



Purpose limitation under the GDPR: can Article 6(4) be automated?

Master Thesis

Study: Law and Technology, LLM

University: Tilburg University, The Netherlands

Author: Zhasmina Radkova Kostadinova

ANR: 661183

Supervisor: prof. dr. R.E. Leenes

Co-supervisor: A. (Aviva) de Groot LLM

Abstract

GDPR's principles of processing personal data are essential for the complete and effective protection of individuals and their personal data. Of particular importance is the purpose limitation principle which symbolizes the shift of responsibility from the weak individuals (data subjects) to the more powerful party (data controllers) – the more powerful one has to pre-emptively adhere to the principle and demonstrate compliance. Although the purpose limitation principle is a subject of guidance from Working Party 29, local Data Protection Authorities and scholars, there is yet no comprehensive and unitary way of determining when a purpose obeys to the purpose specification and furthermore in which cases a further processing of personal data would be considered compatible and why. In order to address this lack of clarity, this work researched whether it is possible to automate Article 6(4) GDPR on the basis of a body of knowledge about it. Upon establishing a body of knowledge regarding Article 6(4) and acknowledging the challenges which needed to be met by a legal knowledge engineering method in order to meet the aim of this work, a dataset was created. This dataset was then used to train several supervised machine learning classifiers, the results of which are highly promising but do not provide for a complete automation of Article 6(4) GDPR.

Table of Contents

Abstract	2
Chapter 1	5
1.1 Introduction	5
1.2 Research Questions	9
1.3 Methodology.....	10
Chapter 2	14
2.1 The definition of compatible further processing	14
2.2 Privacy and data protection in our computerized societies.....	16
2.3 Fair Information principles.....	19
2.4 The body of knowledge surrounding Article 6(4) GDPR	23
2.4.1 Purpose limitation within the GDPR	25
2.4.2 WP29’s opinion on the purpose limitation principle.....	29
2.4.3 Data Protection Authorities’ guidance and consultations.....	45
2.4.4 Case law.....	52
2.4.5 Scholars	61
2.5 The body of knowledge surrounding Article 6(4) GDPR	65
Chapter 3	74
3.1 Challenges in the legal knowledge engineering	76
3.1.1 Ambiguity and vagueness of legal texts	76
3.1.2 One-to-many & Many-to-one representations of knowledge	79
3.1.3 Ex ante vs. ex post	79
3.1.4 Legal concepts are never fully determined.....	80
3.2 The challenges of automating Article 6(4) GDPR	82
3.3 Methods of legal knowledge engineering.....	84
3.2.1 Expert systems and logic programming	84
3.2.2 Machine learning algorithms	87
3.2.3 Ontologies and taxonomies for legal text analysis	92
Chapter 4	99
4.1 The method to automate Article 6(4) GDPR.....	99
4.1.1 The dataset.....	99
4.1.2 The classification and evaluation of its performance	102
4.2 Discussion and Conclusion.....	108
Annex	110

Bibliography121

*To know the laws is not to memorize their letter but
to grasp their full force and meaning.*
Marcus Tullius Cicero (106–43 B.C.)

Chapter 1

1.1 Introduction

When I made a purchase account at Zalando I was not expecting to learn, a few months later, that Zalando had shared my email address with Facebook¹. Online retailers, among other companies, upload lists with consumer emails on Facebook, and other social platforms, to check whether the social media has an account corresponding to the email address. Upon a match, the consumer would start receiving custom advertisements from the retailer or reminders to complete an order – the so-called ‘custom’ or ‘matched’ audiences features.

Now, post-facto, I am informed about this processing of my personal data, but this was not the case by default. In the context of data protection, transparency is an important and long established feature of the European Union’s legislation. It brings trust to consumers by helping them understand how their personal data is processed and enables them to challenge unexpected processing operations (WP260 rev.01, 2018, p. 4). Although transparency is codified in the Treaty of the European Union (TEU) and strongly reinforced in the General Data Protection Regulation (GDPR), full transparency about why and how my personal data has been used, in practice, still has a long way to go.

The GDPR is “perhaps the most comprehensive and forward looking piece of legislation to address the challenges facing data protection in the digital age” (Zarsky, 2016, p. 995). Under the GDPR, the data controller, or the person who determines the purposes and the means of the personal data processing, must always be able to demonstrate that the data are processed lawfully, fairly and in a transparent manner in relation to the data subject (WP260 rev.01, 2018, p. 5). In the example given above, I am the data subject and Zalando is the data controller. Therefore, Zalando must be able to prove that they shared my email address with Facebook in a lawful, fair and transparent manner. One step in enabling such a proof is to demonstrate adherence to the purpose limitation principle. The principle states that personal data shall be “collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes” (Article 5(1b) GDPR). Hence, Zalando has to prove that the uploading of my email address to Facebook is compatible with the purpose for which it was collected. By giving me an opportunity to make

¹ Which I only got to know because the public and politicians were increasing pressure on Facebook after the Cambridge Analytica revelations

a distinction between processing for collection and further use, the purpose limitation principle supports the idea of transparency and binds the more powerful one to a set of rules (Koning, 2015, p. 2). As such, the principle is a core trait of the European personal data protection regime and it “plays an important role in the protection of human rights and the safeguarding of the free flow of personal data” (Koning, 2015, p. 5).

In this specific case, supposedly, only my email address was shared with Facebook. Even though it is *just* an email address “[t]he risk to data protection comes from the purpose(s) of processing” and not only the categories of the personal data collected (WP 194, p. 5). Thus, solely my email address may not reveal a great amount of personal information about me, however, when combined with other data, for example my social medial account, the results could be substantial. Such practices of sharing and combining personal data in return for tailored offerings are the default now. However, the data subjects are not made aware of how those offerings are formed and how their personal data is processed. A power imbalance is created, in which customers give their data for one specific purpose but companies decide to exploit that data for additional purposes beyond the initial one.

The GDPR does not prohibit the use of personal data for other purposes, however, it restricts those other purposes to be within the boundaries of the purpose limitation principle. The strength of the purpose limitation principle lies within its double function – first, being an autonomous principle and second, a condition for Article 5(1) GDPR which codifies all principles of processing of personal data to be demonstrated by the data controller. The autonomous function is that the principle, by itself, sets out an obligation for personal data to be collected for specified, explicit and legitimate purposes and not to be further processed in a way incompatible with those purposes. The principle needs to be complied with, unless there is an applicable exception proven by the data controller. In contrast, the conditional function is that the purpose limitation is dependent on the other principles (such as data minimization, accuracy, integrity and confidentiality) to fully enfold the protection of personal data, and it facilitates *real* protection (separation between initial and further purposes and prohibition of incompatible processing). This is a by-product of the positioning of the principle within Article 5(1) GDPR. Hence, the purpose limitation facilitates transparency and fairness of any processing of personal data. Moreover, it allows for the application of the other principles defined in Article 5(1) GDPR – data minimization, accuracy, storage limitation, integrity and confidentiality, and accountability. Thus, the purpose limitation is at the core of data protection and any “erosion of the conception of the purpose limitation principle results in the erosion of all related data protection principles” (Koning, 2015, p. 5).

According to data protection scholars, the GDPR and specifically the purpose limitation principle are already eroding. Zarsky (2016, p. 996) states that the GDPR “fails to properly address the surge” of data practices² and its provisions are “incompatible with the data environment³”. Similarly, Prins & Moerel (2016, p. 7) argue that the GDPR “does not reflect the reality of everyday life sufficiently in order to be effective and accepted as legitimate”. Specifically, Prins & Moerel (2016, p. 7) criticize the purpose limitation principle on the basis that it only fits the conventional ways of working with personal data (first define a goal and then collect any data needed to achieve the goal); and it does not fit within the current practices of data analytics, where there may not be an original purpose - data is collected in order to subsequently be able to offer potential new services on the basis of the analysis of that data – a practice called data mining. Many companies argue that they collect personal data for the purpose of ‘data collection and analysis’ which is a very broad statement that includes countless processing operations and is self-redundant, thus leaving the purpose limitation no longer meaningful. An why would companies not define such broad purposes? With the current status of the GDPR and the guidance available for the purpose limitation principle, there always is a chance that an existing process violates a provision of the GDPR or it will introduce a violation in the future. In order to proof compliance, companies need to record their analysis of why they are compliant. Thus, an inevitable part of data protection compliance is the obligation to document and hope that once the data protection authorities check the documentation it is actually compliant.

The criticism on the adequacy of the current data protection framework and the continuous opaque use of personal data for purposes other than the collection purposes, makes the purpose limitation principle and the compatibility test for further processing of personal data more relevant than ever. Especially important is that any data controller is confident in their collection and further purposes of processing personal data. Hence, taking into account that an inevitable part of data protection compliance is the obligation to document and be able to proof compliance with the GDPR, a method to proof such compliance would be of great value for data controllers. Therefore, this work researches and analyses multiple sources of information regarding the purpose limitation principle under Article 5(1) and the compatibility test for further processing of personal data under Article 6(4) GDPR. The main purpose of this work is to gather enough information to construct a comprehensive body of knowledge about the purpose limitation principle and to research the possibility of applying an automated method which would be able to distinguish between the compatible cases of further processing of personal data from the incompatible ones with a high degree of accuracy. Such a solution would be of a good fit for all data controllers, who are urgently in the need of guidance on when their

² Large volume of data sets whose size is growing at a vast speed, thus making it difficult to handle such amount of data using traditional software tools available, data mining techniques and database management tools (Mohammed & Humbe, 2016, p. 3).

³ Data environment to be understood as the commonly used methods which facilitate the collection, analysis, sharing, storing and/or destruction of data.

purposes are (in)compatible. Optimally, the solution should also be easy to use, hence there is a pressing need to explore any automation possibilities.

The construction of a comprehensive body of knowledge, sufficient for an automated method to be able to ‘decide’ on the compatibility of an outcome is a challenging task for a number of reasons. For computer programs the concepts of purpose specification and (in)compatibility of further processing are too vague to ‘understand’ – machines need clear (binary) instructions. Thus, in order for a machine to operate in this domain, it should have an explicit ‘knowledge’ of the meaning and applicability of each word and concept. For example, when scholars analyse a legal principle, they search to understand the interpretation of each word first separately, then in conjunction with other terms and last the applicability of the concept within a set of facts and circumstances. After many years of experience and constant learning about the interpretation of a certain rule, legal scholars develop methods, including a gut-feeling, on how to cope with concepts such as the incompatibility of further processing of personal data. The challenge is to bridge the gap between a legal scholar and a machine. This work will simulate the knowledge gathering that a legal scholar would carry out in order to construct a comprehensive body of knowledge about Article 6(4) GDPR in order to evaluate whether the provision can be automated, performed by a junior data protection specialist working in the field of advising companies on how to be GDPR compliant.

Not to get ahead of ourselves, we conclude that there are no methods which can fully automate Article 6(4) GDPR. This is due to the presence of vague terms, the interpretation of which is unclear, highly circumstantial and require judgement rather than logic. Nevertheless, machine learning classification methods offer some support in automating Article 6(4) GDPR. By collecting and forming a body of knowledge which served as the basis of a dataset, different supervised machine learning classifiers were trained to predict the outcome of a further processing of personal data as either compatible or incompatible. Although this method addressed has very encouraging results and it filled in a gap in the literature regarding the automation of any principle under the GDPR, it has its limitations and should be a subject of peer review.

1.2 Research Questions

The focus of this work is the extraction of a comprehensive and explicit knowledge about Article 6(4) GDPR. Various studies have observed that the wording of the purpose limitation principle, especially the notion of compatible use, is “very open-ended, which leaves the concept susceptible to different interpretations” (WP29 203, 2013, p. 5; Korff, 2002, p. 243). Therefore, in order to better understand the notion of compatible use as part of the purpose limitation principle this work will focus on three research questions:

Firstly, what is the definition of the notion of compatible further processing of personal data? Does this definition give us enough information on what is compatible or not?

Secondly, what do we know about the notion of compatible further processing of personal data as part of the purpose limitation principle? Can we better understand it by extracting a comprehensive and explicit body of knowledge about the purpose limitation principle from multiple sources in a manner similar to how a legal scholar would do, in order to become an expert of this domain?

Last but not least, what is the prospect of automation in the domain? Can we use the body of knowledge gathered to reduce the notion of compatible use, under Article 6(4) of the GDPR, to a set of rules and decisions/observations which allow for its automated processing? In other words: can the legal reasoning behind the notion of compatible use be automated? What are the preconditions for artificial intelligence systems in this domain? What should the input look like? What should the output look like? Does the knowledge gathered meet those requirements?

The answers of those research questions are enveloped in a methodology which is detailed in the next section.

1.3 Methodology

This paper focuses on the application area of regulatory compliance (Muthuri, et al., 2017, p. 2). In particular it searches to map out a complete and comprehensive body of knowledge which will allow for the automation of Article 6(4) GDPR.

The nature of this research follows Arthurs' (1983) taxonomy of legal research styles (Figure 1). The vertical axis represents the classic distinction between pure research “undertaken for a predominantly academic constituency”, and the applied research which aims to clarify constituencies to facilitate the professional needs of practitioners and policy makers (Chynoweth, 2008, p. 30). The horizontal axis represents the distinction between doctrinal and interdisciplinary research. Doctrinal research is the “formulation of legal ‘doctrines’ through the analysis of legal rules”, which always requires the relevant legal rules to be applied to the particular facts of the situation under consideration (Chynoweth, 2008, p. 31). Interdisciplinary research makes references to other, external, factors to analyse the existing body of rules and their application. In taking an external view of the law, this research is ‘*about law*’ rather than research ‘*in law*’, thus also adopting epistemology and methodology from the social sciences.

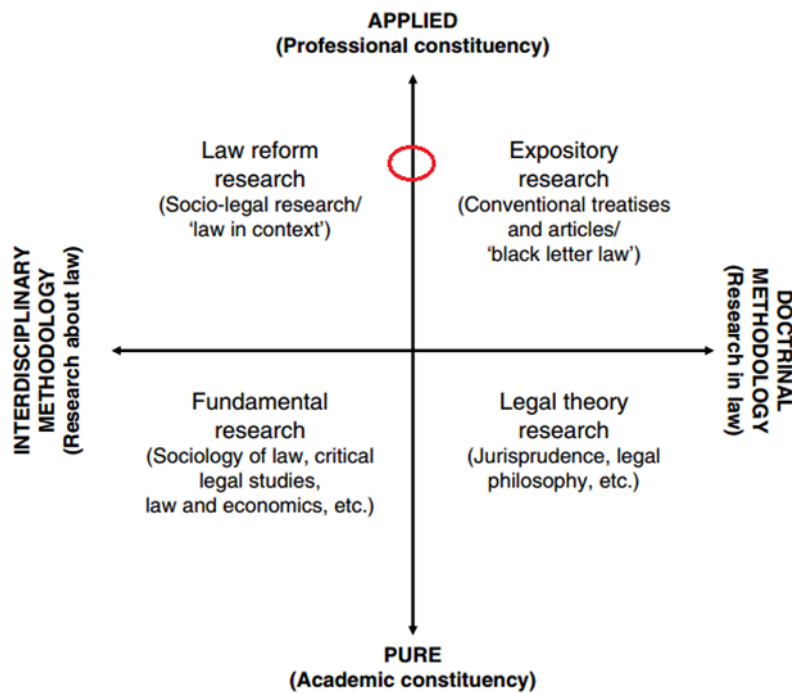


Figure 1: Arthur's taxonomy of legal research styles

This research represents the red circle present at Figure 1. Placed in the middle between interdisciplinary and doctrinal methodology, this work moves up the vertical axis towards applied research, aiming to gather an explicit knowledge of Article 6(4) with a particular purpose in mind. The purpose is to facilitate a future

change - a change in the regulatory environment – namely, how Article 6(4), the compatibility test for further processing of personal data, is to be reduced to a set of rules and decisions that allow for its automation advancement.

Furthermore, this research can be categorized into the classification of Watkins & Burton (2013, p. 12) and Siems & Síthigh (2012, p. 653) for one of three “ideal types” of academic legal research: “law as a practical discipline”, “law as humanities” and “law as social sciences”. In particular, this research fits into the category “law as a practical discipline” because it relies on the objective analysis of the purpose limitation principle and Article 6(4) of the GDPR from regulatory guidance, case law and scholars’ interpretation, in order to achieve a value-free analysis of legal rules, aiming to criticize, explain, correct and possibly direct the way this legal doctrine is to be applied on practice (Birks, 1998, p. 431).

The structure of this research is illustrated at Figure 2. In order to be able to author the target (Article 6(4) GDPR) in a manner similar to which a legal scholar would gather the needed information to become an expert in the domain, we gather information from firstly, the domain - data protection and fair information principles, and secondly, delve deep into any interpretations available about the purpose limitation principle and further processing of personal data.

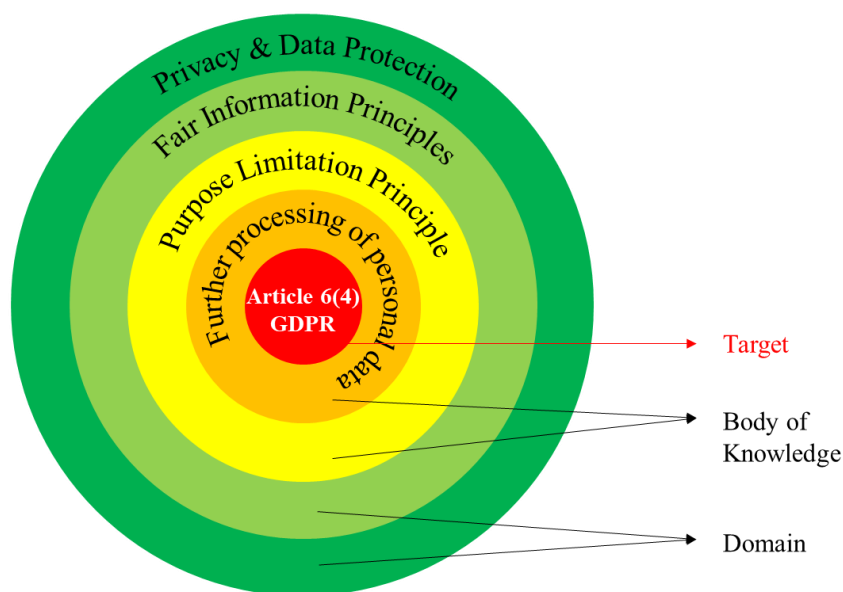


Figure 2: Scope of the research

The structure translates into a body of knowledge which follows the provisions described in the GDPR and the DPD; then adds any interpretations by Working Party 29 and the Data Protection Authorities; and last but not least resorts to external sources of guidance such as case law and domain experts. Overall, any guidance on the interpretation of Article 6(4) will help to create a comprehensive and explicit body of knowledge

about the purpose limitation principle and help reduce Article 6(4) to a set of given facts which will allow for its automation.

In order to meet the goals of this work, an extensive and systematic literature review was conducted. The literature review aims to establish a comprehensive body of knowledge about the purpose limitation principle. The term ‘knowledge’ refers to an organized semantic information about a particular subject-matter domain, and with the use of which a human can refer to procedures and strategies to understand and interpret the domain (Voss & Bisanz, 2017, p. 219).

The literature review started by limiting the search scope to the following key-phrases: ‘purpose limitation principle’, ‘purpose specification principle’ ‘further processing of personal data’ and ‘use limitation’. The purpose limitation principle has been a part of European Union’s legislation on personal data since 1981. This research focuses on finding interpretations about the purpose limitation in the context of two main paradigms. Firstly, purpose limitation within the European regulatory framework and secondly, purpose limitation within the spread and influence of the fair information principles. This is the case because the exact wording of the purpose limitation principle was not altered during the transit between the DPD and the GDPR. Hence, for the purposes of this research, any analysis for the interpretation and applicability of the purpose limitation principle under the 1981 Convention, the DPD or OECD’s fair information principles guideline is to be considered as applicable to the GDPR as well.

To identify the relevant legal and domain materials, the literature review involved the utilization of a wide range of computerized and as printed sources, such databases (e.g., Comparative & International Law Links, be-press Legal Repository, Google Scholar, HeinOnline, JSTOR, LTRC Law Review & Journal Search, Official Journal of the European Union), online and conventional libraries, archives and personal contacts to researchers or other experts in the field of data privacy, innovation, legal interpretation and business ethics. Inclusion criteria for the studies were methodological quality and potential for future research on the topic. The development of the research field of further processing of data is mainly led by the European Commission, thus the literature review incorporated not only academic sources (peer-reviewed journal publications, working papers, and conference papers) but also European Union’s legislation and Working Party 29’s publications (Shaikh & Karjaluo, 2015, p. 544). Searching within the above mentioned information systems, data protection and privacy journals and case law libraries, the main result matches came from interpretations from WP29, local data protection authorities’ guidance, opinions from domain experts working in the field of data protection and last but not least scholars’ work on the topic of purpose limitation across multiple jurisdictions. The observations from this review indicate the current level of thinking and perception of the data protection regulatory framework and guide potential interpretation of Article 6(4) within the current knowledge and understanding of the purpose limitation principle.

The literature review did include a criterion for the year of publishing (Carnevale, 2013). Since the GDPR is a new regulatory instrument, the literature review started from the most recent publications (2018 – 2016) and going backwards to original articles which are milestones regarding the purpose limitation principle, European Union's data protection analysis, fair information principles and most importantly further processing of personal data.

Chapter summary

This next chapter outlines the main issue to be discussed – understanding what Article 6(4) GDPR entails and whether it can be automated. Moreover, the chapter presents a methodology on how the issue at hand will be approached.

