purposes in the future.

Data flow mapping

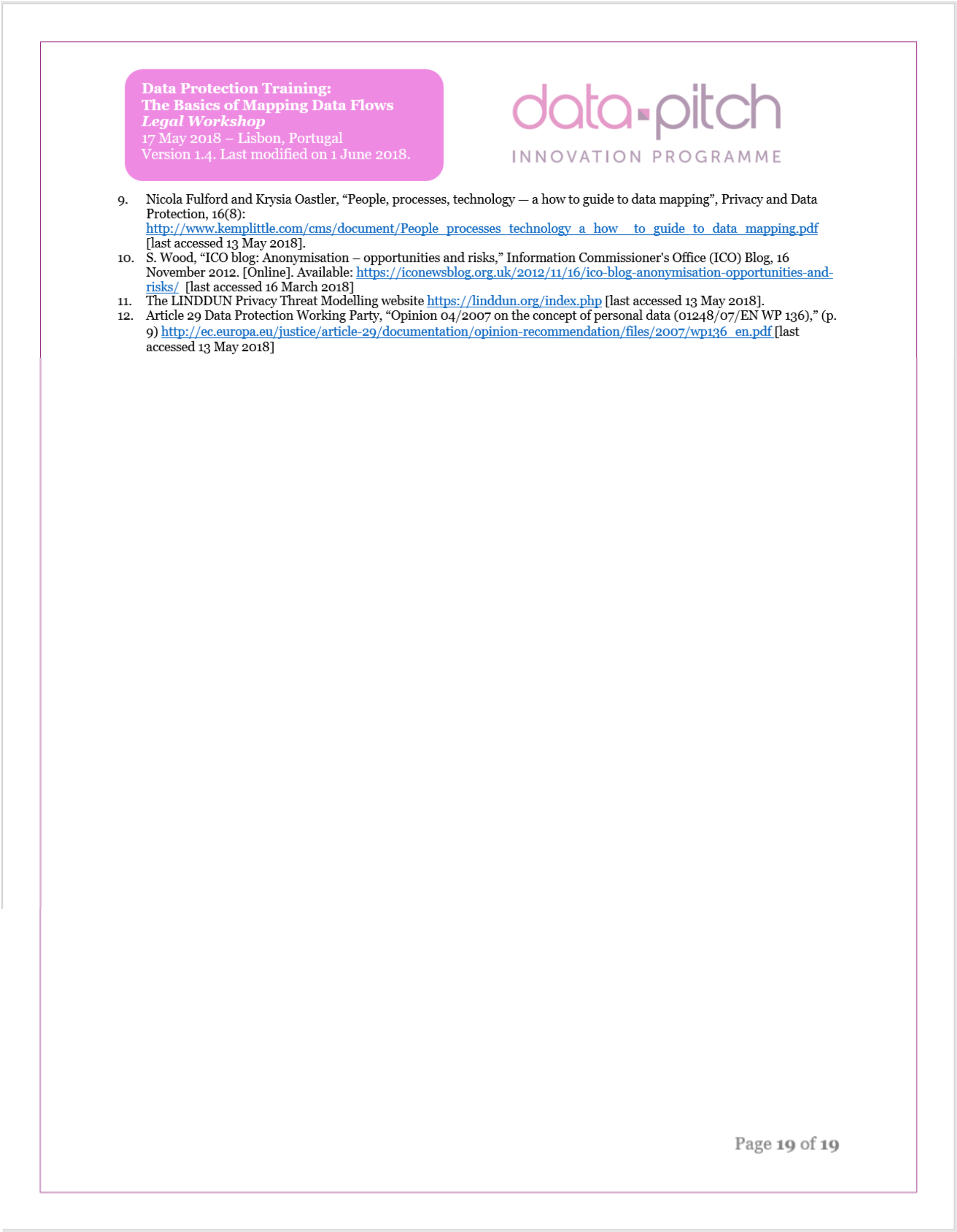
While *data flow mapping* is not a panacea for GDPR-compliance, it is a useful tool to employ so that: gaps between the regulatory framework and how data are processed in practice are revealed; (potential) high-risk data processing activities are identified and risks can be mitigated; individuals are well-positioned to make good decisions about data processing; and, the areas that require further legal training and guidance are exposed.