The driving force of this work is the need for more transparency on how personal data is processed which, is to be achieved only by a full compliance with the purpose limitation principle. Both Article 5(1) GDPR and Article 6(4) establish a number of requirements each data controller needs to comply with, thus allowing for an actual balance of power between the controller and the data subjects. Data subjects do not necessarily have more control under the GDPR, however, the controllers have stricter obligations, including enforcement, which are indented to give a better protection of individuals' personal data. Nevertheless, the wording of both the purpose limitation principle and Article 6(4) have been a subject of criticism. Vagueness and lack of sufficient application guidance may prevent data controllers from complying, without realizing it. To protect data subjects, a comprehensive and explicit body of knowledge which will allow for the automation of Article 6(4) GDPR would/might help data controllers to determine when they are compliant with the provision and when they may not further process certain personal data.

The next three chapters are striving to help both data controllers and data subjects. In Chapter 2 the comprehensive and explicit body of knowledge regarding Article 6(4) GDPR is formed. Chapter 3 analyses the possibility of the body of knowledge being computerized by presenting the main challenges in the legal knowledge engineering and the methods available from Artificial Intelligence to meet the needs of the body of knowledge. This analysis produces a result – an answer to the question whether Article 6(4) GDPR can be done by machines. That answer is discussed in detail in Chapter 4, which elaborates on the proposed outcome. Chapter 5 discusses the strengths and limitations of this research and concludes the findings.

Chapter 2

2.1 The definition of compatible further processing

The answer of the question ‘What is the definition of the notion of compatible further processing of personal data?’ is both a fairly straight forward one and yet non-existent. What we know is that Article 5(1b) GDPR lays down the purpose limitation as one of the six principles relating to the processing of personal data. The principle requires that 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’);

From this definition it can be observed that the purpose limitation principle consists of two components: first, the purpose specification and second, compatibility of further processing of personal data. The second component consists of five factors to be taken into account, defined in Article 6(4) GDPR:

4. Where the processing for a purpose other than that for which the personal data have been collected is not based on the data subject's consent or on a Union or Member State law which constitutes a necessary and proportionate measure in a democratic society to safeguard the objectives referred to in Article 23(1), the controller shall, in order to ascertain whether processing for another purpose is compatible with the purpose for which the personal data are initially collected, take into account, inter alia:

- (a) *any link* between the purposes for which the personal data have been collected and the purposes of the intended further processing;
- (b) the *context* in which the personal data have been collected, in particular regarding the relationship between data subjects and the controller;
- (c) the *nature* of the personal data, in particular whether special categories of personal data are processed, pursuant to Article 9, or whether personal data related to criminal convictions and offences are processed, pursuant to Article 10;
- (d) the *possible consequences* of the intended further processing for data subjects;
- (e) the *existence of appropriate safeguards*, which may include encryption or pseudonymisation.

Article 6(4) defines the factors to be taken into account in order to determine whether a further purpose is compatible or not. However, the presence of the definition does not provide an answer to what is compatible or not. Hence, there is no answer to the above asked question. Additional information is needed in order to understand how to apply the five factors, in order to determine the outcome and understand the notion of compatible further processing of personal data.

Therefore, the next sections will discuss multiple sources of information which will ultimately help form a better understanding of the purpose limitation principle and the compatibility assessment. The first step is to define the domains within which the purpose limitation principle is positioned, namely the domains of privacy and data protection.

2.2 Privacy and data protection in our computerized societies

In 1960, (Licklider, p. 4) anticipated a “man-computer symbiosis” in which men and computers cooperate in making decisions and controlling complex situations without inflexible dependence on predetermined programs:

“In the anticipated symbiotic partnership, men will set the goals, formulate the hypotheses, determine the criteria, and perform the evaluations. Computing machines will do the routinizable work that must be done to prepare the way for insights and decisions in technical and scientific thinking.” (Licklider, 1960, p. 4)

Today, such a symbiosis is a fact due to two developments - the computerization of our society which enhanced human’s knowledge about production and collection of data from different sources; and the tremendous amount of personal data produced from almost every aspect of our lives (Han, et al., 2011, p. xxiii). As a result, the societal focus shifted towards a complex ecosystem between companies and individuals where everyone engages in the aggregation and use of heterogeneous data because data leads to economic opportunities (Libaque-Saenz, et al., 2016, p. 339). Such an explosive growth of stored or transient data generates a constant need for ever newer techniques and smarter algorithms which are able to “intelligently assist us in transforming the vast amounts of data into useful information and knowledge” (Han, et al., 2011, p. xxiii). Hence, it has been widely stated in the academic domain that data is ubiquitous – human societies are data-driven (Pentland, 2013, p. 80; Mortier, et al., 2014).

In the realm of data-driven societies, personal data allows businesses to offer customized solutions and consumers to receive better service in return of their information. Services based on the analysis and aggregation of personal data (further processing), however, raise questions about the degrees of privacy intrusion and data protection of the already collected personal data. The use of those two terms, privacy and data protection, has risen drastically since the wide collection and use of personal data in our data-driven societies and specifically since the adoption of the GDPR. Moreover, privacy and data protection are terms often used inseparably. Although the two concepts are different, they do share similarities which is why their use is often simultaneous. Briefly put, privacy rights can be negatively enforced to prevent others from interfering with one’s private life, while data protection offers positive enforcement to protect any personal data processed.

A closer look on those two terms reveals that privacy and data protection are not black or white concepts, but instead they “intertwine, communicate and overlap in a grey zone” (Koning, 2015, p. 3). Privacy, on the one side, is a value – how much privacy do we want to have, or what of our private information are we willing to share and with whom (Warren & Brandeis, 1890, p. 193). Privacy allows value rationality, because it is the freedom to have a choice even if it is cost-inefficient (Xu, et al., 2014, p. 1149). Data protection, on the other

side, empowers the individual to exercise her right to demand transparency (Koning, 2015, p. 3). Data protection realizes the fundamental right of controlling how your personal data is used, by whom and for what reasons. It especially provides enforcement against processing operations which can have effects on one's private life, such as profiling. The interrelationship of privacy and data protection, their intertwining and overlapping in a grey zone remains a subject of academic and domain focus. Three aspects, about the relationship between data protection and privacy, observed by Koning (2015, p. 3), shed light upon that grey area of intertwinement, communication and overlapping.

First, the objective of the data protection's doctrine revolves around and includes the right to protection of personal data and the safeguard of privacy (Koning, 2015, p. 3). The task of data protection law, through substantive principles and procedural rules, is to balance the gains and threats of personal data processing (the right to protection of personal data) and to ensure that, while the benefits of data processing are taken advantage of, individuals and society at large are shielded from the negative effects (privacy). Hence, this specific grey area of intertwinement is the fact that data protection ultimately protects privacy and it will not be of such value if it was not for the 'right to be let alone' which everyone can relate to (Warren & Brandeis, 1890, p. 195).

Second, the scope of the data protection concept is the result of "decades of case law on the right to private life and communication in automated data processing cases" (Koning, 2015, p. 4). Although, the European Court on Human Rights (ECtHR) did not explicitly acknowledge a general right to protection of personal data, it did recognize aspects of the data protection doctrine under the scope of Article 8 (Right to respect for private and family life) of the European Convention on Human Rights (ECHR) (Koning, 2015, p. 4). It was due to ECtHR's influential case law that the right to protect personal data was explicitly codified in the Charter of Fundamental Rights of the European Union. With the introduction of the Lisbon Treaty, Article 16 TFEU established an explicit, directly applicable to persons right to the protection of personal data. Further reinforcing of this right was brought from revising, in 2009, Article 6 TEU, which recognized that the EU Charter of Fundamental Rights "shall have the same legal value as the Treaties". Thus, the fundamental rights defined in the Charter became directly incorporated into EU law, specifically Article 8 on the Protection of Personal Data. Article 8(1) of the Charter, being identical to Article 16(1) TFEU, established an explicit fundamental right to the protection of personal data, while Article 8(2) of the Charter called upon compliance with the principles for fair processing of personal data. Thus, it can be observed that through the value of privacy, data protection can be protected and vice versa, which translates into an inevitable, but of specific value for individuals, grey area of communication between the two.

Third, the highest court of the European Union (EU), the Court of Justice of the European Union (CJEU), judges on the application of the Charter of Fundamental Rights of the European Union, thus interpreting "data protection and privacy on a fundamental rights level" (Koning, 2015, p. 4). The CJEU prefers a joint

reading of Article 7 and 8 of the Charter - “the right to respect for private life with regard to the processing of personal data” (Joined Cases C-468/10 and C-469/10, 2011). This proves the highest recognition that data protection and privacy are two different concepts, with meaningful overlap, which optimally should be discussed together.

The grey area between privacy and data protection is not a bad thing. In contrast, as the three observations show, the grey area allows for the best of the two worlds. Hence, a protection to the individual on how their personal information is being treated and by whom, thus allowing for a value rational choice. A great example of such a protection is the European Union’s data protection framework. From its earliest efforts with the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data from 1981, until today’s General Data Protection Regulation (GDPR), its essential principles for processing of personal data (one of them being the purpose limitation principle) are based on the so-called ‘fair information principles’ (Koning, 2015, p. 4; Cate, 2006, p. 2). The fair information principles are the true source of intertwinement and overlap for the concepts of privacy and data protection, since they incorporate elements from both concepts. Most importantly, EU legislation is ultimately aiming to create a privacy-friendly environment which protects individuals’ data but and allows businesses to grow and use the economic opportunities created from the processing of personal data. However, what exactly are the fair information principles and how to they relate to Article 6(4) GDPR? The next section discusses that in detail.

2.3 Fair Information principles

“Fair information practices are the building blocks of modern information privacy law” (Schwartz, 1999, p. 1614). They can be found in data processing legislations around the world, especially the European Union (Cate, 2006, p. 343; Koning, 2015, p. 2; Gellman, 2017, p. 1).

A set of fair information principles was established for the first time in the early 1970s, when concerns about computerized databases prompted the US government to examine the technological and legal issues of such databases. With a report, “Records, Computers and the Rights of Citizens”, from 1973, US Congress was urged to adopt a “Code of Fair Information Practices,” based on five principles (Cate, 2006, p. 345):

1. There must be no personal data record-keeping systems whose very existence is secret.
2. There must be a way for a person to find out what information about the person is in a record and how it is used.
3. There must be a way for a person to prevent information about the person that was obtained for one purpose from being used or made available for other purposes without the person’s consent.
4. There must be a way for a person to correct or amend a record of identifiable information about the person.
5. Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take precautions to prevent misuses of the data.

In 1980 the Committee of Ministers of the Organization for Economic Cooperation and Development (OECD) revised the US government’s principles in a document which became influential internationally – the Guidelines on the Protection of Privacy and Trans-border flows of Personal Data (Gellman, 2017, p. 6). The Guidelines outlined eight elaborate and detailed principles⁴ for data protection and the free flow of

⁴ 1. Collection Limitation Principle – There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject.

2. Data Quality Principle – Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date.

3. Purpose Specification Principle – The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfilment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.

4. Use Limitation Principle – Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with [the Purpose Specification Principle] except: (a) with the consent of the data subject; or (b) by the authority of law.

5. Security Safeguards Principle – Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorized access, destruction, use, modification or disclosure of data.

information, which became the standard for fair information principles. Those principles brought for transparency, purpose limitation, data quality, security and data subject rights. Interestingly, the purpose limitation principle here is split into two separate principles (purpose specification and use limitation), thus putting a specific emphasis not only on the importance of the use of personal data strictly for the purposes collected but and the further-processing rule. Furthermore, one of the principles codified is accountability - a very important requirement within the current EU's data protection legislation, symbolizing the shifting responsibility of protecting personal data towards the powerful party (data controllers) and protecting the data subjects – the presumed weaker party. Those Guidelines brought for a change in the regulatory mind-set, however, they had no legal force to protect individuals' privacy and personal data processing, thus could not be enforced (Cate, 1994, p. 348).

Inspired by the Guidelines, a year later, in 1981, the Council of Europe promulgated a Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (hereafter, the Convention). The Convention codified for the first time basic and comprehensive principles⁵ for data protection to be followed by each EU member state and to enact conforming national laws. Those principles are brief and vaguely worded, including only principles for purpose limitation, access and correction and data quality. For example, the purpose limitation principle here is written-down in such a way that it seems that a specified and legitimate purpose is only needed for the storing of personal data, hence any further use is only tied to the purpose of storing the personal data and not its collection or processing. This version of the principle differs drastically from the OECD's purpose specification. Although, the Conventions' principles did call for broad standards on personal data protection, in comparison to the OECD Guidelines, they are very brief,

6. Openness Principle – There should be a general policy of openness about developments, practices and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.

7. Individual Participation Principle – An individual should have the right: (a) to obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to him; (b) to have communicated to him, data relating to

him within a reasonable time; at a charge, if any, that is not excessive; in a reasonable manner; and in a form that is readily intelligible to him; (c) to be given reasons if a request made under subparagraphs (a) and (b) is denied, and to be able to challenge such denial; and (d) to challenge data relating to him and, if the challenge is successful to have the data erased, rectified, completed or amended.

8. Accountability Principle – A data controller should be accountable for complying with measures which give effect to the principles stated above

⁵ Article 5 – Quality of data

Personal data undergoing automatic processing shall be:

- a) obtained and processed fairly and lawfully;
- b) stored for specified and legitimate purposes and not used in a way incompatible with those purposes;
- c) adequate, relevant and not excessive in relation to the purposes for which they are stored;
- d) accurate and, where necessary, kept up to date;
- e) preserved in a form which permits identification of the data subjects for no longer than is required for the purpose for which those data are stored.

vague and incomplete – they do not cover topics such as transparency and security. Despite being a step in the right direction, the Convention permitted broad variances among national regimes and, most detrimentally, only ten countries ratified it (Cate, 1994, p. 350). Thus, neither the Guidelines nor the Convention brought unitary application among national data protection laws within the EU because they were legally not enforceable.

Working towards a common level of personal data protection in the late '90s, while the ECtHR and CJEU were paving the road for direct enforcement of data protection and privacy rulings, the European Union became the first regulative entity to pass a comprehensive law regulating how personal data should be processed across its internal market - the Data Protection Directive (DPD) (Buttarelli, 2016). The DPD set an international landscape for the protection of individuals with regard to the processing and free movement of personal data by laying down, among others, five principles relating to data quality (Art.1, DPD). DPDs' principles are more extensive than the Conventions' ones, making it clear that personal data cannot be processed without complying with those rules. Although the Directive was exalted for its core principles, its across-member-states implementation was "a frequent subject of criticism and discontent" (Robinson, et al., 2009, p. 38; Cuijpers, et al., 2014, p. 1). The DPD did not affect the evolution of the internet, nor did it prevent 'surveillance becoming the internet's prevailing business model' (Buttarelli, 2016). Due to pressure from (independent) data protection authorities, civil societies, and academia, the European Commission launched, in 2012, a campaign to update the legal instruments which formed the European data protection framework (Buttarelli, 2016). The aim was to make the existing legal framework "more relevant at the age of instantaneous communication, ubiquitous data and potential indefinite storage of those data" and "to make Europe fit for the digital age" (Buttarelli, 2016; EC [4], 2016). As a result, the European Commission, Parliament and Council of Ministers created the General Data Protection Regulation (GDPR) to regulate the processing of personal data to be performed only under strict conditions, improving the flaws of the DPD (EC [7], 2017; De Hert & Gutwirth, 2009, p. 7).

GDPR's stricter conditions include clearer and more specific principles relating to the processing of personal data, codified in Article 5(1). Moreover, GDPR's principles of data processing are more detailed in comparison to the DPD's ones and they resemble to a great extent the fair processing principles from OECD's guideline. Both the GDPR and the OECD include the requirements for fair, lawful, and transparent processing of personal data, followed by the purpose limitation principle, data accuracy, storage limitation, security, and accountability. In comparison, the principles of accountability and transparency were only implicit requirements within the DPD and it was the GDPR which elevated their significance. One principle, however, which was introduced with the DPD and kept at the GDPR, but was neither in the OECD's guideline nor the original US fair information principles, the data minimization principle. Thus, the GDPR has a very strong selection of data protection principles, combining the most restrictive ones from of all sources available.

It can be observed, by comparing the principles defined in the GDPR and the other mentioned documents, that early fair information principles “were broad, aspirational, and included a blend of substantive (e.g., data quality, use limitation) and procedural (e.g., consent, access) principles”, thus reflecting the need for “both individual privacy and the promise of information flows in an increasingly technology-dependent, global society” (Cate, 2006, p. 343). Once translated into the regulatory framework of the European Union and several countries around the world, fair information principles were shaped to become more narrow legalistic principles, reflecting “a procedural approach to maximizing individual control over data rather than individual or societal welfare” (Cate, 2006, p. 343). According to Cate (2006), the more narrow legalistic principles, however, have proven unsuccessful in practice.

Businesses and other data users are burdened with legal obligations while individuals endure an onslaught of notices and opportunities for often limited choice. Notices are frequently meaningless because individuals do not see them or choose to ignore them, they are written in either vague or overly technical language, or they present no meaningful opportunity for individual choice. In short, the control-based system of data protection, with its reliance on narrow, procedural FIPs, is not working. The available evidence suggests that privacy is not better protected. The flurry of notices may give individuals some illusion of enhanced privacy, but the reality is far different. The result is the worst of all worlds: privacy protection is not enhanced, individuals and businesses pay the cost of bureaucratic laws, and we have become so enamoured with notice and choice that we have failed to develop better alternatives. The situation only grows worse as more states and nations develop inconsistent data protection laws with which they attempt to regulate increasingly global information flows (Cate, 2006, p. 344).

Indeed, the GDPR, although trying to cut on red tape and unnecessary obligations, still poses the question of how to comply with certain principles - one of them being Article 6(4) or the compatibility test under the purpose limitation principle. The information obtained from the domains of the principle do not provide insights into how to apply it on practice, however they do offer an insight on its importance and development. Therefore, it is of particular importance that Article 6(4)'s interpretation is clear and uniform in order to secure consistent application among EU nation states and all companies which fall under the scope of the GDPR. In the search of clarity, the next section will discuss the principle's interpretation in a step-by-step fashion as an expert in the domain would do in order to grasp its complete meaning.

2.4 The body of knowledge surrounding Article 6(4) GDPR

The body of knowledge to enable the automation of Article 6(4) GDPR follows the information sources visualized in Figure 3. It aims to frame the complete knowledge surrounding the topic of compatibility of further processing of personal data in a similar fashion as a scholar would do. The starting point is the purpose limitation principle and its components. What do the two parts of the principle entail? How should they be applied in practice? In order to grasp the complete interpretation of the principle, layer by layer we will add additional information which should unveil the explicit knowledge on what are specified, explicit and legitimate purposes with clear distinction on which processing activities would be considered further processing of personal data and which ones will be compatible in accordance with Article 6(4) GDPR.

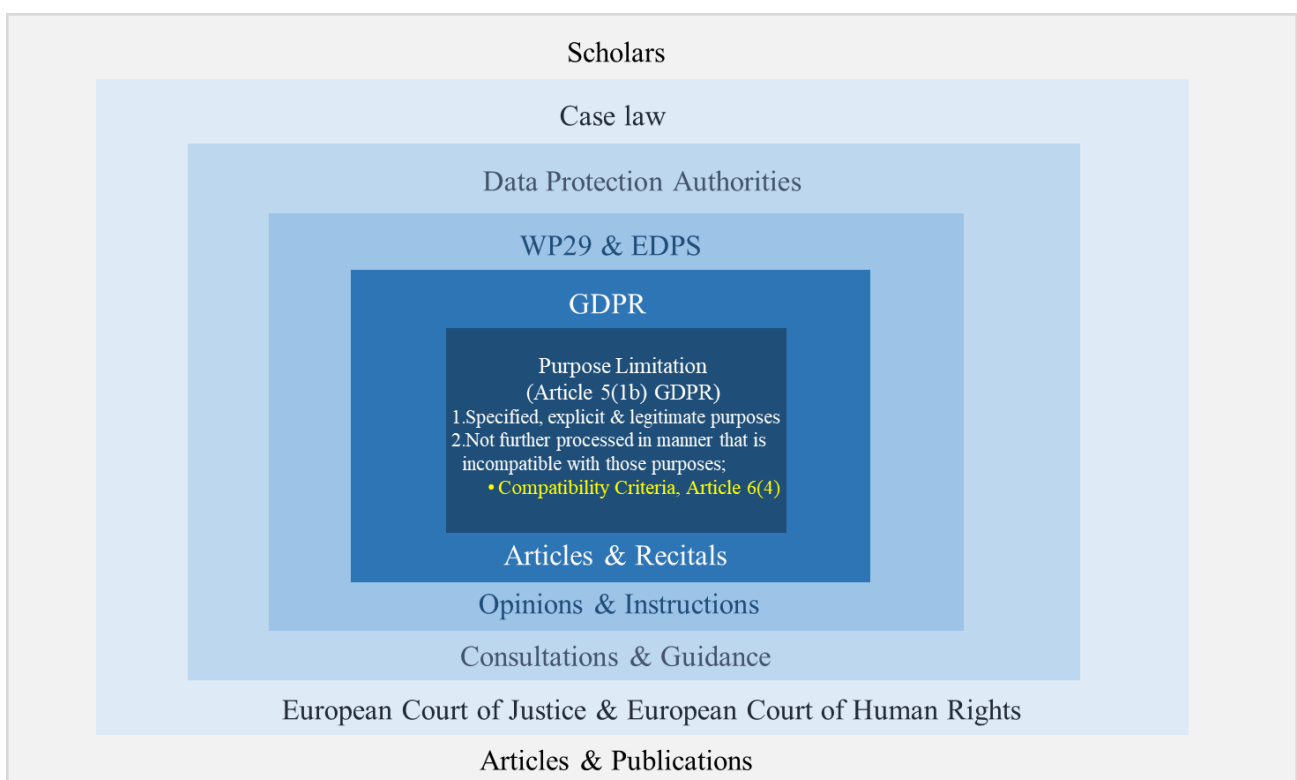


Figure 3: Body of knowledge for the purpose limitation

The first layer is any information provided about the purpose limitation principle and the compatibility criteria within the text of the GDPR. This layer sets the basis of the body of knowledge – the definitions from the GDPR need to be clarified. This should be achieved by adding more information from the next layer – opinions from the advisory body for the DPD & the GDPR, Article 29 Working Party; and from the layer above – guidance from Data Protection Authorities’ guidance. The information to be gained from those sources should give practical advice on how to apply the compatibility criteria in often occurring data controller, data subject cases. To evaluate the usefulness of such opinions and guidance, case law relating to further processing of personal data from the EC(t)HR and the CJEU will be analysed. Last, but not least,

scholars' analysis of the principle and its components will be added to confirm or contradicts the information obtained so far. Overall, those sources and their interpretations and/or guidance on what is compatible further processing of personal data should form a complete body of knowledge regarding this concept.

Along the way of this information gathering, we will map-out the most important findings which then will be used as the basis for the automation of Article 6(4) GDPR.

2.4.1 Purpose limitation within the GDPR

GDPR's purpose limitation principle states that personal data should be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes" (Article 5(1b) GDPR). The predecessor of the GDPR, the Data Protection Directive (DPD), also included this principle using the exact same wording. The GDPR, however, introduced the novelty of four key factors to determine the compatibility of further processing of personal data within Article 6(4). The text of provision 4, Article 6 GDPR reads as follows (emphasis added):

4. Where the processing for a purpose other than that for which the personal data have been collected is not based on the data subject's consent or on a Union or Member State law which constitutes a necessary and proportionate measure in a democratic society to safeguard the objectives referred to in Article 23(1), the controller shall, in order to ascertain whether processing for another purpose is compatible with the purpose for which the personal data are initially collected, take into account, inter alia:

- (f) *any link* between the purposes for which the personal data have been collected and the purposes of the intended further processing;
- (g) the *context* in which the personal data have been collected, in particular regarding the relationship between data subjects and the controller;
- (h) the *nature* of the personal data, in particular whether special categories of personal data are processed, pursuant to Article 9, or whether personal data related to criminal convictions and offences are processed, pursuant to Article 10;
- (i) the *possible consequences* of the intended further processing for data subjects;
- (j) the *existence of appropriate safeguards*, which may include encryption or pseudonymisation.

Provision 4, Article 6 GDPR follows the guidance of Working Party 29 (WP29) in their Opinion 203 on Purpose Limitation, adopted in 2013 under the Data Protection Directive (DPD), after strong criticism on the vagueness and equivocality surrounding the concept of (in)compatible further processing of personal data. Even after the criticism, the key factors of compatibility, as presented by WP29 and Article 6(4), are still defined as an open norm. This means that the data controller might choose to take into consideration additional criteria, but at minimum the decision whether a processing activity is incompatible needs to be based on these five factors (De Hert & Papakonstantinou, 2016, p. 186).

Simply looking at the provisions of the GDPR, one can observe that there is a principle about purpose limitation which consists of two components. The second one includes a set of criteria provided in a separate provision. Figure 4 summarizes this.

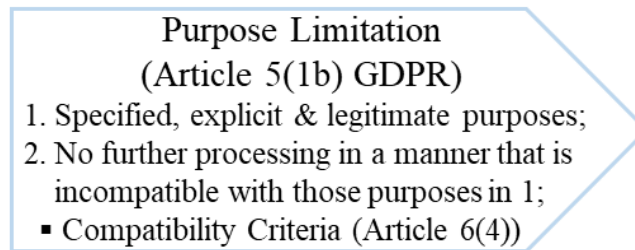


Figure 4: Simple overview of the purpose limitation principle

Additional information about the principle can be obtained from the recitals of the GDPR and its draft versions.

Regarding the first component of the principle (purpose specification) the GDPR lays down the requirement that the data collection purposes need to be specified, explicit and legitimate. However it is not clear from this definition how such purposes should look like. Recital 39 GDPR briefly re-enforces that specified purposes must be “explicit and legitimate and determined at the time of the collection of the personal data”, without providing additional guidance. What is clear from the text of the GDPR is that the purpose limitation principle and the other principles of data processing from Article 5(1)⁶ accept no derogations. This is due to the connection of those principle with the requirements of foreseeability under Article 8(2) of the European Convention on Human Rights (ECHR) and transparency under the Treaty of the European Union (TEU) (Koning, 2015). Although the purpose limitation principle does not accept derogations the first component adheres to that but the second one does not. Hence, there must always be a purpose which is written down and made available to the data subject before or during the collection and processing of the personal data and that purpose should always be specified, explicit and legitimate. However, In contrast, the second component of the purpose limitation principle - the further processing of personal data, includes several exceptions. Such exceptions are explicitly codified within the text of the GDPR, namely Article 6(4) and the Recital 50. The recital offers specific clarity on the additional circumstances to which the compatibility factors are to be applied, including scenarios for which the need to conduct an compatibility assessment are excluded.

⁶ Lawfulness, fairness and transparency; Purpose Limitation; Data Minimization; Accuracy; Storage limitation; Integrity and confidentiality;

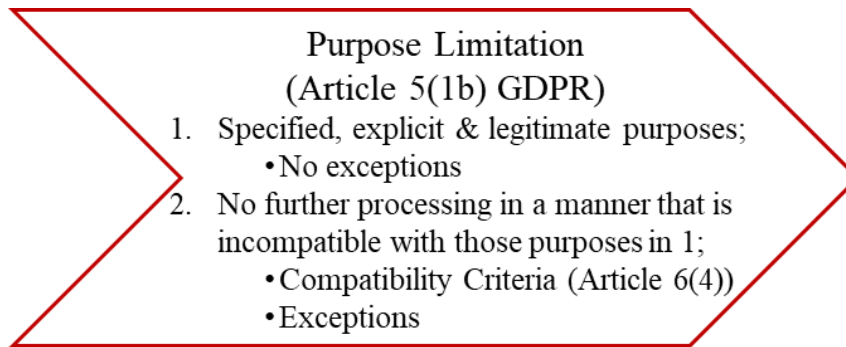


Figure 5: Additional knowledge obtained about the purpose limitation principle

Recital 50 reinforces the principle codified in Article 6(4) – “the processing of personal data for purposes other than those for which the personal data were initially collected should be allowed only where the processing is compatible with the purposes for which the personal data were initially collected”. Therefore, in all compatible cases “no legal basis separate from that which allowed the collection of the personal data is required” (Recital 50 GDPR). The final text of the GDPR does not allow for the further processing of personal data in an incompatible way. This is in contrast with most draft versions, which allowed for the further processing of personal data in an incompatible way if there was a new legal basis applicable to compensate for the incompatibility. Regulator’s decision not to allow such a broad derogation keeps the principle as a credible source of transparency, foreseeability and reasonable expectations by the data subjects.

Furthermore, Recital 50 elaborates on three exceptions to the compatibility assessment, two of which are briefly mentioned in provision 4, Article 6 GDPR. The compatibility assessment is required, unless the new purpose:

- is necessary for “the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller, Union or Member State law may determine and specify the task and purposes for which the further processing should be regarded as compatible and lawful”;
- is a subject for “archival purposes in the public interest, scientific or historical research purposes or statistical purposes” or
- has been consented to by the data subject (Recital 50, GDPR).

Recital 50, also briefly instructs that for cases of possible criminal acts or threats to public security, data processing to “a competent authority should be regarded as being in the legitimate interest pursued by the controller”, if in compliance with a legal, professional or other binding obligation of secrecy. Last but not least, Recital 50 specified that “the application of the principles set out in this Regulation and in particular the information of the data subject on those other purposes and on his or her rights including the right to object, should be ensured”.

The information obtained from the provisions of the GDPR relevant to the purpose limitation principle is a good start for the complete and explicit body of knowledge. Specifically, valuable insights were obtained about the structure and exceptions of the purpose limitation principle. An overview is presented in Figure 6.

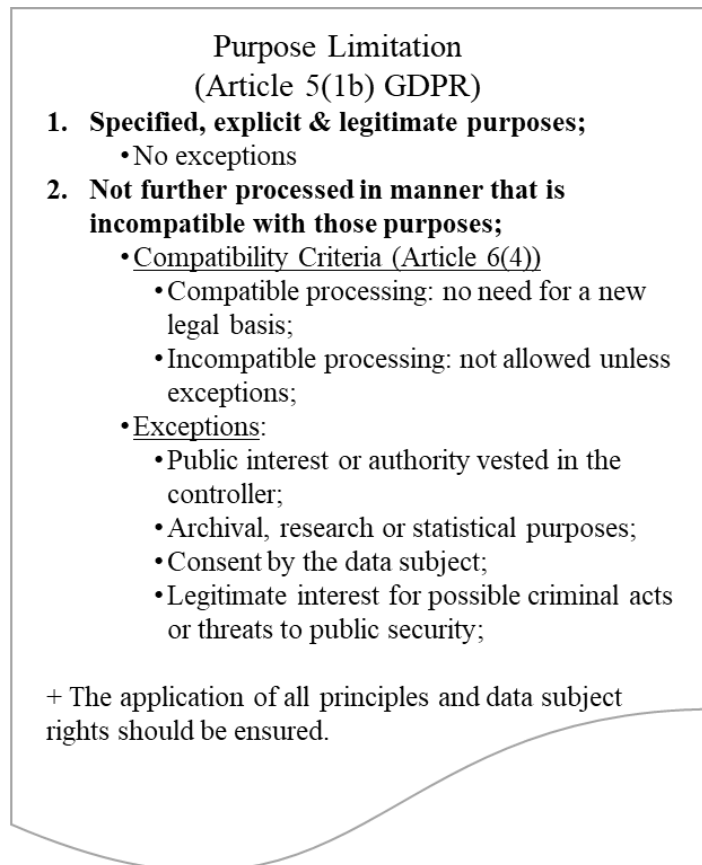


Figure 6: Overview of the information obtained about the purpose limitation after an analysis of GDPR's text

Nevertheless, this overview does not provide an answer yet to the question of what entails compatible processing of personal data. Additional knowledge is needed in order to grasp the practical application of the principle and its building blocks. The next section will provide detailed insights about the principle from the opinions of WP29.

2.4.2 WP29's opinion on the purpose limitation principle

As previously explained, the wording of the purpose limitation principle was left unchanged during the transition between the Data Protection Directive (DPD) and the General Data Protection Regulation (GDPR) (WP29 203, 2013). An analysis of the purpose limitation principle, under the DPD, was provided by 'Working Party on the Protection of Individuals with regards to the Processing of Personal data' (WP29) in its advisory role to ensure a clear and consistent application of the principles codified within the DPD and the back-then draft GDPR. Although, WP29 is comprised of representatives from all EU Data Protection Authorities, the European Data Protection Supervisor (EDPS) and the European Commission, the opinions issued by WP29 reflect only the views of the body itself, and they do not reflect the position of the European Commission (EC[9], 2016). As a result, the Court of Justice of the European Union (CJEU) seems to be ignoring the interpretations given by WP29 (Bird&Bird, 2016). Nevertheless, any opinions issued by WP29 are highly influential among Data Protection Authorities and data controllers because WP29 is the most comprehensive and specific source of guidance on how the provisions of the DPD and the GDPR are to be read and applied (Robinson, et al., 2009, p. 9). Specifically, WP29's opinion on the purpose limitation is an important source of information for the interpretation of the principle, because the text of Article 6(4) GDPR it is based on this opinion.

WP29's opinion on purpose limitation follows the structure of the principle itself, hence first clarifying what specified, explicit and legitimate purposes are, and then giving guidance on what incompatible is and what are the requirement to prove compatibility. Thus, at the upcoming sections we will discuss the guidance given by WP29 and how it contributes to the body of knowledge which may enable the automation of Article 6(4) GDPR.

2.4.2.1 *Specified, explicit and legitimate*

Specified purposes

Specified is a purpose "clearly and specifically identified", "detailed enough to determine what kind of processing is and is not included within the specified purpose" (WP29 203, 2013, p. 15). A specified purpose enables the assessment of:

- the compliance with an applicable law,
- the relevance of data protection safeguards applied, and
- the scope of the processing operation.

This explanation, however, does not enable the reader to be able to draw a line on what is a specified purpose. WP29 further explains that a specific purpose should not be vague or general, neither too detailed nor overly legalistic. The specified nature of a purpose should give the data subject, and anyone reading the

purpose, an opportunity to detect a separate process for which his/her personal data is collected, including the positioning of this process within both the legal and technical safeguards. Hence, a data collection process may include more than one purpose – the so-called “related purposes” and or “separate purposes”. The concepts of a related and separate purposes is tightly connected to the reasonable expectations of the data subjects and are usually explained using examples.

When registering for a shopping account at Zalando I gave them my email address, among other personal data, for the following purposes explicitly mentioned on Zalando’s website:

- Create a customer account;
- Access the customer account with information about purchases and use;
- Confirm the receipt and the sending of any orders;
- In general communicate with the customer;

Those purposes are related. Together they serve the general purpose of providing me with a service. All four purposes fit under the same legal ground of performance of a contract, thus supposedly also using the same technical standards to protect my personal data. However, not only that there is no specific, explicit mentioning of sharing my personal data with 3rd parties but and I have not separately consented to it. Thus I had, and still have, no reasonable expectations that my email address would be uploaded to Facebook on the basis of those purposes specified by Zalando.

However, if there was to be a separate line stating “sharing of your personal data with 3rd parties for marketing reasons”, this would have been a separate purpose (because it has nothing to do with enabling me to use Zalando’s shopping website) and it would require a separate legal ground (because it is not part of the performance of the contract).

Overall, the information gathered about what is a specified purpose does not provide an exact definition. Instead, a number of indications are obtained as summarized in Figure 7. It seems like an abstract process to apply the indications ex ante, however ex post seems to be easier. Using Zalando’s purposes as an example, it can be argued that they are not too detailed or legalistic, but are clear enough to understand that the legal ground applicable (performance of a contract), the scope of the processing operations and protection safeguards, and to enable certain reasonable expectations. However, if the purposes are to be drafted from scratch, the indications summarized in Figure 7 are not specific enough to make sure that indeed a purpose would be ‘specified’.

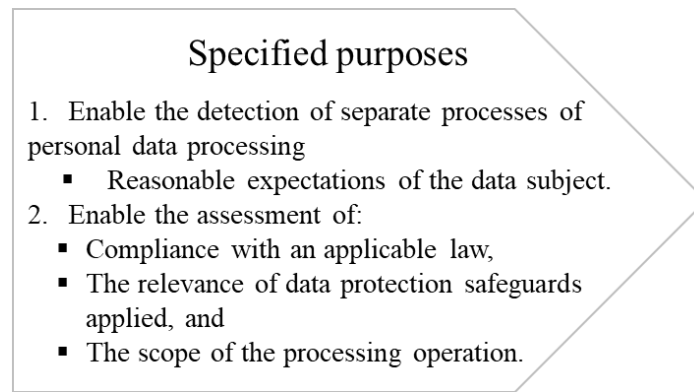


Figure 7: Minimum requirements to be met for a purpose to be specified

Explicit purposes

Explicit is a purpose “sufficiently unambiguous and clearly expressed” in some intelligible form. An explicit purpose is a specific purpose, ensuring no vagueness or ambiguity as to its meaning or intent. “What is meant must be clear and should leave no doubt or difficulty in understanding”, leading to a common understanding of how the data can be used, irrespective of who is the reader – data controller(s), data subject, third party processors or the data protection authorities (WP29 203, 2013, p. 17). Therefore, this requirement enables transparency and predictability. It “reduces the risk that the data subjects' expectations will differ from the expectations of the controller” (WP29 203, 2013, p. 17). The purposes must be expressed and explained in some form available to the data subject, whether it will be in writing or orally.

Moreover, WP29 clarifies that this requirement is distinct from the requirement of information to be given to the data subject, although the two are closely related and both serve to enable transparency. Overall, in the context of further processing of personal data, an explicit and specific purpose would provide “proof of the original purpose and allow comparison with subsequent processing purposes” (WP29 203, 2013, p. 18).

To put this definition into practice, Working Party 29 has given the following phrases as examples of purposes of processing which are neither sufficiently clear, nor explicit (WP260 rev.01, 2018, p. 9).

- *“We may use your personal data to develop new services”;*
- *“We may use your personal data for research purposes;*
- *“We may use your personal data to offer personalized services”.*

The common ground of unspecificity among those purposes is that language qualifiers such as “may”, “might”, “some”, “often” and “possible” should be avoided, unless the data controllers can demonstrate the need of such language and how it will not undermine the fairness of the personal data processing. Particularly, regarding the first bullet of the examples it is unclear what the services are or how the personal

data collected will help to develop those services. The second bullet provides no explanation for what kind of research the personal data will be used. The third bullet does not clarify what personalization entails.

If the data controller does not specify the purposes of the processing in sufficient detail, or in a clear and unambiguous language, or the information provided may not correspond to the facts of the case, or it could contain inconsistencies about the purpose, then the reasonable expectations of the data subjects cannot be met and the data controller cannot process the personal data for any purposes at its discretion. Where the purposes are specified inconsistently or the specified purposes do not correspond to reality, the data controller or a Data Protection Authority should take into account all factual elements, as well as the common understanding and reasonable expectations of the data subjects based on such facts, in order to determine the actual purposes (WP29 203, 2013, p. 18).

Accordingly, WP29 also provides the correct alternatives for those three incorrect purposes:

- *“We will retain your shopping history and use details of the products you have previously purchased to make suggestions to you for other products which we believe you will also be interested in”;*
- *“We will retain and evaluate information on your recent visits to our website and how you move around different sections of our website for analytics purposes to understand how people use our website so that we can make it more intuitive”;*
- *“We will keep a record of the articles on our website that you have clicked on and use that information to target advertising on this website to you that is relevant to your interests, which we have identified based on articles you have read”;*

It can be clearly observed from the corrected purposes that they are very affirmative and give a clear understanding on what types of data will be processed, how the personal data will be processed and what the result towards the data subject will be (WP260 rev.01, 2018, p. 9). Nevertheless, it should be noted that as for the guidance on ‘specified’, the guidance on ‘explicit’ becomes tangible only on the light of the examples provided by WP29. Hence, the indications summarised in Figure 8 become explicit only in the light of ex post examples.

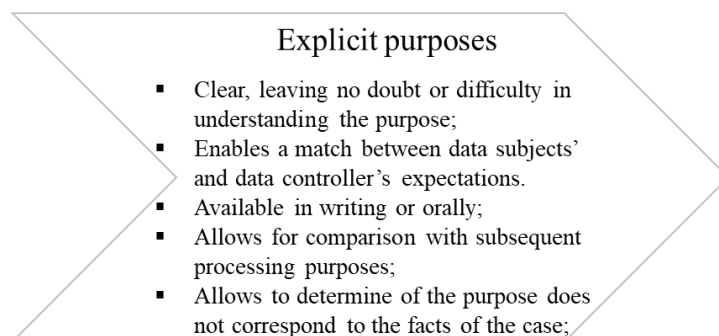


Figure 8: Minimum requirements to be met for a purpose to be explicit

Legitimate purposes

Legitimate is a purpose which is at “all different stages and at all times” based on at least one of the legal grounds provided for in Article 6(1)⁷ GDPR (WP29 203, 2013, p. 19). Moreover, legitimate should be interpreted as “in accordance with the law” in the broadest sense, which extends to, firstly other areas of law including “all forms of written and common law, primary and secondary legislation, municipal decrees, judicial precedents, constitutional principles, fundamental rights”; secondly “[w]ithin the confines of law, other elements” such as customs, codes of conduct, codes of ethics, and contractual arrangements; and thirdly the “general context and facts” of each case, including “the nature of the underlying relationship between the controller and the data subjects” (WP29 203, 2013, p. 20). Figure 9 summarizes what legitimate purposes consist of.

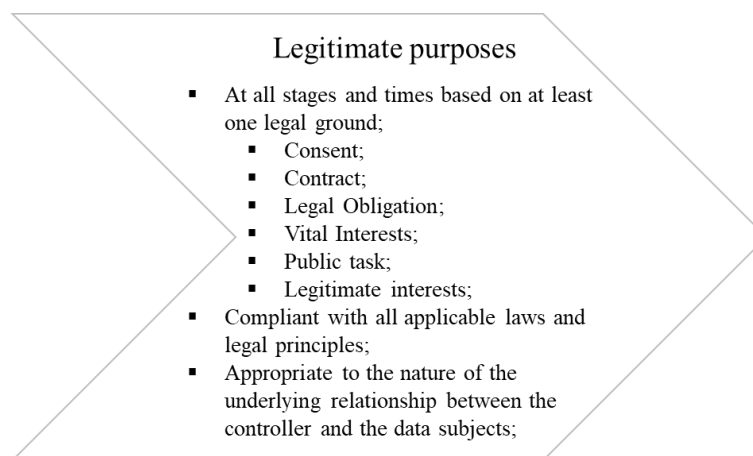


Figure 9: Minimum requirements to be met for a purpose to be legitimate

The burden of proof, for compliance with the legitimate requirements mentioned above, is on the data controller. Not only that data controllers need to comply with the applicable laws and legal principles, but they also need to communicate this to the data subjects. The condition for a legitimate processing enables data subjects to enforce their rights against the data controllers. However, depending on the type of legal

⁷ 1.Processing shall be lawful only if and to the extent that at least one of the following applies:

- (a) the data subject has given consent to the processing of his or her personal data for one or more specific purposes;
 - (b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;
 - (c) processing is necessary for compliance with a legal obligation to which the controller is subject;
 - (d) processing is necessary in order to protect the vital interests of the data subject or of another natural person;
 - (e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;
 - (f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.
- Point (f) of the first subparagraph shall not apply to processing carried out by public authorities in the performance of their tasks.


ground (e.g. consent, contract, legal obligation) and the relationship between the data controller and subjects (e.g. employment relationship, tax obligations) the enforcement of some data subject rights might be limited. Therefore, it is of particular importance to data subjects that the applicable legal grounds are clearly communicated for both related and separate purposes. Multiple related purposes are bound together by a single legal ground. In contrast, any separate purposes, which will use the same personal data collected, would be based on a different legal ground.