Feedback and next steps

Many thanks for your participation in today's workshop.
We would be grateful for any feedback you may have on this guidance. Please email Dr Sophie Stalla-Bourdillon: S.Stalla-Bourdillon@soton.ac.uk or Dr Laura Carmichael: L.E.Carmichael@soton.ac.uk.

References

1. Article 29 Data Protection Working Party, "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 (17/EN WP 248)," 4 April 2017 (Adopted). [Online]. Available: http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=50083 [Accessed 26 February 2018].
2. Boris Lubarsky, "Re-Identification of "Anonymized" Data", 1 GEO. L. TECH. REV. 202 (2017) https://perma.cc/86RR-JUFT [last accessed 13 May 2018].
3. Data Flow Mapping Tool, the IT Governance website https://www.itgovernance.co.uk/shop/Product/data-flow-mapping-tool [last accessed 13 May 2018].
4. Information Commissioner's Office (ICO) templates to document processing activities within an organisation https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/documentation/how-do-we-document-our-processing-activities/ [last accessed 13 May 2018].
5. Information Commissioner's Office (ICO), "Consultation: GDPR DPIA guidance" 22 March 2018 https://ico.org.uk/media/about-the-ico/consultations/2258459/dpia-guidance-v08-post-comms-review-20180208.pdf [last accessed 13 March 2018].
6. Information Commissioner's Office (ICO) , "What is personal data? – A quick reference guide (V1.1)," https://ico.org.uk/media/for-organisations/documents/1549/determining_what_is_personal_data_quick_reference_guide.pdf [last accessed 13 May 2018].
7. Information Commissioner's Office (ICO), "Determining what is personal data" (v1.1) (2012121) https://ico.org.uk/media/for-organisations/documents/1554/determining-what-is-personal-data.pdf [last accessed 13 May 2018].
8. Mark Elliot et al., "Anonymisation Decision-Making Framework" (2016) http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf [last accessed 13 May 2018].



Data Protection Training:
The Basics of Mapping Data Flows
Legal Workshop
17 May 2018 – Lisbon, Portugal
Version 1.4. Last modified on 1 June 2018.

data·pitch
INNOVATION PROGRAMME

9. Nicola Fulford and Krysia Oastler, "People, processes, technology — a how to guide to data mapping", Privacy and Data Protection, 16(8): http://www.kemplittle.com/cms/document/People_processes_technology_a_how__to_guide_to_data_mapping.pdf [last accessed 13 May 2018].
10. S. Wood, "ICO blog: Anonymisation – opportunities and risks," Information Commissioner's Office (ICO) Blog, 16 November 2012. [Online]. Available: https://iconewsblog.org.uk/2012/11/16/ico-blog-anonymisation-opportunities-and-risks/ [last accessed 16 March 2018]
11. The LINDDUN Privacy Threat Modelling website https://linddun.org/index.php [last accessed 13 May 2018].
12. Article 29 Data Protection Working Party, "Opinion 04/2007 on the concept of personal data (01248/07/EN WP 136)," (p. 9) http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf [last accessed 13 May 2018]

Page **19** of **19**

Annex 3: Future of the prototype tool

While there are currently no formal plans in place to further develop the prototype tool (accessible via http://pz-wp-test.synote.io/ [last accessed 4 June 2018]), there are three main areas for future improvement:

1. **More written content.** It would be very useful to integrate more of the written guidance provided in the workshop hand-out alongside the interactive decision-trees; including further pop-up boxes with more information. It would also be beneficial to include the three fictional mapping scenarios (located in Part B).

2. **Include images of the legal decision-trees.** It would be valuable to display the whole legal-decision tree to the user – after they have used each section of the tool – in order to show how purpose and context affects the status of the planned data processing activity and the appropriate controls required.

3. **Display relevant parts of the legal decision-tree.** It would be helpful to display relevant parts of the decision-tree to users alongside the particular question they are answering.

Annex 4: Basic data flow mapping examples

This annex provides some very basic examples of data flow maps based on the types of data environments (potentially) present as part of the Data Pitch programme. Note: while it follows guidance provided by [1], it is very basic in the sense that it does not offer any detail on the distinct *"configuration"* of *"people, other data, infrastructure and governance structures"* [1, p. 69] that pertain to each data environment.

Types of datasets

There are two types of datasets (re)used by the start-ups for the purposes of the Data Pitch programme:

1. **Data provider datasets –** i.e. datasets that are directly supplied by data providers who have a contractual relationship with Data Pitch for the purpose of re-use in response to a specified challenge.
2. **Self-sourced datasets –** i.e. datasets that are obtained by the start-ups through internal data collection processes and/or from other third parties who do not have a contractual relationship with Data Pitch as a data provider.

Map the data flow (i): Data-provider data – data storage and access

The Data Pitch programme outlines the four possible options for data storage and access to data shared by Data Providers: (1) the University of Southampton hosts the data; (2) the Data Provider hosts the data; (3) the commercial cloud hosts the data (under the direction of the Data Provider); or, (4) the Participating SME (chosen to process the relevant data) hosts the data. 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.[34]

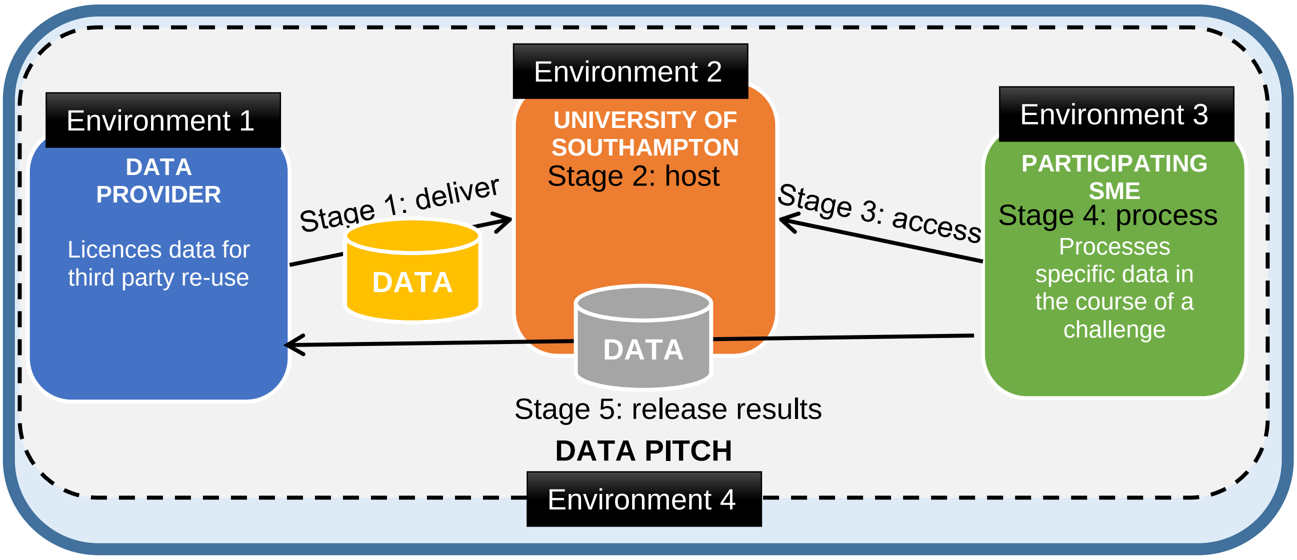


Figure 9 Data Situation A: simple share between Data Provider and Participating SME where University of Southampton hosts the data

[34] For further evaluation of these four possible options for data storage and access as part of the Data Pitch programme, see: [4, pp. 3-4, 52-53].