This can be illustrated with an example. In the context of an employment relationship between the data subject and the data controller there are multiple related purposes to collect the personal data of the employee - the creation of a contract, payment of a salary, pension-related activities, etc. For all those related purposes the legal grounds are performance of a contract and compliance with legal obligations. There are also separate purposes which are not part of the performance of the employment contract and to which a new additional contract needs to be formed – if the employee has a right to a leased car, the car contract does not influence in any way the employment relationship, but the employee needs to enter into this new contract for that specific purpose. Hence, at the moment of collection of the personal data (when the contract is being formed), the data subject (employee) is presented with several related purposes for the employment contract and a separate group of related purposes for the lease car contract. It could be, however, quite difficult for the employee to determine, from all purposes presented, which are the further purposes and which are not.

According to WP29, the need to distinguish between the different types of purposes (related and specific) is of a specific importance because it helps to avoid an incompatible processing of personal data. An incompatible processing “cannot be remedied simply by adopting a new legal ground” - the “processing of personal data in a way incompatible with the purposes specified at collection is unlawful and therefore not permitted” (WP29 203, 2013, p. 36). However, Working Party 29 pointed out that an initial purpose can, nevertheless, change and that change can be compensated by the adoption of a new legal ground:

“... in some situations, after assessment of all relevant factors, including the availability of safeguards and/or the availability of an appropriate new legal basis to compensate for the change of purpose, the controller may find that further processing for a changed purpose can comply both with the compatibility requirement and the requirement of a legal ground under [Article 6 GDPR].”

Overall, a further purpose will be compatible when it is connected to the initial purpose(s) and is based on a new legal ground, in order to compensate for the change. Figure 10 summarizes this.

- 
 At the moment of collection or prior to it:
 - **Separate purposes** ⇔ Require separate legal grounds
 - **Related purposes** ⇔ Require a common legal ground

- 
 After collection:
 - **Further purposes**
 - Incompatible further processing *cannot* be remedied by adopting a new legal ground;
 - Compatible further processing *can* have a new legal ground to compensate for the change;


- 
 Distinguishing between the different types of purposes helps to avoid an incompatible processing of personal data.

Figure 10: Overview of the types of purposes and legal grounds

Although this information is very useful, it is still unclear what is further processing of personal data and what is the connection between the applicable legal grounds and the change of purpose.

2.4.2.2 Further processing of personal data

According to WP29, the Data Protection Directive's notion of compatibility does not specifically refer to processing for the 'originally specified purposes' and processing for 'purposes defined subsequently'. Instead, it differentiates between:

... the very first processing operation, which is collection, and all other subsequent processing operations (including for instance the very first typical processing operation following collection - the storage of data) (WP29 203, 2013, p. 21)

Hence, WP29 offers a very broad reading on what constitutes a further processing of personal data:

... any processing following collection, whether for the purposes initially specified or for any additional purposes, must be considered 'further processing' and must thus meet the requirement of compatibility (WP29 203, 2013, p. 21)

This interpretation is also laid down at the GDPR. Article 6(4) states that the compatibility assessment applies to purposes 'other than that for which the personal data have been collected'.

Such a broad and enveloping definition of further processing of personal data seems to be a very practical solution to an otherwise impossible task. For example, according to the purpose specification, the initial purposes defined, at the moment of collection or prior to it, also include implicit purposes determined using the reasonable expectations of the data subjects. Hence, if there is no strict and clear distinction between initial and further purposes, it would be rather impossible to determine where one ends and the other starts.

Nevertheless, such a broad definition raises a different type of inquiry. Namely, are the related and separate purposes compatible, as long as they comply with the purpose specification requirements or do they need to pass a compatibility assessment? If the answer is affirmative, then the question of where related and separate purposes end and where other further processes start emerges (again). There is no guidance on this issue from WP29, instead they point out that the legislator introduced “a double negation: it prohibited incompatibility” (WP29 203, 2013, p. 21):

By providing that any further processing is authorised as long as it is not incompatible (and if the requirements of lawfulness are simultaneously also fulfilled), the legislators intended to give some flexibility with regard to further use. Such further use may fit closely with the initial purpose or be different. The fact that the further processing is for a different purpose does not necessarily mean that it is automatically incompatible: this needs to be assessed on a case-by-case basis (WP29 203, 2013, p. 21).

The question is still pending. Does this mean that a compatibility assessment needs be performed for every processing activity after collection? Working Party 29 does not explicitly state that. Instead, they distinguish three categories of further processing of personal data:

- obviously compatible;
- compatibility is not obvious and needs additional analysis;
- obviously incompatible;

A closer look on what each of those categories entails, follows.

Obviously (in)compatible

According to WP29, an obviously (in)compatible purpose of further processing should be easily identified by comparing it to the initial purposes defined before or at the moment of data collection. Further processing would be compatible when the personal data are processed specifically to achieve the purposes clearly specified at collection, in a way customary to achieve those purposes so that the processing activity can clearly meet the reasonable expectations of the data subjects “even if not all details were fully expressed at the start” (WP29 203, 2013, p. 22). In contrast, if the personal data are to be further processed in an incompatible way it would be for additional purposes that a reasonable person would find unexpected, inappropriate or otherwise objectionable. Thus, the processing would clearly not meet the expectations of a reasonable person. This would include, personal data collected for commercial purposes to be further processed for anti-terrorism purposes using secret algorithms, without an appropriate lawful ground and without transparency towards the data subjects.

From this explanation it follows that an assessment needs to be performed for each further processing of personal data. This answers the question raised above in the affirmative, and it brings for a new inquiry. How should the obviousness assessment look like? Obviously, it is not a simple string matching. To this problem, WP29 proposes the use of a formal assessment – a comparison of all initial purposes defined by the data controller with any further purposes, in order to find out whether the further ones are covered (explicitly or implicitly) (WP29 203, 2013, pp. 21-23). Unfortunately, WP29 provides no additional guidance on the application of this formal method, besides the strong emphasis on the reasonable expectations of the data subject as a main criterion.

WP29 does not provide for an objective solution. Instead, it relies on concepts which need to be evaluated by “a reasonable person in the situation of the data subject” and may vary since there is no common method of evaluation. Even, if WP29’s words are to be followed that “[o]nly in marginal cases of doubt, would further analysis be useful”, the question what ‘marginal cases of doubt’ are emerges (WP29 203, 2013, pp. 21-23).

Compatibility is not obvious and needs further analysis

Where there might be a ‘connection’ between the initial purposes and the way the personal data are to be further processed but the (in)compatibility is not obvious, a multi-factor analysis is needed to determine the outcome. The analysis can differ in its depth - “the greater the distance between the initial purpose specified at collection and the purposes of further use, the more thorough and comprehensive the analysis will have to be” (WP29 203, 2013, p. 24).

In contrast to the obvious cases which use a formal assessment, the multi-factor analysis should be based on a substantive assessment. Such an assessment goes beyond the formal approach of identifying the original purposes, the further purpose and comparing them in light of the reasonable expectations of the data subject. Instead it takes into account the way the purposes should be understood from their context and whether a number of additional criteria are adopted to compensate for the change of purpose (WP29 203, 2013, p. 26).

The substantive assessment makes use of the of four key factors defined by WP29 within their opinion 203 on the purpose limitation principle. The aim of those factors is to enable a pragmatic approach using 'rules of thumb' based on “what a reasonable person would find acceptable under any given circumstances” (WP29 203, 2013, p. 49). WP29 developed these factors based on Member States’ specific legal provisions and practice. An overview is present at the left side of Table 1. Most importantly, these factors are the foundation of the text of Article 6(4) GDPR, on the right side of Table 1. The difference between the two sets of factors is marginal, as presented in the table. Therefore, for the purposes of this research we will presume that guidance about the interpretations of one set of factors applies to the other one and vice versa.

Table 1: Overview compatibility factors WP29 and GDPR

WP29	Article 6(4)
a) the <i>relationship</i> between the purposes for which the data have been collected and the purposes of further processing	a) any <i>link</i> between the purposes for which the personal data was collected and the purposes of intended further processing
b) the <i>context</i> in which the data have been collected and the <i>reasonable expectations</i> of the data subjects as to their further use	b) the <i>context</i> of the data collection and the <i>relationship</i> between the data subject and the controller
c) the <i>nature</i> of the data and the <i>impact</i> of the further processing on the data subjects	c) the <i>nature</i> of the personal data d) the possible <i>consequences</i> of the processing towards the data subjects
d) the <i>safeguards</i> applied by the controller to ensure fair processing and to prevent any undue impact on the data subjects	e) the existence of appropriate <i>safeguards</i>

The five factors (based on Article 6(4) GDPR) give an actual structure to the compatibility assessment, in comparison to the formal assessment. They are, nevertheless, open-ended which leaves the assessment susceptible to different interpretations. In order to elucidate the logic applied and to understand how to use the factors in practice, WP29 provided twenty-two practical examples. Those examples enable the shaping of a hunch about the type of questions one ought to ask and the way in which one ought to think when assessing the compatibility of a further purpose.

The table below summarizes the sub-factors, which indicate either compatibility or incompatibility, based on the examples of WP29. Each sub-factor has been rewritten as a statement, such as ‘the further purpose is not part of the initial purposes’; and it can be answered either affirmatively or in disagreement, thus allowing for an objective evaluation of any measures taken to compensate for the change of purpose. In the opinion of the author, those explicit sub-factors have the potential to make the compatibility assessment not only easier to be applied but and more robust.

Table 2: An overview of Article 6(4)'s compatibility factors and their corresponding sub-factors extracted from WP29's examples

Substantive factor	Indicates compatibility	Indicated incompatibility
A. any link between the purposes for which the personal data was collected and the purposes of	The further purpose is implied from the initial purposes. For example: - The further purpose is the next logical step in the processing	The further purpose is not implied from the initial purposes. For example: - The further purpose is not part of the initial purposes.

<p>intended further processing</p>	<p>according to the purposes.</p> <ul style="list-style-type: none"> - The further purpose is within the reasonable expectations of the data subjects. - The further purpose is commonly understood in the same way by relevant stakeholders; <p>The further purpose is intended for research and afterwards not processed for any other purpose.</p>	<ul style="list-style-type: none"> - The further purpose is unexpected, inappropriate or otherwise objectionable. - The further purpose is not commonly understood in the same way by relevant stakeholders; <p>The further purpose is intended for research purposes but the research outcome will be used for commercial gains.</p>
<p>B. the context of the data collection and the relationship between the data subject and the controller</p>	<p>The context of the further use is the same as the collection context. For example:</p> <ul style="list-style-type: none"> - Context of collection is commercial service, the further purpose is also for commercial purposes. - Context of collection is professional service (medical, legal), the further purpose is also for professional services. - Context of collection is compliance with a legal obligation, the further purpose is also for compliance with the legal obligation. <p>Within the context of data collection the data subject was informed about the purposes, legal grounds, how the personal data will be used and any consequences. For example:</p> <ul style="list-style-type: none"> - An understandable and easy to find privacy statement. - Clarity on the methods used to process the personal data. <p>There is a balance of power between the data subject and controller. For example:</p> <ul style="list-style-type: none"> - The data subject is more or equally powerful to the data controller. - The data controller does not process personal data in a secret or vague way. - The personal data processing is not bound to a professional secrecy. <p>The further processing of the personal data is based on a valid legal ground. For example:</p>	<p>The context of the further use is different from the collection context. For example:</p> <ul style="list-style-type: none"> - Context of collection is commercial service, the further purpose is counter terrorism. - Context of collection is professional service (medical, legal), the further purpose is for commercial purposes. - Context of collection is compliance with a legal obligation, is based on the legal ground of legitimate interests of the data controller. <p>Within the context of data collection the data subject was not informed about the purposes, legal grounds, how the personal data will be used and any consequences. For example:</p> <ul style="list-style-type: none"> - No transparency about the collection of the personal data, neither for the (potential) further purposes. - No clarity on the methods used to process the personal data. <p>There is an imbalance of power between the data subject and controller. For example:</p> <ul style="list-style-type: none"> - The data controller is more powerful than the data subject. - The data controller processes personal data in a secret or vague way. - The personal data processing is bound to a professional secrecy. <p>The processing of the personal data is not based on a valid legal ground. For example:</p>

	<ul style="list-style-type: none"> - When the further purpose is a subject to a legal obligation, statutory responsibility, public task or formal role of the controller, those always adhere to the principle of legal certainty, transparency, necessity and proportionality. - When the applicable legal ground is contract, for the data subject is relatively easy to terminate the contract. - When the applicable legal ground is consent, the consent is both freely given and informed. 	<ul style="list-style-type: none"> - There are no legal obligations or other legal grounds to enable the data collection. - There are legal obligations to enable the further processing, but they do not constitute a necessary and proportionate measure in a democratic society. - When the applicable legal ground is contract, for the data subject is not easy to terminate the contract. - When the applicable legal ground is consent, the consent is not freely given.
C. the nature of the personal data	<p>The personal data processed does not require special protection.</p> <p>The further processing envisages using a part of the personal data collected.</p> <p>Combining large and different data sets of personal data in a foreseeable and transparent manner.</p> <p>The further processing will use alternative methods which are less intrusive or do not involve personal data processing.</p> <p>The methods used to process the personal data are explained in detail to the data subject.</p>	<p>The personal data processed is sensitive and/or it requires special protection.</p> <p>The further processing envisages using all personal data collected.</p> <p>Without any foreseeability at the time of collection, combining large and different data sets of personal data.</p> <p>The further processing will use new methods to analyse data, without appropriate mitigating measures and quality check of any results.</p> <p>The methods used to process the personal data are secret or not clearly communicated to the data subject.</p>
D. the possible consequences of the processing towards the data subjects	<p>The further processing will be conducted by the same data controller.</p> <p>The further purpose is conducted for the benefit of the data subject or the public in general.</p> <p>The consequences of the further processing are foreseeable and are communicated clearly to the data subject</p> <p>The personal data which will be further processed would be available to only a limited amount of people.</p>	<p>The further processing will be conducted by a different controller.</p> <p>No sensitive data will be processed for the further purpose, but the impact will be sensitive.</p> <p>Unknown consequences and unforeseen consequences (exclusion, discrimination, emotional impact).</p> <p>The personal data will be publicly disclosed and/or made accessible to a large number of persons.</p>
E. the existence of appropriate	Additional measures are applied by the controller to serve as a compensation	The data controller does not take any additional measures to compensate for

safeguards	<p>for the change of purpose. For example:</p> <ul style="list-style-type: none"> - The personal data to be further processed is a subject of partial or full anonymization, pseudonymization and/or aggregation of the data - The methods used to process the data do not involve automated decision making and there are no risks of wrong assumptions being made. - The methods used to process the data involve automated decision making but the data subject has given a valid consent for it. - The data subject has been informed and is given the possibility to object to processing. - Where the further processing is based on consent, it must be really freely given, specific and informed. - The further processing is based on either contractual safeguards regulating the transfer of the personal data between the parties or other formal arrangements. - The initial and the further purposes both comply with the storage limitation principle; 	<p>the change of purpose. For example:</p> <ul style="list-style-type: none"> - The personal data to be further processed is not anonymized or aggregated. - The methods used to process the data involve automated decision making which is not communicated to the data subjects. - The further processing intends to use data analytics algorithm to analyse personal data and the results of the analysis may be biased, but the data subject is not made aware of that. - The data subjects are not informed about the methods of further processing, thus not given an opportunity to object. - The further processing is not based on a valid consent. - The relationship between the data controllers for the further processing is unclear and there are no contractual safeguards applicable. - The personal data processed (either for collection or further purposes) do not adhere to the storage limitation principle
------------	---	--

Besides the formation of the explicit statements, another observation, based on the practical examples, is that the multi-factor assessment does not follow the order in which the factors are laid down. Instead, all examples start by discussing the factual situation and then, on the basis of that, a major part of the explicit statements which indicate (in)compatibility are answered in a random order. Hence, the compatibility assessment does not have a method within which the outcome should be determined. Moreover, the outcome will depend on the assessment as a whole, because there could be deficiencies at certain points (e.g. unexpected further processing of mobile phone location data for speeding prevention) which can be compensated by adequate measures at other points (e.g. effective anonymization of the personal data or freely given, informed and specific consent).

Overall, it can be observed that the substantive assessment relies heavily on the facts of the case. Plausibly, that is the reason why WP29 states that the outcome needs to be assessed on a case-by-case basis. Nevertheless, the explicit statements inferred out of the opinion on the purpose limitation principle are not facts based, and help to establish some ‘rules of thumb’ thus making the assessment less susceptible to different interpretations.

2.4.2.3 Exceptions

The way in which WP29 treats the exceptions to the compatibility rule is by incorporating them into the multi-factor assessment. Each of the exceptions can be perceived as an additional layer of appropriate safeguards. There are exceptions available for further purposes which would be considered incompatible. Namely, a public interest or authority vested in the controller, archival/research, consent, and reporting possible criminal acts or threats to public security towards a competent authority. However, only the ‘correct’ version of each exception will allow for the outcome to be labelled as compatible. An explanation follows.

For the first exception, further purposes necessary for the performance of a task carried out in the public interest are only applicable to data controllers which have a formal role, a statutory responsibility or a legal obligation of safeguarding public interests. Hence, this exception is only applicable to a limited amount of data controllers. For example, supermarkets who further process the personal data of their ‘unhealthy’ customers as part of a public health initiative promoted by the local government will not be exempted under the public interest exception. This is because supermarkets do not have any formal role in safeguarding public health. Similarly, any processing for the exercise of official authority vested in the controller, Union or Member State law, must meet the requirements set in Article 23(1) GDPR. This entails that any authority and/or legal obligation needs to be a necessary and proportionate measure in a democratic society and it safeguards one of the purposes of Article 23(1)⁸.

For the second exception, when a further purpose is a subject for “archival purposes in the public interest, scientific or historical research purposes or statistical purposes” it needs to have any of the following additional appropriate safeguards:

- The personal data will not be used to support measures or decisions regarding any particular individual;

⁸ (a) national security;
(b) defence;
(c) public security;
(d) the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security;
(e) other important objectives of general public interest of the Union or of a Member State, in particular an important economic or financial interest of the Union or of a Member State, including monetary, budgetary and taxation matters, public health and social security;
(f) the protection of judicial independence and judicial proceedings;
(g) the prevention, investigation, detection and prosecution of breaches of ethics for regulated professions;
(h) a monitoring, inspection or regulatory function connected, even occasionally, to the exercise of official authority in the cases referred to in points (a) to (e) and (g);
(i) the protection of the data subject or the rights and freedoms of others;
(j) the enforcement of civil law claims.

- Subject to professional codes of conduct;
- Where possible, anonymization or pseudo-anonymization;
- Restriction on access;
- The publication of the research results is in an aggregated and/or fully anonymized form;
- Functional separation between participants in the research and outside stakeholders, especially in the case of different data controllers⁹;
- Perform a Data Protection Impact Assessment¹⁰

For the third exception, further purposes which can be consented to by the data subject should meet all requirements for consent. Hence, consent should be explicit, specific and freely given. Especially for health data, data about children, other vulnerable individuals, or other highly sensitive information, since such data can only be processed on the legal basis of data subject's consent. In order for a consent to be explicit, specific and freely given, it needs to be to a real choice, without a power imbalance¹¹ between the data subject and the controller, and without any negative consequences towards the data subject.

For the last but not least exception, in the cases of possible criminal acts or threats to public security, data processing to a competent authority should be in line with any legal, professional or other binding obligations of secrecy. Of particular importance is that the data subjects are accordingly informed and/or can reasonably expect the further use of their personal data for such purposes.

Overall, if any of the exceptions does not meet the additional safeguards, they will not convert an incompatible further purpose into a compatible one.

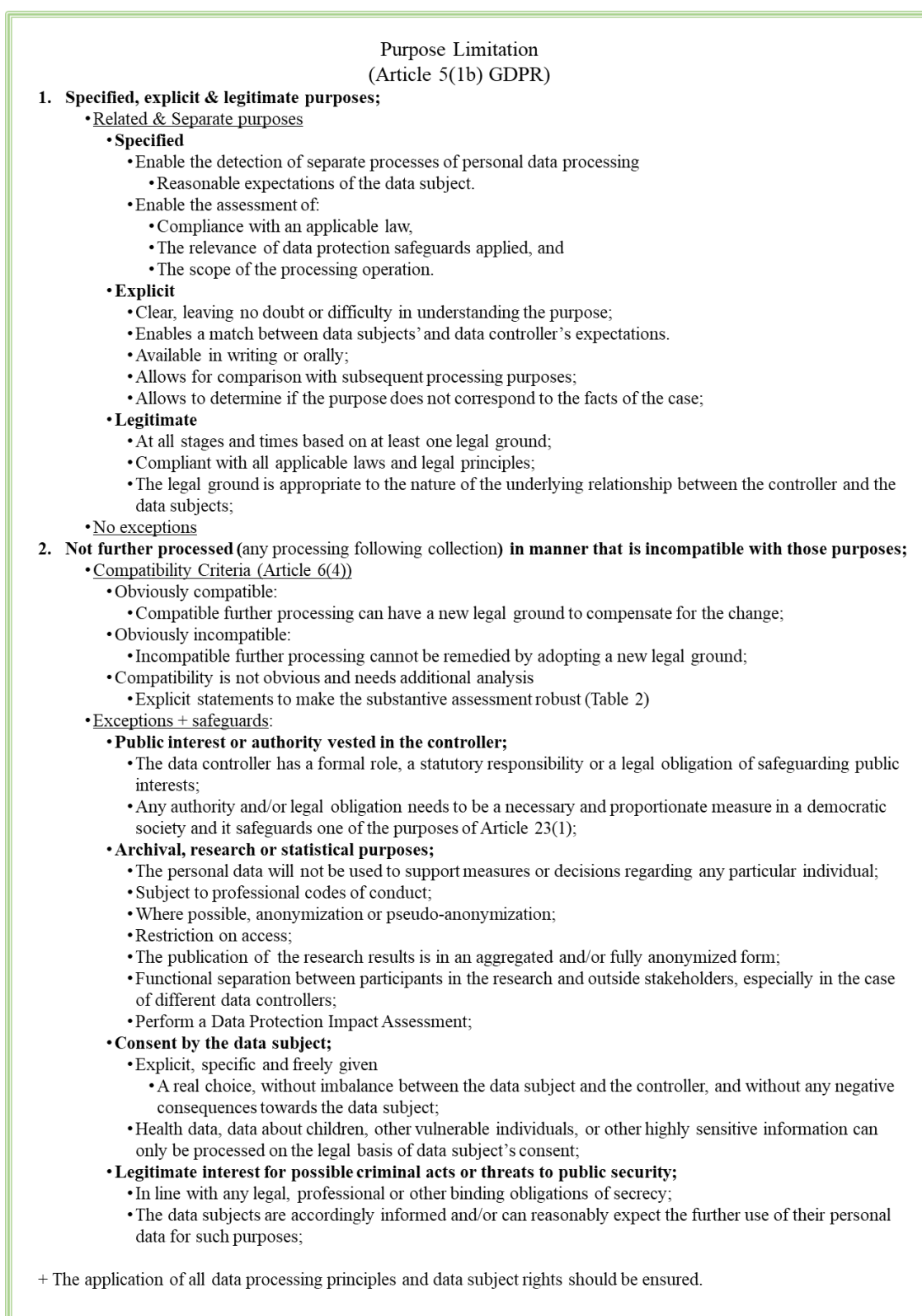
In conclusion, this section gathered a substantial amount of information about the purpose limitation principle and the interpretation of its components from the opinions of Working Party 29. An overview of the information gathered thus far is presented in Figure 11.

⁹ E.g. Cambridge Analytica's research of personality traits and the further use of the results from political parties;

¹⁰ According to the ICO.

¹¹ A power imbalance between a data controller and a data subject entails that the data controller has more power on deciding how and why the personal data will be used. In some occasions the data subject may not even be aware that their personal data is being collected and processed by one or more data processors.

Figure 11: An overview of the information obtained about the purpose limitation principle



Despite the amount of insights gained about the two components of the purpose limitation principle, additional information is still needed. Concepts, such as the formal assessment of obvious further purposes,

specific and explicit, and the possibility to compensate with safeguards for the change of purpose which would be otherwise be incompatible (while an incompatible processing is unlawful), require further guidance and interpretation in order to become explicit.

Such additional guidance might be available beyond the text of the GDPR and WP29's opinions. Namely, local guidance from the data protection authorities. This will be the focus of the next section.

2.4.3 Data Protection Authorities' guidance and consultations

Each EU member state has a data protection authority (DPA). DPAs supervise “through investigative and corrective powers, the application of the data protection law” (EC, 2018). It is within the powers of DPAs to provide advice on data protection issues and handle complaints regarding violations of the GPPR and the relevant national laws, within the private sector (EC, 2018). For the EU institutions and bodies which process personal data, the European Data Protection Supervisor (EDPS) ensures compliance with the GDPR. The EDPS also supervises & advises on all aspects of personal data processing and related policies and legislation, handles complaints and conducts inquiries (europa.eu, 2018).

From the guidance and consultations published by DPAs and the EDPS, valuable information can be obtained about the application of the purpose limitation principle. This section will present a selection of decisions by the EDPS and several countries' DPAs, which will illustrate how those bodies tackle compliance with the purpose limitation principle and its components.

2.4.3.1 Country specific guidance

Country specific guidance can be of two types. First, each Data Protection Authority can publish additional guidance on how to comply with the provisions of the GDPR and second, they can provide advice and consultations on specific real-life cases.

Regarding the former, most DPAs cite and refer to the opinion of WP29 on purpose limitation, without any additional requirements or guidance on how to apply the substantive multi-factor assessment. The Information Commissioner's Office¹² (ICO) however, published a discussion paper on ‘Big Data, Artificial Intelligence, Machine Learning, and Data Protection’, in which it made a number of observations about the purpose limitation principle and it provided guidance on how to determine whether a further processing of data, within the domain of big data and machine learning, is compatible.

¹² The Information Commissioner's Office (ICO) is the independent regulatory office (national data protection authority) dealing with the Data Protection Act 1998 and the Privacy and Electronic Communications Regulations 2003 across the UK; and the Freedom of Information Act 2000 and the Environmental Information Regulations 2004 in England, Wales and Northern Ireland and, to a limited extent, in Scotland.

ICO & Fair compatibility

The ICO (2017, p. 38) stated that “a key factor in determining whether big data analysis is incompatible with the original processing purpose” is fairness. Fairness, according to the ICO (2017, pp. 38-39), has three components: “[first] [t]ransparency – what information people have about the processing, [second] the effects of the processing on individuals, and [third] their expectations as to how their data will be used”:

In particular, this means considering how the new purpose affects the privacy of the individuals concerned and whether it is within their reasonable expectations that their data could be used in this way. This is also reflected in the GDPR, which says that in assessing compatibility it is necessary to take account of any link between the original and the new processing, the reasonable expectations of the data subjects, the nature of the data, the consequences of the further processing and the existence of safeguards. (ICO, 2017, p. 38)

Although ICO’s guidance is clearly intended for further big-data-analytics processing, the choice of fairness as the ‘key factor’ is an interesting one for several reasons.

First, by connecting fairness and the question of ‘how the new purpose [will] affect the privacy of the individual concerned’, the authority made an explicit link between the privacy of the data subject and his/her data protection rights. As discussed earlier, privacy and data protection are concepts which are not black and white but they intertwine, which is the reason why the EU charter recognized both the respect for private and family life (Article 7) and the protection of personal data (Article 8). However, this explicit link when it comes to application of a data protection concept is unusual. After all, within the complete text of the GDPR, privacy is not mentioned even once.

Second, ICO’s definition of fairness does not differentiate itself from GDPR’s principles of data processing. Both fairness and transparency are the first requirements to be met when personal data is processed (Article 5(1) GDPR). Although the ICO recognizes that the importance of fairness is preserved at Article 5(1)(a) it does not assume that any further data-analytics processing will first comply with requirements for fair, transparent and lawful processing and then the purpose limitation. Instead, the ICO explicitly makes fairness part of the substantive multi-factor assessment for further processing of purposes related to big data analytics, artificial intelligence and machine learning domains. Such an approach seems to be at odds with any interpretation of the fair principle by the WP29 or the CJEU.

Third, this guidance does not add more clarity or explicitness to the factors defined in Article 6(4). To the contrary, it adds an more requirements without any explanation on how are those to be applied.

The domains of big data and machine learning are very challenging to the protection of individual’s data. However, in order to mitigate the potential risks emerging from the use of personal data in unknown ways,

the substantive assessment for further processing of personal data under Article 6(4) should be made less challenging and not the opposite. Other DPAs address the issues of processing for any research with personal data, irrespective if will include big-data analytics or not, to be a subject of a Data Protection Impact Assessment (DPIA). In that way no additional factors are assigned to the already existing ones in Article 6(4) GDPR.

Specific cases

Several data protection authorities actively discussed specific cases of further processing of personal data and provide for public decisions on whether the processing was compatible or not. In some cases where incompatibility was detected, the DPAs issued fines to the violators. DPAs often reaffirm what has been observed so far that the two components of the purpose limitation principle go hand-in hand. A brief overview of such cases follows.

The French Data Protection Authority, the CNIL, found that Facebook violated the purpose limitation principle by not clearly explaining to its users that their personal data are systematically collected when they navigate on third-party websites which include social plug-ins (CNIL, 2017). Although, the cookie banners of Facebook did mention that information is collected "on and outside Facebook" they did not allow users to clearly understand what this entails (CNIL, 2017). Furthermore, the CNIL decided that Facebook had violated the purpose limitation principle by further processing personal data of its users by combining all their information to display targeted advertising. The CNIL pointed out that for this specific purpose there is no legal basis which could justify the change of purpose (CNIL, 2017). Similar observations were made by the Belgian Privacy Commission and the Dutch Data Protection Authority (Dutch DPA). In particular, the Dutch DPA found out that Facebook further processed sensitive personal data, such as user's sexual preferences, without explicit consent in order to show targeted advertisements (CNIL, 2017).

Another investigation by the Dutch DPA elucidated a similar violation of the purpose limitation principle (Autoriteit Persoonsgegevens, 2017). Microsoft claimed to have processed customers' data in order to "fix errors, to keep devices up-to-date and secure and to improve its own products and services" (Autoriteit Persoonsgegevens, 2017). This statement by itself is very broad and does not meet the specific and explicit requirements of the purpose specification. Moreover, in practice Microsoft would collect more data - in multiple ways and from multiple sources, in a manner unpredictable to the user. Users could opt-out for some purposes, if they would be aware of them and knew how to do so, however that was not clearly communicated. Hence, the collection purposes did not meet the legality requirement of the purpose specification. The data collection went beyond what is necessary to perform the contract of service and it could not meet the freely given and informed requirements of consent. Additionally, Microsoft further processed customers' data to show personalized advertisements in Windows 10, the browser Edge, all apps for sale in the Windows store, and other apps. For this further purpose also Microsoft did not have a valid

ground - consent would not be considered freely given and informed in such cases and any of the other legal grounds was not applicable either.

Using personal data for more purposes than communicated to the data subject is a common topic of criticism by DPAs. A German state office for Data Protection advised all manufacturers and users of data warehousing and data mining not to create 'general data warehouses' where all collected data, both personal and non-personal, are used as training sets to run machine learning algorithms. Such use of the data, if completely separate from data's collection purposes, will be noncompliant with the purpose limitation principle (LDI, 2000). The use of data mining and machine learning algorithms to search for new, meaningful insights, to create and combine data poses risks towards the fundamental right to informational self-determination, thus potentially resulting in unknown personality profiles, automated predictions of behaviour and long-term storage without a valid legal ground. The German state office explicitly advised that for such practices, consent would not compensate for the change of purposes because consent for indefinite and unlimited purposes would be invalid.

The Data Protection Commissioner of Ireland reaffirmed that it would be unlawful to collect information about people routinely and indiscriminately. A data controller must always have a sound, clear and legitimate purpose for collecting any personal data (Data Protection Commissioner, 2008). For example, where personal data stored about a bank card is collected for the purposes of a transaction, it can be assumed that the purpose for its collection ends following completion of the payment for a product or service. Thus, personal data obtained from a bank card for a particular transaction should not be used subsequently for other transactions without express consent to do so. If the customer has clearly opted in (as opposed to not having opted-out) to their data being retained for future transactions, this would permit further processing e.g. if a customer has consented to this purpose (Data Protection Commissioner, 2008). Another recommendation is that personal images captured on CCTV cameras by a data controller, where the CCTV was in operation solely for security purposes, should not be used by the data controller for any other purpose, especially for staff monitoring. Furthermore, telephone service providers, upon termination of a subscription, should not continue processing the personal data of the data subject, unless the user actively consented to that (Data Protection Commissioner, 2008).

Together these cases provide important insights into a number of incompatible cases. Although, the guidance from the ICO and the consultations from the other DPAs did not provide for additional clarity on how to apply the four factor assessment under Article 6(4) GDPR, a number of specific patterns can be commonly observed.

A common ground of criticism is the lack of adherence to the purpose specification and the lack of a valid legal ground for each further purpose. Figure 12 summarizes the collection purposes which are unspecific and inexplicit to comply with the purpose specification.

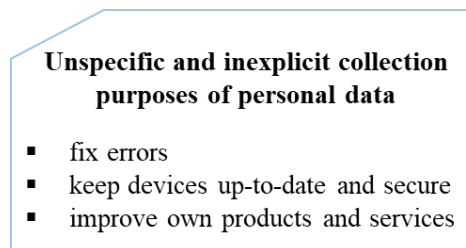


Figure 12: Purposes which would not comply with the purpose specification

In contrast, a common ground of advice to data controllers is to put in place appropriate procedures and security measures to ensure that personal data obtained for one purpose may not be accessed and used for another purpose without an assessment of (in)compatibility. Figure 13 summarizes the observations made from DPAs guidance.

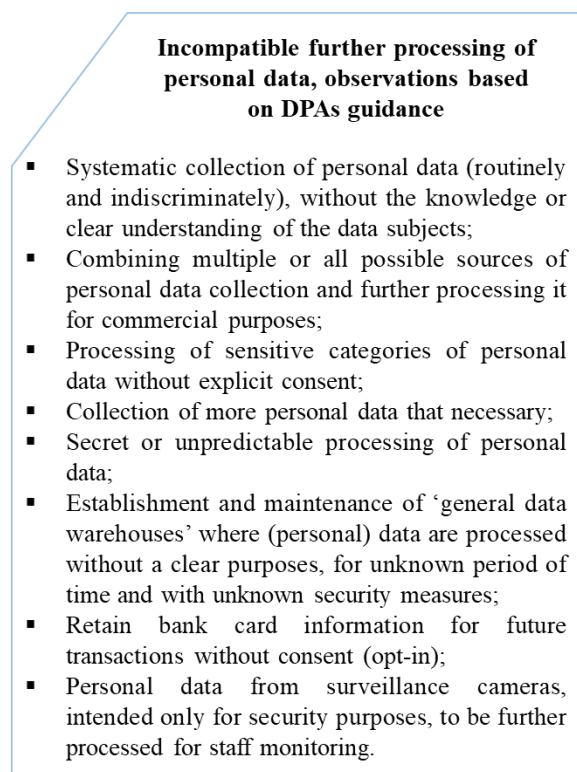


Figure 13: Cases considered incompatible by local DPAs

2.4.3.2 Public bodies' guidance

As already mentioned, for EU institutions and bodies which process personal data, including legislation specifying the purposes of personal data processing, the European Data Protection Supervisor (EDPS) ensures compliance with the GDPR. The EDPS issues opinions, which depending on the nature of the task are public or not. From the publicly available opinions, few touch upon the importance of the purpose limitation principle, including further processing operations which would be considered as compatible. A brief overview, of two cases from the EDPS and one from the joint Supervisory Body of Europol, follows.

In its opinion on the Proposal for a Council Regulation 2006/C 242/14, the EDPS advised on the right wording of the provision so that it would comply with the purpose limitation principle. Regulation 2006/C 242/14 allows national authorities and courts to process personal data for the specific purposes of facilitating the enforcement of maintenance claims. In its opinion, the EDPS requires that the purposes, for which creditors' personal data will be processed, must be precisely and explicitly defined, in accordance with the purpose specification. For example, Article 44 of Regulation 2006/C 242/14 defines the "specific purposes for which information shall be provided by national administrations and authorities to the relevant central authorities: [...] to locate the debtor". The EDPS was on the opinion that such a purpose should be more complete and precise - 'locate the debtor' should be more specific. For example, it should include a location with a certain degree of stability such as debtor's address. Moreover, it should be specified that the use of GPS data to locate the debtor should be excluded as an option. In that way, the kinds of personal data that might be processed according to this proposal would be circumscribed. The EDPS also went to advise on a specific case of compatible further processing. The purposes of 'exercise of official authority' and 'protection of the data subjects or of the rights and freedoms of others' would be considered compatible, if they are absolutely necessary and based on proportionate legislative measures.

The EDPS also advised on the compatibility of further processing of personal data under Regulation (EC) No 45/2001 (on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data). The question was whether personal data originating from an access security system or a time management system can be used for investigative purpose, such as to instruct a disciplinary process (EDPS, 2013). If, under the application of a specific legal instrument, there are rules which govern the disciplinary procedures and fraud investigations and those rules allow for the use of specific types of data in the context of disciplinary investigations, then the processing would be compatible. That being said, according to the EDPS, the authorization must be understood restrictively. When a disciplinary process is launched it should be clear and open to objections, and for a specific case (misconduct). Moreover, the further processing for the disciplinary process must be proportional and necessary to the purpose, and it should use only personal data which are relevant and adequate to the case. Overall, any further systematic or structured use of access-security or time-management

data in the context of administrative enquiries or disciplinary investigations would need to be based on a specific internal rule in compliance with Article 6.1 of Regulation (EC) No 45/2001¹³.

Another exception to the purpose limitation principle was granted in Opinion 06/22 of the joint Supervisory Body of Europol. In it was decided that law enforcement authorities and Europol can have access and use visa information system (VIS) data for purposes other than the ones for which the data was collected. According to Europol, such an exception could be justified under certain conditions. Namely, no routine access. Any access is limited to specific cases or for Europol to a specific task. Moreover, there should be sufficient control over the use of any VIS data. Last but not least, this exception should be proportionate and substantially contributing to the purpose or case identified. Only then, it would be compatible.

From those opinions it becomes evident that public bodies need to adhere to the purpose limitation principle, just as any private data controller. Each opinion presents a purpose, or a set of purposes, which will be considered as compatible only when they include a number of appropriate safeguards. In comparison to advices targeted to private controllers, the EDPS aims to make it explicitly clear in advance which further processing would be compatible or not. This enables transparency and creates legal certainty.

Overall, those opinions add no additional clarity on the application or interpretation of Article 6(4) GDPR. However, they do point out several appropriate safeguards which valuable input for Table 2 – an overview at Figure 14.

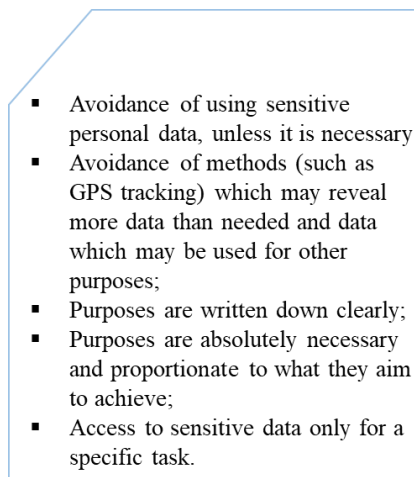
- 
- Avoidance of using sensitive personal data, unless it is necessary;
 - Avoidance of methods (such as GPS tracking) which may reveal more data than needed and data which may be used for other purposes;
 - Purposes are written down clearly;
 - Purposes are absolutely necessary and proportionate to what they aim to achieve;
 - Access to sensitive data only for a specific task.

Figure 14: Appropriate safeguards inferred from EDPS's opinions

¹³ Regulation 45/2001, Article 6, Change of purpose: Without prejudice to Articles 4, 5 and 10:

1. Personal data shall only be processed for purposes other than those for which they have been collected if the change of purpose is expressly permitted by the internal rules of the Community institution or body.
2. Personal data collected exclusively for ensuring the security or the control of the processing systems or operations shall not be used for any other purpose, with the exception of the prevention, investigation, detection and prosecution of serious criminal offences.

2.4.4 Case law

“Data protection legislation originating in the DPD or GDPR is only a fraction of the law” relevant for the further processing of personal data (Koops & Leenes, 2014, p. 5). An essential source of guidance and interpretation of data protection statutes is case law. In the domain of data protection, two courts - the Court of Justice of the European Union (CJEU) and the European Court of Human Rights (ECtHR) - have the greatest influence on the interpretation and application of data protection’s principles.

The CJEU strives to achieve harmonization of national legislation and to rule on the legality of EU laws, while the ECtHR provides a minimum human rights protection (Butti, 2013). “If the function of the CJEU is to help build unity ... the [one of the] ECtHR is to help build a community” (Claire, 2011, p. 1433). The effects of the rulings of the two courts are different. When a national piece of legislation is found to be in violation of the ECHR, the court’s decisions are advisory, hence, non-binding. The ECtHR has no jurisdiction to annul domestic laws or administrative practices which violate the Convention. It can only advise on repealing or amending them. The CJEU, in contrast, refers to the EU principles of supremacy, direct effect and state liability, hence imposing that national legislation, found violating the EU laws, is going to be changed or annulled (Butti, 2013). The two courts are not connected, but they share essentially identical provisions for privacy and data protection. Therefore, in order to secure legal certainty the two courts often cross-reference each other, contributing to the creation of a "uniform human rights standard" (Butti, 2013).

2.4.4.1 ECtHR

The right to privacy was for the first time explicitly introduced in EU legislation with the changes brought by EU Charter of Fundamental Rights. Before that privacy used to be primarily protected under Article 8 (Right to respect for private and family life) of the European Convention on Human Rights (ECHR). ECtHR was the first international institution available for individuals to challenge national legislation for violating the private life of individuals. Although, the European Court on Human Rights (ECtHR) has never explicitly acknowledged a general right to protection of personal data, it has recognized aspects of the data protection doctrine under Article 8 ECHR. In particular, Article 8 includes within the concept of private life, aspects relating to personal identity, which is interpreted to include how and why personal data is processed, what are the (potential) results of the processing and does it influence an individual’s rights and freedoms.

In the context of the purpose limitation principle, the European Court of Human Rights (ECtHR) has addressed specific cases of ‘re-use’ of personal data. In its judgments, the ECtHR would not mention the purpose limitation principle directly, nor would it make a distinction between the purpose of collection and further purposes. Instead, the court would look at specific cases of processing of personal data and decide whether they are in a violation of Article 8 or not. For example, the ECtHR decided that the copying of

documents containing banking data and their subsequent storage by local authorities (*M.N. and Others v. San Marino*) is allowed by Article 8, when it is justified by a purpose, which is specific and adhered to, without collecting more than what is necessary. In contrast, at another example, the ECtHR found that Article 8 would be violated when one's personal life is intruded upon by a systematic surveillance or transfer of personal data with the intention to realize negative actions against that individual. In *Luordo v. Italy* the court decided that the unrestricted monitoring of one's correspondence, although permitted by local bankruptcy law, violated Article 8. Furthermore, in *Rotaru v. Romania*, the ECtHR expanded the scope of private life to include injuries to an applicant's reputation if they were caused by a systematic collection and storing of 'false' personal data (data relating to another person with the same name). Another interference with Article 8 was an obligation towards private companies to provide tax auditors with access to individuals' personal data without a concrete and specific reason (*Bernh Larsen Holding AS and Others v. Norway*, § 106).