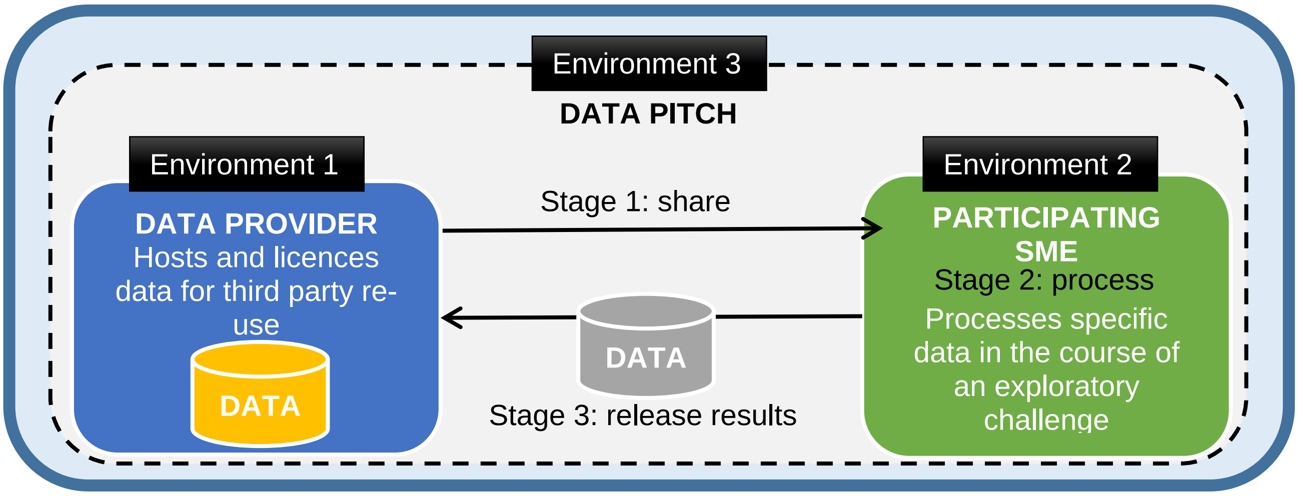


Figure 10 Data Situation B: simple share between Data Provider (host) and Participating SME

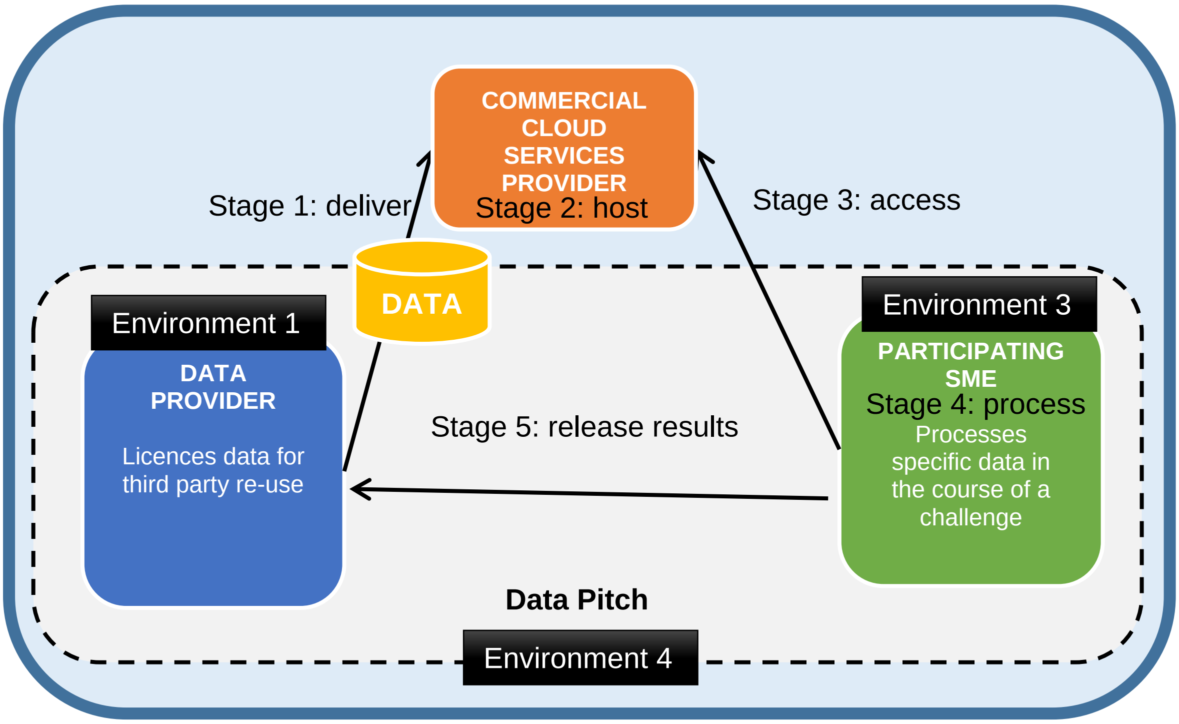


Figure 11 Data Situation C: simple share between Data Provider (host) and Participating SME via third party host (commercial cloud services provider)

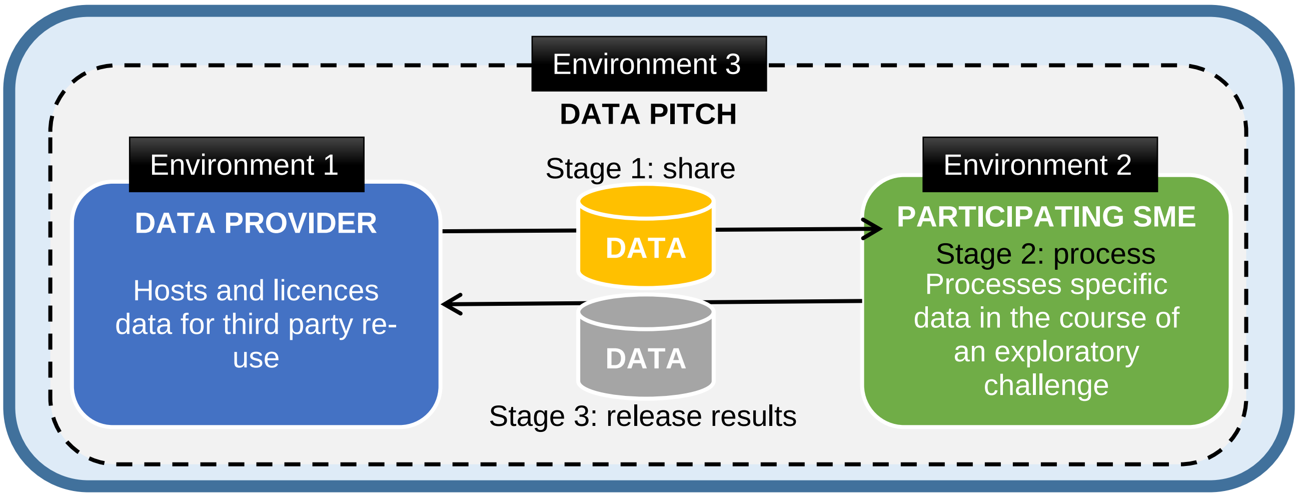


Figure 12 Data Situation D: simple share between Data Provider and Participating SME (host)

Map the data flow (ii): Self-sourced data

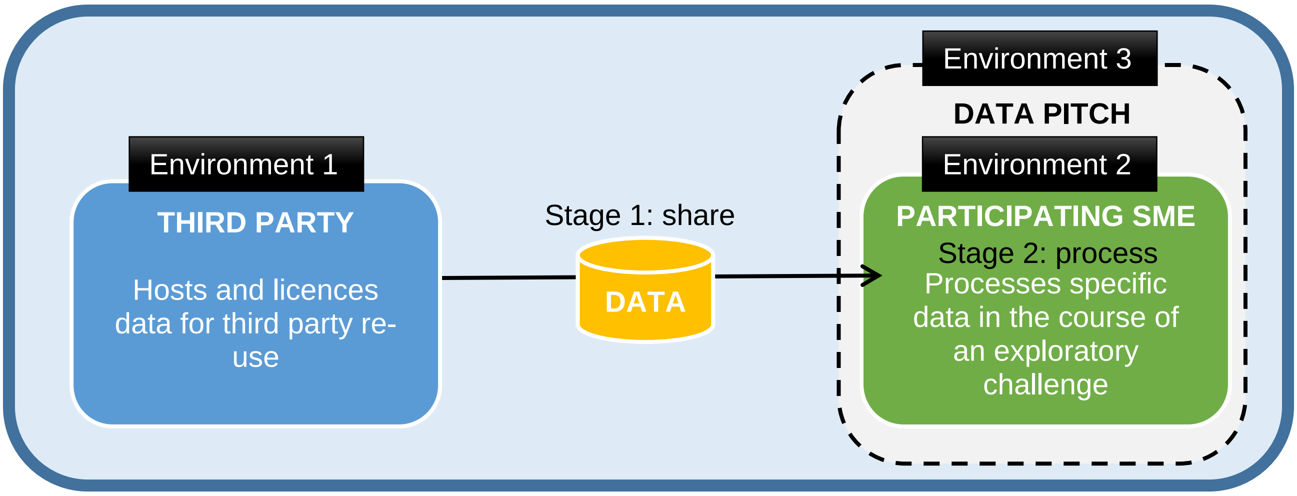

Environment 3
DATA PITCH
Environment 1
THIRD PARTY
Hosts and licences data for third party re-use
Stage 1: share
DATA
Environment 2
PARTICIPATING SME
Stage 2: process
Processes specific data in the course of an exploratory challenge

Figure 13 Data situation 1: simple single-share between Participating SME and Third Party