From those selected examples it can be inferred that the ECtHR would often look at a specific processing case at hand (based on a legal ground of legal obligation), analyse what the purpose of the data processing was, were there any exceptions or safeguards to justify the processing and if there were none - declare the processing to be in a breach of individual's right to respect for private and family life. ECtHR case law does not contribute to the explicit understanding of the purpose limitation principle. This has to do with the fact that the court has never addressed the purpose limitation principle per se and has not made a distinction between collection purposes and further processing purposes. Moreover, ECtHR cases would judge at processing operations based on the legal ground of compliance with a legal obligation, which is only one of the six possible legal grounds available under Article 6(1). Hence, those judgements offer only a one-sided view on how to deal with the 're-use' of personal data. Nevertheless, such cases do help to get a better understanding of which processing purposes would violate the ECHR. Overall, it can be concluded that processing operations which collect more personal data than needed, in a systematic way and the results of which are negative towards the individual, would be considered in violation of the right to respect for private and family life.

2.4.4.2 CJEU case law

The CJEU has been extremely influential in the domain of data protection. It has interpreted multiple principles of the DPD, although it has never directly accessed the applicability of the purpose limitation principle. Neither, similarly to the ECtHR, does it make a distinction between collection and further processing of personal data. Instead, the CJEU would assess the (in)compatibility of a certain processing of personal data, heavily relying on the facts of the case, by the applicability of other provisions of the DPD. An overview follows.

(In)compatibility assessed through the applicability of the right to be forgotten

Probably the most important case tackling the (in)compatibility of a processing of personal data is case C-131/12 (Google Spain SL, Google Inc. v AEPD, Mario Costeja González). The court decided that data subject's rights under the DPD (in particular, Article 12b and Article 14a) enable an individual to have search engine results about him altered even though the information was true and lawfully published by third parties. The importance of this case lays within first, the acknowledgement that a data-subject-right of erasure could be exercised only if a personal data's processing is incompatible with Article 6 (principles of data quality) and 7 (making processing legitimate) DPD, and second, the definition of incompatibility provided by CJEU.

The court's analysis was very systematic. First, it was established that the activity of a search engine, finding information on the internet and indexing it, can be classified as processing of personal data within the material scope defined in Article 2(b) DPD. Second, the court decided that the operator of a search engine can be regarded as a data controller because it determines which results to be shown and it stores them temporarily. Thus, any processing of personal data by the search engine must comply with the principles relating to data quality set out in Article 6 DPD and must meet one of the criteria for making the processing of personal data legitimate as listed in Article 7 DPD. Up until this point, CJEU's analysis was very detailed and it inspected the concepts addressed – processing of personal data and being a data controller. Nevertheless, with the statement that Articles 6 and 7 must be complied with, the analysis of DPD's provisions stopped. The court did not analyse whether those two articles are complied with. Instead it continued to address the processing in light of the facts of the case.

According to the facts, the search engine (further) processed personal data collected lawfully in the context of a legal action. A Spanish newspaper published the details of a legal claim, including data subject's personal data, on the basis of an order by Spain's Ministry of Labour and Social Affairs. In contrast, the (further) processing of the personal data by the search engine was based on the legal ground of legitimate interest. Fifteen years after the publishing of the newspaper article, although there were no outstanding claims against the data subject anymore, the newspaper could not remove the article, thus the data subject opposed to the indexing performed by the search engine as being prejudicial to him and his fundamental rights. The CJEU accepted the case and recognized that such an objection would be only possible if the personal data was processed in an incompatible way.

The court did not analyse whether the processing in question is compatible with Articles 6 and 7 DPD. Nor did it follow WP29's substantive assessment of compatibility. Instead, the CJEU relied on the relationship between the principles of data processing and data subject rights: "the question whether the processing complies with Articles 6 and 7(f) [DPD] [...] may be determined in the context of a request as provided for

in Article 12(b)” (C-131/12, 2014). Article 12(b) DPD, the right to have a rectification, erasure or blocking of one’s personal data is only applicable when the processing of the personal data in question is incompatible. Hence, the court has to address what incompatibility is, which it did by presenting a definition:

... incompatibility may result not only from the fact that such data are inaccurate but, in particular, also from the fact that they are inadequate, irrelevant or excessive in relation to the purposes of the processing, that they are not kept up to date, or that they are kept for longer than is necessary unless they are required to be kept for historical, statistical or scientific purposes. (C-131/12, 2014).

Furthermore, the CJEU went to clarify that an “initially lawful processing of accurate data may, in the course of time, become incompatible where the data are no longer necessary in the light of the purposes for which they were collected or processed” (C-131/12, 2014). This foundation laying set the scene for the analysis of whether the facts of the case fit the definition of incompatibility.

For this specific case, the personal data, at that specific point in time (fifteen years later), were no longer relevant. The (further) processing made the irrelevant data ubiquitous online, which infringed upon data subject’s right to privacy. Moreover, the person, whose data was processed did not have a pronounced public role hence there was no interest of the general public to keep those search results (C-131/12, 2014). Thus, the processing, was found incompatible. It must, nevertheless, be acknowledged that if some facts were to be different, such as if the role played by the data subject in the public life was to be more pronounced so that an interference with the subject’s fundamental rights would be justified by the overriding interest of the general public; or if the search engine had not made the results accessible to anyone, the outcome of the case would have been different.

This case is the only one at which the CJEU explicitly mentioned and defined the incompatibility of a processing of personal data. Interestingly enough, the definition of the incompatibility of a processing is tightly connected to the principle of accuracy. Hence, personal data which violates the principle of accuracy would be considered incompatible with the purpose limitation principle. Furthermore, of particular importance is also the time-factor of the processing: an initially lawful processing may become incompatible where the data are no longer necessary to the collection or processing purposes. The CJEU did not elaborate much on how to determine such a necessity, besides discussing few factors such as an overriding public interest, the amount of time passed and the reach of the processing (ubiquitous). This ruling clearly demonstrates the difference, between the CJEU and WP29, in the logic of assessing (in)compatibility. The CJEU explicitly did not follow the distinction between collection and any further processing, and the substantive assessment introduced by WP29. Instead, the court preferred to address the concept of

incompatibility through data subject rights. This case is not the only one at which the court decides in such a way. The next case demonstrates this.

(In)compatibility assessed through the applicability of information to be given to the data subject

In case C-201/14 (Bara case), the CJEU decided that a processing of personal data was incompatible when the data subject was not informed about it. In this case, the CJEU, similarly to the previous one, did not assess whether the purpose limitation principle and the requirement for legitimate processing under Article 7 DPD conditions were fulfilled, despite recognizing their importance. Instead the court assessed the compatibility of a data transfer between the two public authorities in the light of Articles 10 and 11 (information to be given to the data subject) and whether an exemption to these rights was applicable under Article 13 DPD.

In order to determine whether “persons earning income through self-employment qualify as insured persons”, a transfer of personal data between two public bodies was made obligatory by a local member-state provision. That provision, however, was never publicized, thus unknown to anyone else but the public bodies. Moreover, the transfer included more personal data than it was necessary, which clearly violated the data minimization principle. Last but not least, the local public authorities, both the one collecting and transferring the data and the one receiving it, failed to inform the data subject of those processing operations, and there were no exceptions under Article 13(1)(e) and (f) DPD to exclude the data controllers from their obligation to provide information. As a result, CJEU concluded that the transfer did not comply with neither the conditions laid down in Article 10 and 11, nor in Article 13.

The court did not explicitly state that the transfer was incompatible, although that might have been a more efficient route to determine the incompatibility of the processing, instead of discussing the violations of Article 10, 11 and 13. If the court was to first assess that Article 6 and 7 of the DPD were not met (the personal data transfer was not lawful, both the purpose limitation principle and the data minimization principles were not met) then it would not be needed to analyse the violation of the information to be given to the data subject when no exceptions apply. Although in this and the previous case the CJEU did not first analysed the applicability of Articles 6 and 7 DPD, but instead analysed the processing in light of the data subject rights involved, in the next case, it did focus firstly on the applicability of Article 7 DPD.

Compatible processing based on legitimate interest

In Case C-13/16 (Valsts policijas Rīgas reģiona pārvaldes Kārtības policijas pārvalde v Rīgas pašvaldības SIA "Rīgas satiksme"), the CJEU pointed out that Article 7(f) DPD allows for a strictly necessary (further) processing of personal data in order to realize a third party's legitimate interests (Case C-13/16, 2016).

More specifically, the CJEU stated that “there is no doubt” that the interest of a third party to obtain “the personal information of a person who damaged their property in order to sue that person for damages can be qualified as a legitimate interest”. The personal data in question were collected for the purpose of realizing a civil claim. The (further) processing consisted of a data transfer to the injured party. That transfer included merely the first name and the surname of the person who caused the damage, not allowing for the precise identification of that person in order to bring an action against him. The (further) processing was limited to what was strictly necessary because the third party did not have, under local law, a legal ground to receive all details about the person who caused the damage, but it did have a legitimate interest to receive some of the personal data¹⁴. Moreover, the legitimate interest would have been fully satisfied (with all personal data needed to bring an action against the person who damaged their property) under local administrative law, the right of which was not exercised by the third party.

Therefore, such a limited, in compliance with local law, (further) processing of personal data to realize a third party’s legitimate interest is in line with Article 7(f), hence compatible. In this case the CJEU did not explicitly mention that this processing was compliant with the purpose limitation and data minimization principles, however it did point out within the analysis of the facts that the collection had a clear and specified purpose of collection, that the further purpose also had such a purpose, only the minimal amount of data was transferred and there was a new legal ground to compensate for the purpose change. This type of analysis is similar enough to resemble the substantive assessment under WP29’s opinion. This case was the only one at which CJEU was vaguely similar to the substantive assessment of WP29. Although the purpose

¹⁴ The legitimate interest to bring an action against a person who damaged their property does not allow for the sharing of the

As the taxi driver was initially held responsible for that accident, Rīgas satiksme sought compensation from the insurance company covering the civil liability of the owner and lawful user of the taxi. However, that insurance company informed Rīgas satiksme that it would not pay Rīgas satiksme any compensation on the basis that the accident had occurred due to the conduct of the passenger in that taxi, rather than the driver. It stated that Rīgas satiksme could bring civil proceedings against that passenger.

14 Rīgas satiksme then applied to the national police asking it to provide information concerning the person on whom an administrative penalty had been imposed following the accident, to provide copies of the statements given by the taxi driver and the passenger on the circumstances of the accident, and to indicate the first name and surname, identity document number, and address of the taxi passenger. Rīgas satiksme indicated to the national police that the information requested would be used only for the purpose of bringing civil proceedings.

15 The national police responded by granting Rīgas satiksme’s request in part, namely by providing the first name and surname of the taxi passenger but refusing to provide the identity document number and address of that person. Nor did it send Rīgas satiksme the statements given by the persons involved in the accident.

16 The decision of the national police was based on the fact that documents in the case file in administrative proceedings leading to sanctions may be provided only to the parties to those proceedings. Rīgas satiksme is not a party to the case at issue. Under the Latvian Administrative Infringements Code, a person may at his express request be given the status of victim in administrative proceedings leading to sanctions by the body or official responsible for examining the case. In the present case Rīgas satiksme did not exercise that right.

limitation principle here was not explicitly discussed it becomes evident that if the principles of data processing and the criteria for making the processing of personal data legitimate are met, then a further processing operation would be compatible.

The next case further elucidates this observation, but this time in light of the proportionality principle.

(In)compatibility assessed through the applicability of the proportionality principle

The CJEU declared Directive 2006/24/EC (Data Retention Directive) to be invalid in Joined Cases C-293/12 and C-594/12. The Data Retention Directive required EU member states to store citizens' telecommunications data in order for police and security agencies to be able to request such data under the general interest of fight against serious crime and public security. The CJEU invalidated the Directive because the EU-legislature had exceeded the limits of the principle of proportionality by allowing for the wide-ranging and particularly serious interference with data subjects' fundamental rights (CJEU, 2014).

In particular, the CJEU pointed out three arguments why the purposes defined in the Data Retention Directive did not comply with the proportionality principle. Firstly, the directive's scope was extremely broad – “all individuals, all means of electronic communication and all traffic data without any differentiation, limitation or exception” (CJEU, 2014). Secondly, the directive failed to lay down any objective criteria for the competent national authorities to gain access to the data retained and use them. Access to the data was not made dependent on the prior review by a court or by an independent administrative body and use was allowed when it would refer to ‘serious crime’, which is defined differently by each Member State (CJEU, 2014). Thirdly, the data retention period was set at a minimum of six and a maximum of twenty-four months, but there were no objective criteria on which period was to be applied when and why, hence there was no assurance that both the period and the data received are limited to what is strictly necessary (CJEU, 2014). Last, but not least, the directive did not provide sufficient safeguards to ensure protection of the data against abuse and unlawful access and use (CJEU, 2014).

Therefore, the Data Retention Directive, by not clearly defining strict criteria for collection, use and retention of the personal data in question, did neither comply with any of the components of the purpose limitation principle nor with any of the other principles of data processing. Thus the processing was not necessary nor proportionate to the objective to be achieved. As explained earlier, this clearly illustrates the importance of the compliance with all principles of data processing as one. Without them there will be no fair and proportionate processing of personal data.

CJEU's analysis in this case can be compared to the advices and consultations by the Member States Data Protection Authorities discussed in the previous section - any rules which set out for the further processing of personal data, need to be compliant with the purpose limitation principle and the other principles of data processing, otherwise they will not constitute a necessary and proportionate measure in a democratic society.

Further cases pending

More cases, specifically relating to the purpose limitation principle are currently pending in front of the CJEU. Those cases will be reviewed in the light of the text of the GDPR, hopefully Article 6(4). One case to keep a close eye on is T-881/16 HJ v EMA. In it, a British citizen claims that “documents in his personal file, which were made public and accessible to any member of staff of the European Medicines Agency for a period of time, were not processed fairly and lawfully but were processed for purposes other than those for which they were collected without that change in purpose having been expressly authorized by the applicant”. Although this question is one step closer to Article 6(4) GDPR, the court is asked to assess the ‘other’ processing within the concepts of fair and lawful processing. This may include determining whether the other purpose is compatible, but it may also be discussed in the light of Article 5(1a) GDPR¹⁵. Nevertheless, it will be definitely a case worth discussing.

Conclusion

Overall, the CJEU did not address further processing directly, however it did acknowledge that each processing of personal data needs to comply with the purpose limitation, among the other principles of fair processing, and the legitimate processing provisions. The presented cases did not provide additional knowledge on how to interpret Article 6(4) of the GDPR. They did not shed much light on the interpretation of the purpose limitation principle, nor the guidance of WP29, especially the substantive multi-factor analysis for determining the (in)compatibility of further processing. However, the CJEU provided two definitions of incompatibility. It specifically connected incompatibility to the principle of accuracy and time as a factor to determine necessity of purposes. That included the presence of an overriding public interest, the amount of time passed and the reach of the processing. It also clearly confirmed the statement by WP29 that ‘the application of all data processing principles and data subject rights should be ensured’. Figure 15 summarizes those findings.

¹⁵ Personal data shall be: (a) processed lawfully, fairly and in a transparent manner in relation to the data subject (‘lawfulness, fairness and transparency’);

Indicates compatibility

- All principles of data processing are met;
- At least one of the criteria for making the processing of personal data legitimate is applicable;
- Data subject rights are ensured;

Indicates incompatibility

- **Definition incompatibility:** Processing of personal data which is inaccurate, inadequate, irrelevant or excessive in relation to the purposes of the processing, is not kept up to date, or is kept for longer than is necessary unless they are required to be kept for historical, statistical or scientific purposes;
- Initially lawful processing of accurate data may, in the course of time, become incompatible where the data are no longer necessary in the light of the purposes for which they were collected or processed;
 - Factors contributing to 'no longer necessary':
 - There are no overriding public interests;
 - A large amount of time has passed;
 - The reach of the processing is ubiquitous;
- More information is processed than needed;
- Data subject rights are not ensured and there are no exceptions to allow that;
- One or more of the principles for processing personal data are not met;
- No legal grounds to make the processing of personal data legitimate is applicable;

Figure 15: Knowledge obtained from case law on processing of personal data

2.4.5 Scholars

Scholars working in the domain of privacy, similarly to the CJEU, recognize the importance of the purpose limitation principle, but do not take it into account when arguing for or against the compatibility of a further purpose (Zarsky, 2016; Koops & Leenes, 2014; Prins & Moerel, 2016). Instead, most scholars, discussing the purpose limitation principle, criticize its adequacy to safeguard reuse of personal data (Zarsky, 2016, p. 996; Koops & Leenes, 2014, p. 2; Prins & Moerel, 2016, p. 2; De Hert & Papakonstantinou, 2016, p. 185). In general, research on data protection rarely discusses the purpose limitation principle in-depth and specifically for Article 6(4) there was been little to no research published to date.

2.4.5.1 General observations

Scholars would usually approach the topics of GDPR/DPD compliance and/or specific principle's application (privacy by design or privacy by default, DPIAs) by first mentioning general statements such as that the purpose limitation is one of the fundamental principles of data protection, that it has been enshrined in the EU Charter and it consists of two elements (De Hert & Papakonstantinou, 2016, p. 185; Zarsky, 2017, p. 1006). Some scholars have acknowledged that Article 6(4) GDPR is defined as an open norm and it consists of five factors to be evaluated in order to determine whether the new purpose is compatible, or not, with the original purpose (De Hert & Papakonstantinou, 2016, p. 185; Koops & Leenes, 2014, p. 7; Prins & Moerel, 2016, p. 19). De Hert & Papakonstantinou (2016, p. 185) argue that the data controller might choose to consider additional criteria, but at minimum the compatibility evaluation needs to be based on the five factors. However, no specific additional criteria were explicitly presented by the De Hert & Papakonstantinou (2016).

Prins & Moerel (2016, p. 43) acknowledge that, ultimately, the compatibility assessment allows for flexibility in the processing - any further processing needs not to be compatible with the original purpose, but it must not be incompatible. Hence, “[i]t would [...] be more accurate to speak of a *prohibition on incompatibility*¹⁶” (Prins & Moerel, 2016, p. 43). However, no scholar has elaborated on the application of the incompatibility assessment, nor on the substantive method presented by WP29, instead, the greatest part of the research on the purpose limitation is criticism as laid out in the following paragraph.

2.4.5.2 Criticism

A common ground of critique among scholars is that the purpose limitation is vague and left to the data controllers for interpretation (Koops & Leenes, 2014, p. 8; De Hert & Papakonstantinou, 2016, p. 186; Kuner, et al., 2016, p. 259). Koops & Leenes (2014, p. 7) stated that the purpose specification simply

¹⁶ Emphasis added

requires the processing purposes to have a legitimate ground and after that the data controllers can process personal data provided that they have “specified ‘explicit and legitimate’ purposes for the processing”.

The substantive multi-factor assessment for compatibility has also been criticized - Article 6(4) GDPR, is simply indicative. Controllers can take into consideration additional criteria on their own discretion and make the necessary evaluations as to whether the new, further purposes are compatible or not (De Hert & Papakonstantinou, 2016). Because Article 6(4) is so indicative, even where there is a genuine desire to comply with the purpose limitation principle, “actual compliance can never be more than a guess” until the authorities check it (Kuner, et al., 2016, p. 159). According to Kuner, et al., (2016, p. 259) and Koops & Leenes (2014, p. 5) the fact that the text of GDPR’s purpose limitation principle still contains vague and uncertain concepts “is not good lawmaking – clearer, more precise, language would have been helpful”.

Another major source of criticism among scholars is that the purpose limitation does not meet the advanced analytics needs of 21st century. The purpose limitation principle relies on the premise that the purposes of data processing can be determined before the processing occurs (Prins & Moerel, 2016, p. 47). However, analysing large and diverse datasets will involve methods which “neither the entity collecting the data nor the data subject considered or even imagined at the time of collection” (Zarsky, 2016, p. 1006). Knowing in advance why and how the data will be processed would limit the size and use of data analytics’ tools. The ultimate goal of data science - to discover novel trends, patterns, and relationships - would be obstructed, which will lead to “substantial loss of economic and social benefits” (Custers & Uršič, 2016, p. 5). Some scholars state that the purpose limitation principle “fundamentally contradict the business logic of fast-growing online platforms like Alibaba and eBay” (Bendiek & Schmiege, 2016).

While the added value of Big Data and the Internet of Things resides in their potential to uncover new correlations for potential new uses once the data have been collected. These new uses may not have anything to do with the original purposes for which the data were collected. There may not even be an original purpose; the data are often first collected in order to subsequently be able to offer potential new services on the basis of an analysis of those data. The primary purposes would be data collection and analysis, as a result of which the purpose limitation test would no longer limit what types of data may be collected in the first place (Prins & Moerel, 2016, p. 7).

Prins & Moerel (2016) presented an interesting argument by demonstrating the divide between practice and law. They claimed that “[d]ata collection and analysis are themselves the purposes for collecting data” (Prins & Moerel, 2016, p. 7). Data controllers set the purposes to be only ‘collection and analysis’, otherwise the purpose limitation principle will limit the types of data which can be collected. Moreover, Prins & Moerel

(2016, p. 7) declared that “if the purpose coincides with data collection and analysis, purpose limitation is no longer meaningful”, because then all further purposes will be obviously compatible.

Defining a very broad and vague purposes of collection, in order to be able to use the data for any future uses would “not resolve this matter” (Zarsky, 2016, p. 1006). Such purposes would not be specific. “Furthermore, stating an unnecessarily broad purpose might even be considered as ‘illegitimate’ and thus lead to unacceptable processing” (Zarsky, 2016, p. 1006). Instead, in order to comply with the purpose limitation principle, data controllers engaging in data analysis will need to inform data subjects of the future acts of processing they will engage in (which must still be legitimate by nature) and closely monitor their practices to assure they did not exceed the permitted realm of analyses. “Carrying out any one of these tasks might prove costly, difficult and even impossible” (Zarsky, 2016, p. 1006). That is why, it can be concluded that data mining not only does not fit the purpose specification element, but and the notion of compatibility.

However, Article 5(1)(b) GDPR allows further processing for “statistical purposes”. Hence, if Big Data analytics falls within this category, it will be compatible. Nonetheless, this exception is further detailed in article 89(1), which states that “appropriate safeguards” must be applied. Such safeguards could be data minimization and pseudonymization. But relying on the exception of “statistical purposes for Big Data analytics is further restricted under Recital 162 - “statistical purposes” implies that the results of such processing are not to be used “in support of measures of decisions regarding any particular natural person” (Recital 162 GDPR). Yet the Big Data practices usually directly impact individuals. Hence, any such further processing needs to meet the substantive multi-factor assessment. Zarsky (2016, p. 1008) states that:

[A]rticle 6(4)(b) calls for considering the context in which the data was collected—a notion counter to that of Big Data, which calls for analyzing data in different and distant contexts. Article 6(4)(c) calls for considering the “nature of the personal data”—yet another factor that is constantly in flux when applying Big Data measures. Finally, article 6(4)(e) calls for the use of possible safeguards such as pseudonymization—a measure which can substantially undermine the quality of the data and the insights it can provide given the loss of identifiable data which adds to the process’s precision and accuracy.

As a result, the harshest critics go as far as stating that the principle of purpose limitation will have to be abandoned (Prins & Moerel, 2016, p. 12). To the contrary, WP29 has stated that it “has no reason to believe that the EU data protection principles [...] are no longer valid and appropriate for the development of big data, subject to further improvements to make them more effective in practice” (WP29, 2014). However, those improvements are yet to be seen and to be specified.

Despite the abundant source of scholarly criticism on the purpose limitation principle, no research has so far focused on analysing the formal and/or substantive assessments for further processing of personal data,

which WP29 presents. Koops & Leenes (2014, p. 7) criticize the assessments in the light of how difficult it would be to hard-code them, but do not criticize the assessments' implementation by a data controller or a court.

In conclusion, this section did not provide additional explicit knowledge on how to interpret Article 6(4) GDPR. Instead, it demonstrated the great need for the development of a body of knowledge and the structured analysis of the existing publications on the matter.

2.5 The body of knowledge surrounding Article 6(4) GDPR

Overall, from the sections on privacy and data protection, fair information principles and scholars' research, it can be observed that the purpose limitation principle is an essential principle of fair information processing and that it serves "the idea of transparency and binding of the more powerful to predetermined conditions" (Koning, 2015, p. 3). However, those three sections did not provide any explicit knowledge on how Article 6(4) GDPR is to be read or applied. They simply helped to understand better the importance of the purpose limitation principle and the need for an explicit knowledge. In contrast, the sections on WP29's opinion, Data Protection Authorities' guidance, and case law added valuable knowledge on how to interpret the text of the GDPR. Specifically, those sources formed the actual body of knowledge regarding Article 6(4):

Any processing of personal data following collection is further processing and needs to pass the factor's assessment provided for at Article 6(4) GDPR. Thus, further processing can be of two kinds – compatible and incompatible. Whether a purpose is one of those two will depend on the initial purposes defined and the appropriate safeguards applied to compensate for the change.

The initial purposes need to be specified, explicit and legitimate. This entails that they should be made available in writing or orally, and are so clear that they leave no doubt or difficulty to their meaning, scope and methods. In particular, the purposes should also be specific enough to enable the detection of separate processes of personal data processing (related & separate purposes) and to enable the assessment of compliance with any applicable laws, any safeguards applied, and the scope of the processing operations. Moreover, the purposes must enable for a match between data subjects' and data controller's expectations, allow for comparison with any subsequent processing purposes and to determine if the purposes do not correspond to the facts of the case. Any initial purpose needs to be at all stages and times based on at least one legal ground, be compliant with all applicable laws and legal principles. For example, the following purposes do not comply with the requirements presented above are: 'develop new services'; 'research purposes'; 'offer personalized services'; 'fix errors'; 'keep devices up-to date and secure'; 'improve own products and services'. There are no exceptions to the specified, explicit and legitimate requirements as presented in Figure 16.

Purpose Specification

- Related & Separate purposes
 - **Specified**
 - Enable the detection of separate processes of personal data processing
 - Reasonable expectations of the data subject.
 - Enables the assessment of:
 - Compliance with an applicable law,

- The relevance of data protection safeguards applied, and
- The scope of the processing operation.
- **Explicit**
 - Clear, leaving no doubt or difficulty in understanding the purpose;
 - Enables a match between data subjects' and data controller's expectations.
 - Available in writing or orally;
 - Allows for comparison with subsequent processing purposes;
 - Allows to determine if the purpose does not correspond to the facts of the case;
- **Legitimate**
 - At all stages and times based on at least one legal ground;
 - Compliant with all applicable laws and legal principles;
 - The legal ground is appropriate to the nature of the underlying relationship between the controller and the data subjects;
- No exceptions

Figure 16: The body of knowledge about the purpose specification requirement

There are, however, exceptions to the obligation that personal data shall be not further processed in a manner which is incompatible with the initial purposes. They are defined within the text of the GDPR and are a subject of additional safeguards which are part of the substantive assessment presented by WP29 in their opinion 203 on the purpose limitation principle. If the further purpose is based on one of the exceptions available and the appropriate safeguards are accordingly met, then the further processing would be compatible.

Compatible, at minimum, are cases at which all principles of data processing are met; at least one of the criteria for making the processing of personal data legitimate is applicable; all data subject rights are ensured and are within the reasonable expectations of the data subjects to achieve the purposes clearly specified at collection, in a way customary to achieve those purposes. Initially lawful processing of accurate data may, in the course of time, become incompatible where the data are no longer necessary in the light of the purposes for which they were collected or processed. The factors contributing to 'no longer necessary' are that there are no overriding public interests; a substantial amount of time has passed; and the reach of the processing is ubiquitous or similar.

In contrast, incompatible is any processing of personal data which is inaccurate, inadequate, irrelevant or excessive in relation to the purposes of the processing, is not kept up to date, or is kept for longer than is necessary unless they are required to be kept for historical, statistical or scientific purposes. At minimum, incompatible are processing activities for which more information is processed than needed; data subject rights are not ensured and there are no exceptions to allow that; one or more of the principles for processing

personal data are not met; and there are no legal grounds to make the processing of personal data legitimate; because incompatible further processing cannot be remedied by adopting a new legal ground;

Compatibility Criteria (Article 6(4) GDPR)

No further processing (any processing following collection) of personal data in manner that is incompatible with the initial purposes;

- **Incompatible**

- Processing of personal data which is inaccurate, inadequate, irrelevant or excessive in relation to the purposes of the processing, is not kept up to date, or is kept for longer than is necessary unless they are required to be kept for historical, statistical or scientific purposes.
 - Incompatible further processing cannot be remedied by adopting a new legal ground;
- Initially lawful processing of accurate data may, in the course of time, become incompatible where the data are no longer necessary in the light of the purposes for which they were collected or processed.
 - Factors contributing to ‘no longer necessary’:
 - There are no overriding public interests;
 - A large amount of time has passed;
 - The reach of the processing is ubiquitous;

- **Compatible**

- Compatible further processing can have a new legal ground to compensate for the change;
- Exceptions & appropriate safeguards:
 - *Public interest or authority vested in the controller;*
 - The data controller has a formal role, a statutory responsibility or a legal obligation of safeguarding public interests;
 - Any authority and/or legal obligation needs to be a necessary and proportionate measure in a democratic society and it safeguards one of the purposes of Article 23(1);
 - *Archival, research or statistical purposes;*
 - The personal data will not be used to support measures or decisions regarding any particular individual;
 - Subject to professional codes of conduct;
 - Where possible, anonymization or pseudo-anonymization;
 - Restriction on access;
 - The publication of the research results is in an aggregated and/or fully anonymized form;
 - Functional separation between participants in the research and outside stakeholders, especially in the case of different data controllers;
 - Subject of a Data Protection Impact Assessment;
 - *Consent by the data subject;*
 - Explicit, specific and freely given
 - A real choice, without imbalance between the data subject and the controller, and without any negative consequences towards the data subject;
 - Health data, data about children, other vulnerable individuals, or other highly sensitive information can only be processed on the legal basis of data subject’s consent;
 - *Legitimate interest for possible criminal acts or threats to public security;*
 - In line with any legal, professional or other binding obligations of secrecy;
 - The data subjects are accordingly informed and/or can reasonably expect the further use of their personal data for such purposes;

Figure 17: The body of knowledge about the compatibility criteria defined under Article 6(4) GDPR

Whether a further purpose will be compatible or not, as already mentioned, will depend on the initial purposes defined and the appropriate safeguards applied to compensate for the change. When it comes to the appropriate safeguards applied, Table 2 summarizes the sub-factors on the basis of which a decision, on the level of (in)compatibility, can be taken. They indicate which appropriate safeguards are taken to compensate for the change of purpose.

Table 3: Complete overview of all sub-factors which indicate the (in)compatibility of a further processing for each of the factors under Article 6(4) GDPR

Substantive factor	Indicates compatibility	Indicated incompatibility
<p>a) any <i>link</i> between the purposes for which the personal data was collected and the purposes of intended further processing</p>	<p>All purposes (collection and further) are written down clearly;</p> <p>All purposes (collection and further) are absolutely necessary and proportionate to what they aim to achieve;</p> <p>The further purpose is implied from the initial purposes. For example:</p> <ul style="list-style-type: none"> - The further purpose is the next logical step in the processing according to the purposes. - The further purpose is within the reasonable expectations of the data subjects. - The further purpose is commonly understood in the same way by the relevant stakeholders; <p>The further purpose is intended for research and afterwards not processed for any other purpose.</p>	<p>One or more purposes (collection and further) are not written down clearly;</p> <p>One or more purposes (collection and further) are not absolutely necessary and proportionate to what they aim to achieve;</p> <p>The further purpose is not implied from the initial purposes. For example:</p> <ul style="list-style-type: none"> - The further purpose is not part of the initial purposes. - The further purpose is unexpected, inappropriate or otherwise objectionable. - The further purpose is not commonly understood in the same way by the relevant stakeholders; <p>The further purpose is intended for research purposes but the research outcome will be used for commercial gains.</p>
<p>b) the <i>context</i> of the data collection and the <i>relationship</i> between the data subject and the controller</p>	<p>The context of the further processing is the same as the collection context. For example:</p> <ul style="list-style-type: none"> - Context of collection is commercial service, the further purpose is also for commercial purposes. - Context of collection is professional service (medical, legal), the further purpose is also for professional services. - Context of collection is compliance with a legal obligation, the further purpose is also for compliance with 	<p>The context of the further processing is different from the collection context. For example:</p> <ul style="list-style-type: none"> - Context of collection is commercial service, the further purpose is anti-terrorism. - Context of collection is professional service (medical, legal), the further purpose is for commercial purposes. - Context of collection is compliance with a legal obligation, the further purpose is based on the legal ground of legitimate interests of the data

	<p>the legal obligation.</p> <p>Within the context of data collection the data subject was informed about the purposes, legal grounds, how the personal data will be used and any consequences. For example:</p> <ul style="list-style-type: none"> - An understandable and easy to find privacy statement. - Clarity on the methods used to process the personal data. <p>There is a balance of power between the data subject and controller. For example:</p> <ul style="list-style-type: none"> - The data subject is more or equally powerful to the data controller. - The data controller does not process personal data in a secret or vague way. - The personal data processing is not bound to a professional secrecy. <p>The further processing of the personal data is based on a valid legal ground. For example:</p> <ul style="list-style-type: none"> - When the further purpose is a subject to a legal obligation, statutory responsibility, public task or formal role of the controller, those always adhere to the principle of legal certainty, transparency, necessity and proportionality. - When the applicable legal ground is contract, for the data subject is relatively easy to terminate the contract. - When the applicable legal ground is consent, the consent is both freely given and informed. <p>Bank car information is retained for future transactions with the valid consent of the data subject (opt-in);</p>	<p>controller.</p> <p>Within the context of data collection the data subject was not informed about the purposes, legal grounds, how the personal data will be used and any consequences. For example:</p> <ul style="list-style-type: none"> - No transparency about the collection of the personal data, neither for the (potential) further purposes. - No clarity on the methods used to process the personal data. <p>There is a misbalance of power between the data subject and controller. For example:</p> <ul style="list-style-type: none"> - The data controller is more powerful than the data subject. - The data controller processes personal data in a secret or vague way. - The personal data processing is bound to a professional secrecy. <p>The processing of the personal data is not based on a valid legal ground. For example:</p> <ul style="list-style-type: none"> - There are no legal obligations or other legal grounds to enable the data collection. - There are legal obligations to enable the further processing, but they do not constitute a necessary and proportionate measure in a democratic society. - When the applicable legal ground is contract, for the data subject is not easy to terminate the contract. - When the applicable legal ground is consent, the consent is not freely given. <p>Bank car information is retained for future transactions without consent (opt-in);</p> <p>Personal data from surveillance cameras, intended only for security purposes, which are further processed for staff monitoring.</p>
--	--	---

<p>c) the nature of the personal data</p>	<p>The personal data processed does not require special protection.</p> <p>The further processing envisages using only a part of the personal data collected.</p> <p>Combining large and different data sets of personal data in a foreseeable and transparent manner.</p> <p>The further processing will use alternative methods which are less intrusive or do not involve personal data processing.</p> <p>The methods used to process the personal data are explained in detail to the data subject.</p> <p>Sensitive personal data is not being processed.</p> <p>Sensitive personal data is being processed only strictly necessary to achieve the purpose;</p> <p>Access to the sensitive personal data processed is only for a specific task;</p> <p>The legal ground to compensate for the change of purpose is adequate to the data subject-data controller relationship.</p>	<p>The personal data processed is sensitive and/or it requires special protection.</p> <p>The further processing envisages using all personal data collected.</p> <p>Without any foreseeability at the time of collection, combining large and different data sets of personal data.</p> <p>The further processing will use new methods to analyse data, without appropriate mitigating measures and quality check of any results.</p> <p>The methods used to process the personal data are secret or not clearly communicated to the data subject.</p> <p>Sensitive personal data is being processed, not strictly necessary to achieve the purpose;</p> <p>The sensitive personal data is being processed by controllers who should not have had access to it.</p> <p>The legal ground to compensate for the change of purpose is not adequate to the data subject-data controller relationship.</p>
<p>d) the possible consequences of the processing towards the data subjects</p>	<p>The further processing will be conducted by the same data controller.</p> <p>The further purpose is conducted for the benefit of the data subject or the public in general.</p> <p>The consequences of the further processing are foreseeable and are communicated clearly to the data subject.</p> <p>The personal data which will be further processed would be available to only a limited amount of people.</p> <p>Systematic collection of personal data (routinely and indiscriminately), with</p>	<p>The further processing will be conducted by a different controller.</p> <p>No sensitive data will be processed for the further purpose, but the impact will be sensitive.</p> <p>Unknown consequences and unforeseen consequences (exclusion, discrimination, emotional impact).</p> <p>The personal data will be publicly disclosed and/or made accessible to a large number of persons.</p> <p>Systematic collection of personal data (routinely and indiscriminately), without the knowledge or clear</p>

	<p>the valid consent of the data subjects;</p> <p>No personal data collection is further processing it for commercial purposes without the data subject's valid consent;</p> <p>Establishment and maintenance of 'general data warehouses' only (personal) data are processed with a clear purposes, for known periods of time, with known security measures and the data subject's valid consent;</p>	<p>understanding of the data subjects;</p> <p>Combining multiple or all possible sources of personal data collection and further processing it for commercial purposes;</p> <p>Secret or unpredictable processing of personal data;</p> <p>Establishment and maintenance of 'general data warehouses' where (personal) data are processed without a clear purposes, for unknown period of time and with unknown security measures;</p>
<p>e) the existence of appropriate safeguards</p>	<p>Additional measures are applied by the controller to serve as a compensation for the change of purpose. For example:</p> <ul style="list-style-type: none"> - The personal data to be further processed is a subject of partial or full anonymization, pseudonymization and/or aggregation of the data - The methods used to process the data do not involve automated decision making and there are no risks of wrong assumptions being made. - The data subject has been informed and is given the possibility to object to processing - Where the further processing is based on consent, it must be really freely given, specific and informed. - The further processing is based on either contractual safeguards regulating the transfer of the personal data between the parties or other formal arrangements. - Avoidance of methods (such as GPRS tracking) which may reveal more data than needed and data which may be used for other purposes; - Adherence to the purpose minimization principle; - The initial and the further purposes both comply with the storage limitation principle; 	<p>The data controller does not take any additional measures to compensate for the change of purpose. For example:</p> <ul style="list-style-type: none"> - The personal data to be further processed is not anonymized or aggregated. - The methods used to process the data involve automated decision making which is not communicated to the data subjects. - The results of the further processing may not be for sure accurate. - The data subjects are not informed about the methods of further processing, thus not given an opportunity to object. - The further processing is not based on freely given and informed consent. - The relationship between the data controllers for the further processing is unclear and there are no contractual safeguards applicable. - Use of of methods (such as GPRS tracking) which may reveal more data than needed and data which may be used for other purposes; - Collection of more personal data than necessary; - The personal data processed (either for collection or further purposes) do not adhere to the storage limitation principle

In conclusion, if the initial purposes defined are specific, explicit and legitimate, and the appropriate safeguards are applied to compensate for the change of purpose, the further processing will be compatible. Otherwise, if the definition of incompatibility is met, the initial purposes do not meet the purpose specification and there are no exceptions, the processing would be incompatible.

Chapter summary

Multiple and different sources of information regarding the purpose limitation principle were discussed in detail. Not all sources of information contributed to the establishment of an explicit and comprehensive body of knowledge regarding Article 6(4) GDPR. However, those which did provide such guidance, allowed for the creation of an actual body of knowledge, presented in the section above.

On the basis of this chapter, and each of its sections, a substantial, objective understanding of the purpose limitation principle was obtained. We started with the basics – how it is positioned with the domains of data protection and privacy, that it is an essential part of each set of fair information principles and that it has an important part within the GDPR, although it does not lack criticism on its wording and positioning.

Of the greatest value, for the creation of a body of knowledge, were the opinions of WP29 and CJEU's case law. Although the two do not follow each other's approaches for interpreting the (in)compatibility of further processing, their analysis were the most detailed ones. Nevertheless, additional guidance and jurisprudence, would be extremely useful to enrich the current body of knowledge. An explicit and complete body of knowledge will enable the automation of Article 6(4), which will optimally relieve the current burden on data controllers of proving compatibility.

Chapter 3

In the domain of legal automatic-interpretation, a “man-computer symbiosis” is yet to be achieved (Licklider, 1960, p. 4). Human societies are data-driven. In contrast, legal provisions are notorious for being a major challenge to strict rationality. The demand is present - precise and correct legal information is a vital resource which needs to be utilized. That is why legal rules, representation and resolution are currently undergoing major transformations in order to meet the needs of the ‘man-computer symbiosis’. This includes changes in the legal rules, for example adopting new legislative tools (such as the GDPR), and changing how the (new or old) rules are applied in practice, for example more control is given to individuals when it comes to processing activities which affect them (Pentland, 2013, p. 78; Mortier, et al., 2014, p. 1; Pagallo & Durante, 2016, p. 17).

Lawyers have spent decades training how to read, interpret, and apply legal rules from multiple domains, one being the domain of data protection. However, not a single formula or approach, for those inherently human activities, has been established. Notwithstanding, lawyers, upon years of work, have been able to establish a number of rules and procedures which, in line with the principle of legal certainty, will always be a factor for the determination of a problem’s outcome. Such work enables the automation of legal texts.

“Legal automation” is a broad notion. It refers to any method using computer programs to represent a legal text in an algorithmic manner: “data (such as facts) are transformed into outputs (agreements or litigation stances) via application of set rules” (Pasquale, 2018, p. 1). Methods to automate legal texts can be ‘simple’, such as the turning of legal arguments into computer readable syntax “through combinatorial analysis, probability calculus, and binary arithmetic” by the German philosopher G.W. Leibniz, and they can be more elaborate such as text mining and machine learning algorithms (Pagallo & Durante, 2016, p. 18). The method to be used will depend highly on the task at hand, because different types of legal work are more or less susceptible to automation. Pasquale & Cashwell (2015, p. 30) distinguished between “High vs Low Susceptibility to automation” and “High vs Low Intensity of Legal Regulation¹⁷”. Low intensity and high susceptibility to automation cases would be precise legal rules such as rules on how tall building can be in a specific city. High intensity and high susceptibility cases would be a breach of contract claim with damages clearly described in the contract. Low intensity and low susceptibility cases would be contracts with unclear provisions referring to unsettled law. Last, but not least, high intensity and low susceptibility would be so-called *difficult cases* where the text contains ambiguous terms. Such difficult cases are usually present in still

¹⁷ According to Pasquale & Cashwell (2015, p. 47) the Intensity of Legal Regulation means the “degree to which legal tasks are simple or complex”, with simple translating into low intensity and complex translating into high intensity.

growing regulatory domains such as cybersecurity and data protection, mainly because they are largely unsettled (Pasquale & Cashwell, 2015, p. 38).

The domain of focus of this work is data protection specifically the GDPR. As previously mentioned, the domain of data protection offers *difficult cases* when it comes to automation of any provisions. Hence, any method to be automate Article 6(4) will have to deal with the high intensity of the text – ambiguity, vagueness; and with the low susceptibility to automation – translate the compatibility assessment into a set of rules. The next two sections will focus on those two issues. The first section will discuss the main challenges which Article 6(4) is facing and second section will discuss the methods available to meet those challenges and whether they are suitable for the automation of Article 6(4).