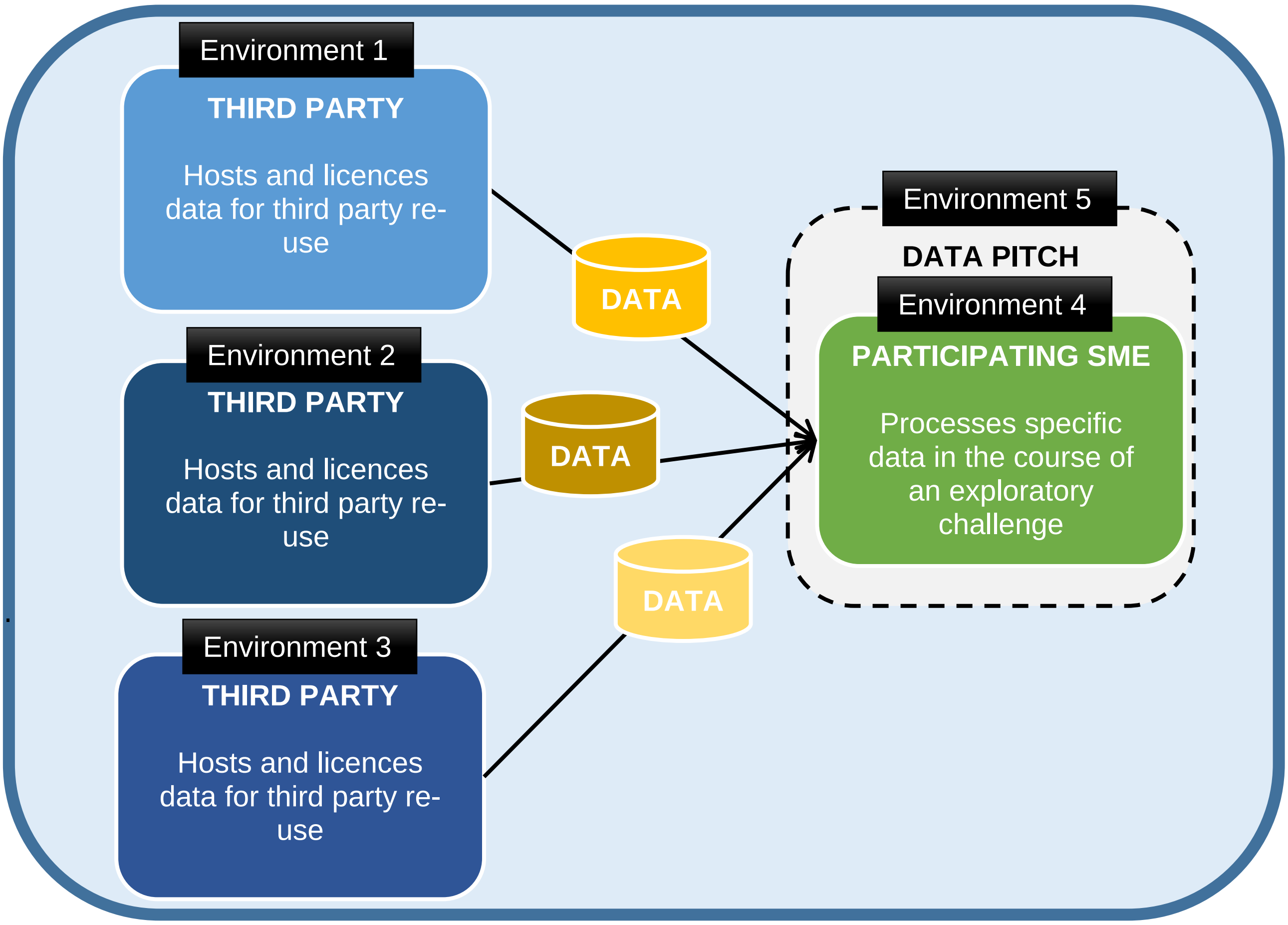

Environment 1
THIRD PARTY
Hosts and licences data for third party re-use
Environment 2
THIRD PARTY
Hosts and licences data for third party re-use
Environment 3
THIRD PARTY
Hosts and licences data for third party re-use
DATA
DATA
DATA
Environment 5
DATA PITCH
Environment 4
PARTICIPATING SME
Processes specific data in the course of an exploratory challenge

Figure 14 Data situation 2: simple multi-share between Participating SME and several Third Parties