3.1 Challenges in the legal knowledge engineering

Companies are under the obligation to comply with multiple and different legal requirements and regulations, among which is the GDPR. Compliance is a risky business, hence there is always a chance that an existing process violates a provision of the GDPR or it will introduce a violation in the future. In order to proof compliance, companies need to record their analysis of why they are compliant. Thus, an inevitable part of data protection compliance is the obligation to document. This includes the relevant business processes which process personal data (inventory), any concrete risks resulting from the processes and a set of measures and controls to minimize those risks (data protection impact assessment), including a procedure to secure that the measures applied are really working as intended (Scheer et al., 2006, p. 146, Ashley, 2017). The obligations to document and be able to proof compliance is a substantial part of the text of the GDPR.

Legal obligations, including the GDPR, are written in sets of rules. Almost any rule can be expressed logically and be translated into simple phrases which follow from one another deductively like steps. Hence, legal rules can be automated. In an ideal situation, one simply inputs a question to a computer program and the program determines whether or not the problem is compliant with the rule. However, in real life, automatic statutory reasoning presents multiple challenges, especially when the statutes are “vague, syntactically ambiguous as well as semantically ambiguous, and subject to structural indeterminacy” (Ashley, 2017, p. 38). Specifically, there are four major challenges which prevent legal rules from their automation; namely, ambiguity and vagueness, multitude of information sources, circumstantiality, and incomplete determination. Each of those four challenges will be discussed in detail.

3.1.1 Ambiguity and vagueness of legal texts

“The legal profession¹⁸ holds itself out to the public as expert in the art of communication through language, and yet, it is well known that there has been an old and continuing problem of using language effectively to communicate the mandates of the legal system” (Allen & Engholm, 1977, p. 380). Hence, it is useful to examine the different types of uncertainty in the meaning of what is written in a law. Following the distinction introduced by Allen & Engholm (1977, p. 381), this section focuses only on imprecise legal texts in the sense that they are written in an uncertain way, and not being incomplete.

¹⁸ Both lawyers and scholars

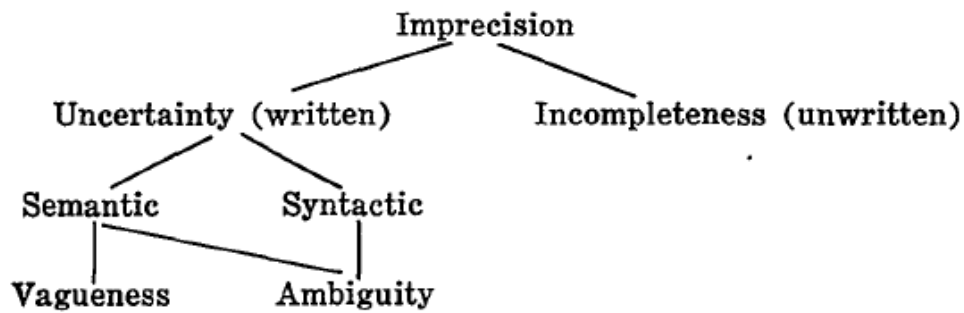


Figure 18: Relationship between imprecisions in law; Allen & Engholm, 1977

Allen & Engholm (1977, p. 382) determines that ‘uncertainty’ should be understood as having a narrower sense than ‘imprecision’ but broader than ‘ambiguity’. Uncertainty envelops both semantic – “how the meaning of the overall sentence is influenced by the range of meaning given to individual words and phrases appearing in the sentence” and syntactic – “how the meaning of the overall sentence is influenced by interpretation of the words that express relationships between the semantic words and phrases” impressions of legal texts (Allen & Engholm, 1977, p. 382). Following the scheme, vagueness is defined as ‘a semantic uncertainty about precisely where the boundary is with respect to what a term does and does not refer to’, while ambiguity is the ‘uncertainty between relatively few distinct alternatives’, as ambiguity can also be defined as language variability or “different expressions convey the same meaning” (Allen & Engholm, 1977, p. 382; Kim, et al., 2014, p. 2).

Having identified that legal language does contain vague, general wording (especially the purpose limitation principle), it must also be acknowledged that uncertainty in legal sources is “usually a deliberate matter” and legal sources also include clear absolute statements from which no variation seems possible (Allen & Engholm, 1977; Kuner, et al., 2016, p. 260). The GDPR is an example of legal statutes with both absolute and uncertain statements. Where used appropriately, the first type—the vague, general wording, commonly found as expressions of principles that require a flexible interpretation – is a the desirable tool if the draftsman wishes to express rules that need to cover unforeseen circumstances (Allen & Engholm, 1977, p. 383; Kuner, et al., 2016, p. 260). Article 6(4) is the embodiment of a flexible provision. Moreover, concepts such as ‘adequacy’ and ‘proportionality’ are also examples of it. The second type—the clear absolute statements—are often (appropriately) found as expressions of fundamental rights, such as the right to privacy (Kuner, et al., 2016). Within the GDPR, such absolute statements are the processing operations which trigger

a DPIA¹⁹ and the requirement that processing of sensitive personal data, such as religious beliefs or health, is prohibited, unless the data subject has explicitly consented to the processing.

A major challenge in the legal domain is the fact that legal interpretation calls for normative judgments (Slocum, 2017, p. 5). Even for legal texts without ambiguity or vagueness, their linguistic meaning may differ from the legal meaning. “Classical logic is unforgiving towards inconsistencies. The reason for this is known as the principle of explosion – from an inconsistent set of premises, every conclusion can be reached” (Schafer, 2017, p. 1). The virtue of language normalization is that it provides for the evolvement of legal systems in the direction of more orderly expression of legal norms (Allen & Engholm, 1977, p. 382). Various approaches to open-texture have been explored. One approach is to apply canons of construction, is a skilled legal activity, and attorneys can and do argue about the applicability of competing canons (Brudney & Ditslear, 2005). To the best knowledge of the author, no legal engineering project has attempted to formalize the process of selecting and applying canons of construction.

The need for normative judgements, however, does not exclude the importance of language’s descriptive insights for legal interpretation – “the meaning of a legal text is generally dependent on objective determinants of meaning that relate to how people normally use language (both inside and outside of the legal context), and which may be said in part to constitute the linguistic meaning of the text” (Slocum, 2017, p. 5). Hence, all interpretations of a rule need to be taken into account. The next section explains more on that topic.

¹⁹ Recital (91) GDPR: This should in particular apply to large-scale processing operations which aim to process a considerable amount of personal data at regional, national or supranational level and which could affect a large number of data subjects and which are likely to result in a high risk, for example, on account of their sensitivity, where in accordance with the achieved state of technological knowledge a new technology is used on a large scale as well as to other processing operations which result in a high risk to the rights and freedoms of data subjects, in particular where those operations render it more difficult for data subjects to exercise their rights. A data protection impact assessment should also be made where personal data are processed for taking decisions regarding specific natural persons following any systematic and extensive evaluation of personal aspects relating to natural persons based on profiling those data or following the processing of special categories of personal data, biometric data, or data on criminal convictions and offences or related security measures. A data protection impact assessment is equally required for monitoring publicly accessible areas on a large scale, especially when using optic-electronic devices or for any other operations where the competent supervisory authority considers that the processing is likely to result in a high risk to the rights and freedoms of data subjects, in particular because they prevent data subjects from exercising a right or using a service or a contract, or because they are carried out systematically on a large scale. The processing of personal data should not be considered to be on a large scale if the processing concerns personal data from patients or clients by an individual physician, other health care professional or lawyer. In such cases, a data protection impact assessment should not be mandatory.

3.1.2 One-to-many & Many-to-one representations of knowledge

“Laws and regulations generally do not specify what measures should be implemented to match the requirements that they state” (Bartolini, et al., 2016, p. 1). Legal rules are sometimes clear, but most of the times they are ambiguous, hence having multiple possible meanings. A provision, the meaning of which, is “plain and uncontested [when it] is represented by just one rule” (Poulin, et al., 1993, p. 93). However, such legal provisions are rare. Instead, most legal terms have many interpretations or sources of information.

A complete body of knowledge, as the one this work attempted to create in Chapter 2, typically accommodates several different meanings of each provision and is not limited to the most plausible or often used meaning. Hence one legal provision is being assigned a one-to-many representation of knowledge. Such a method is quite congenial to the open texture of the law. For example, for each of the factors under Article 6(4) many additional sub-factors were identified, which allow for a better understanding of what the factor entails. The precision of most statutory texts may convey to the non-lawyer the idea that legal concepts and provisions are unequivocal and that the only real difficulty is to design a procedure to determine the meaning (Poulin, et al., 1993, p. 91). However, as evident from Chapter 2, even if all representations of a rule are detected at a certain point of time, there is no guarantee that the list is complete and new, different representations may emerge at a later point – legal interpretation is not static but evolves.

Furthermore, a single provision in a legal source can be represented by several rules in the knowledge body. Provisions that can be interpreted in several ways give rise to as many rules in the object level knowledge base as there are defensible interpretations. Each rule is tagged with a label saying which particular rule of interpretation justifies its inclusion in the knowledge base. It may carry further tags, such as references to the corresponding statutory provision or to related or conflicting rules, perhaps some kind of priority expressing how confidently this interpretation can be sustained, and so on. There is thus a many-to-one relationship between rules in the knowledge base and statutory provisions (Poulin, et al., 1993, p. 91). Within the purpose limitation principle ‘legitimate’ would be such a provision with many interpretations, as made evident at Chapter 2.

When designing a computer program one must interpret all legal rules and potential results ahead of such concrete applications. However, such interpretations can only be provisional – the next section will discuss this in more detail.

3.1.3 Ex ante vs. ex post

As already mentioned, it is a common practice within the legal domain to establish the meaning of legal texts in concrete situations. In contrast, computer algorithms ‘look forward’ - interpret the task at hand ahead of any concrete applications (Daniel, et al., 1993, p. 90).

When a human expert makes a decision on the outcome of a statutory problem, she would research all available information, including statutory rules, exceptions, case law, underlying social values and legislative purposes (soft law). In the very same manner, the body of knowledge from Chapter 2 was constructed. The knowledge is based on already available information, on decisions on post-facto cases.

Nevertheless, some of the guidance relating to the purpose specification was directed towards future cases. The advice of the European Data Protection Supervisor (EDPS) regarding the compatibility of regulation's legal texts were intended to be applied *ex ante* and thus, very useful for the future application of the purpose limitation principle. Hence, the purpose specification can be defined *ex ante*, however the body of knowledge regarding the Article 6(4) is based on *ex post* knowledge. This difference in the two components of the purpose limitation is due to the fact that the compatibility assessment is not fully determined, which makes its application challenging. The next section will explain what that entails.

3.1.4 Legal concepts are never fully determined

“Law is agonistic - a position develops from confrontation with another position, usually one body of evidence against another, but when money is at stake, one view of law against another interpretation of that same law” (Leith, 2016, p. 98). For data controllers complying with the purpose limitation principle, taking into account the enforcement fines in the GDPR, is a money issue. Hence, they want one rule which they can follow without the possibility of a non-compliance. However, legal provisions inherently contain evaluative terms, i.e. terms calling explicitly for a judgement. It is a practice among legislators to leave room for evaluation and weighing of the particular case circumstances. That is why, Article 6(4) is indicative and the end decision is left to the data controllers. This issue can be described by the term ‘open texture’. Open texture is inherent to all words referring to empirical concepts in a natural language. It indicates that a word's meaning is not fixed: each word has a core of certainty and a ‘penumbra’ of doubt (Hart, 1961, p. 120). The term ‘open texture’ comes originally from Waismann (1968, p. 41) and has become well-known since Hart (1961) used it for the famous example of the term ‘vehicle’ and the rule prohibiting the use of vehicles in a park. Vehicle has a core which presupposes anything which has 4 wheels and a driver. However, riding a bicycle in the park is prohibited, hence that type of vehicle falls within the core. But what about a wheelchair, a toy-car or a tank serving as a war-memorial? Those are examples of the ‘penumbra’ of doubt, where disagreements arise (Smith, 1994, p. 12).

A similar observation can be made about the factors under Article 6(4). One of the factors is “any *link* between the purposes for which the personal data was collected and the purposes of intended further processing”. ‘Link’ for the obviously compatible cases may be that the further purpose is part of the initially defined purposes, however it is also tied to the reasonable expectations of the data subject, which can be very wide. Hence, the penumbra of ‘link’ can incorporate any connection between two or more purposes.

Indeed, the purpose limitation principle has an open texture, together with most of the GDPR, because without such a texture the principle will not be applicable for several decades and to methods of personal data processing yet to be discovered. Although it is a challenge to comply with the purpose limitation principle, there are some methods available to avoid this. For example, one method of legal engineering - legal experts systems, circumvent open-texture by following explicitly any guidance from case law as the definition of what a certain rule means (Smith, 1994, p. 12). The same logic was used in the Google Spain (C-131/12) case as already discussed in Chapter 2 – the definition of incompatibility was taken as the basis of the analysis and the core of the term and the factual circumstances were decided to either be part of the penumbra related to the core.

In conclusion, there are several major challenges to the codification of any legal text. As it becomes obvious from the description of the challenges that Article 6(4) has an element from each one of them. Article 6(4) contains vague and ambiguous legal terms, which may have multiple interpretations because they are not fully determined and the only way to know with certainty anything about those specific terms is from ex post case law. Such legal challenges are major obstacles to the automation of any legal rule, specifically Article 6(4). Yet, for reasons outlined in the next section, Article 6(4) GDPR has a number of additional challenges which complicate even further the task at hand.

3.2 The challenges of automating Article 6(4) GDPR

The European ENDORSE project attempted to hard-code the data protection provisions laid down at the Data Protection Directive (DPD). Although it was anticipated that legal rules' translation into software rules would be a significant challenge, "it turned out to be a far more complex issue because hard-coding data protection law involves more than simply transforming and representing rules" (Koops & Leenes, 2014, p. 4). The project acknowledged that the main complicating issue with many legal requirements, as outlined in the previous section, is that they have been formulated in such a way as to allow flexible application in practice. The flexibility is realized by the wording of the provisions which does not stipulate precisely what must be or cannot be done. Koops & Leenes (2014, p. 8) recognized that "this is due to the dual aims of the DPD [and the GDPR]: facilitating the free flow of information and guaranteeing an adequate level of privacy protection". Moreover, Koops & Leenes (2014, p. 7) presented specific reasons why the two components of the purpose limitation principle cannot be automated, as part of their analysis of the European ENDORSE project.

The purpose specification cannot be hardcoded because it is fundamentally left "open to data controllers and therefore allows for a wide variety of purposes defined in natural language" (Koops & Leenes, 2014, p. 7). As long as the grounds for processing are legitimate under Article 6(1) GDPR, data controllers can process personal data provided that they have specified "explicit and legitimate" purposes for the processing (Koops & Leenes, 2014, p. 7). The determination of whether a certain text defined by a controller is legitimate and sufficiently explicit is particularly difficult because it concerns the semantics of the purpose. Without clear instructions on what is always explicit and specific, a machine will not be able to distinguish compliant from non-compliant initial purposes. Koops & Leenes (2014, p. 7) suggest a possible solution for this problem. Taking into account that the purpose limitation principle has not changed from the transition between the DPD and the GDPR, under the DPD there was a requirement that data controllers should report the purposes of processing personal data to Data Protection Authorities. This rule is not within the GDPR, however the DPD was in force for more than two decades, hence a list of possibly certified purposes can be devised. From this list, data controllers can pick the relevant one(s) for their case. "This, however, does not do justice to the fact that the system is principally open" (Koops & Leenes, 2014, p. 7). This means that any collection purpose, also the ones which no one has ever come up with, may meet the requirements for specific, explicit and legitimate purpose. Therefore, as it was pointed out in Chapter 2, for the obviously (in)compatible formal method presented by WP29 a simple string matching will not be enough to determine neither whether the initial purposes are specific, explicit and legitimate, nor whether the further purposes are compatible or not.

The prohibition of processing further personal data in an incompatible to the initial purposes manner cannot be hardcoded because "the actual processing has to take place within the frame defined by the purposes as

defined by the controllers themselves” (Koops & Leenes, 2014, p. 7). This would require a semantic mapping of the initial purposes defined and the purposes for further processing. Conducting this semantic mapping in a software environment would require the data controller to specify for each process to be carried out what is the purpose, is it the data already collected or this is a further processing and then during runtime have the software verify whether the personal data in question may be further processed or not, which implies a decision on (in)compatibility of the processing. This approach may be feasible for certain straightforward purposes such as making an order at Zalando, paying for it and delivering it. However it will be difficult to “hard-code” the more vague or open-ended purposes defined by the data controllers such as “for 3rd party quality assessment and improvement” (Koops & Leenes, 2014, p. 7). Even though we discussed in Chapter 2 that such purposes will not be compliant with the purpose specification requirement, they are nevertheless the current practice and they will be substantially difficult to translate into a machine-interpretable process. This translates into the re-occurring theme that human intervention is needed – whether it will be “to interpret whether a particular system process requires processing of data for ‘quality improvement’” or to determine the semantic mapping in general (Koops & Leenes, 2014, p. 7).

Overall, the codification of Article 6(4) GDPR is not only context-dependent but dynamic, as the interpretation of legal norms, such as lawful, transparent, specific and explicit may shift over time.

The idea of encoding legal norms at the start of information processing systems is at odds with the dynamic and fluid nature of many legal norms. Rules need breathing space, and breathing space is typically not something that can be embedded in software. Simple and very specific rules might be suitable for hard-coding in IT systems, but techno-regulation as enforcement of a legal norm is problematic if the norm itself is more representationally complex, be it due to openness, fuzziness, contextual complexity, or to regulatory turbulence (Koops & Leenes, 2014, p. 8).

Nevertheless, legal engineering programs have been designed to meet one or more such challenges. There is, yet, no method which can address all challenges presented, but at least a sufficient part of it. The next section will present a selection of the methods which have contributed enormously to the domain of legal text automation.

3.3 Methods of legal knowledge engineering

The goal of any legal engineering method is to develop a computational model of legal reasoning (CMLR). CMLRs realize processes which evidence “attributes of human legal reasoning” (Ashley, 2017, p. 4). Such processes may include answering a legal question, predicting an outcome, or even drafting a legal argument. CMLRs can simplify a complex intellectual task into a set of computational steps. At its basics, each computational model would have a set of input variables pre-determined by the expert setting up the model and it would output an answer to a legal question. At the black box between the input and the output, the algorithm’s logic calculates which outcome is possible based on the training set from which it has learned which facts of the case correspond to which outcome (Ashley, 2017, p. 4).

The designing of a model which can address legal questions and problems is a great challenge. Nevertheless, legal engineers have developed models which can perform typical lawyer-tasks such as interpreting what a legal rule means, whether a certain answer applies to a situation, how to distinguish ‘hard’ from ‘easy’ legal issues, and how to interpret legal judgement (Ashley, 2017, p. 4). How did they do that? There are several methods to convert a legal requirement into a computer program. One can use rules - similarly to how an expert would argue (if *a* then always *b*); (mathematical) logic to convert each legal concept into an equation; machine learning algorithms at which the legal requirement is broken down to vectors on the basis of which an outcome can be predicted; or ontological logic to elucidate the fundamental concepts of a system and set out the relations among them, just to name a few. A closer look those methods and how they tackle the automation of legal provisions, similar to the Article 6(4), follows.

3.2.1 Expert systems and logic programming

Rule-based and case-based programs are able to perform intelligent tasks such as legal reasoning and argumentation, outcome prediction and explanation. Those programs make use of knowledge structures which represent a statute’s provisions or judicial reasoning using “schemes of inference and argument to process reasons” (Ashley, 2017, p. 33). Any knowledge structure (or body of knowledge), similarly to Chapter 2, is a subject of manual extraction. As one might expect, manual knowledge extraction is a major bottleneck to the wide applicability of such programs - it is time consuming and it requires an expertise to represent the knowledge into a format which the programming language could *understand* (Ashley, 2017, p. 33).

Hence, depending on the type, either rule or case based, expert systems use statutes or case law translated to expert rules to reach an outcome. In this section, case-based programs will not be discussed, for two reasons. First, the only case-based program which can be applied to the problem at hand is Case Based Reasoning (CBR). But for Article 6(4) there is a shortage (none) of specific cases which explicitly analyse the (in)compatibility of a further processing. Without such information, the CBR would not be able to reach a

conclusion. Second, the focus of this work is on statutory compliance and the methods typically used for case-based programs usually result in lower accuracy when they are applied to automation of statutes. Therefore, the focus will be solely on rule-based programs. Specifically, the most commonly used one within the legal domain - expert systems.

Expert systems are computer programs which solve problems, offer advice, and undertake a variety of other tasks. They do so in manner similar to human experts – the program contains representations of knowledge and expertise, within a narrow area of law, having enough “knowledge and expertise” to “ask a client user relevant questions about his/her problem, to customize its answer based on the user’s responses, and to explain its reasons” (Ashley, 2017, p. 8). The ‘expertise’ of such programs is comprised of heuristics or rules-of-thumb represented in a declarative language specifying matches between conditions and conclusions. This type of knowledge is derived through a manual knowledge-acquisition process, as already mentioned (Ashley, 2017, p. 8).

Expert systems usually make use of the so-called isomorphic approach. It follows the formalism inherent in legal texts to translate them and formalize them as executable logic programs (Sergot, et al., 1986, p. 371). To clarify, the theory of legal formalism treats legal rules or statutes like a mathematical equation or a scientific theorem (Francesconi, et al., 2010, p. 2). Hence, formalizing the outcome of whether a certain rule is applicable or not (identify the relevant legal principles, apply them to the facts of a case, and logically deduce a rule that will govern the outcome) is the same as when a scientist would scope the relevant axioms, apply them to the problem, and systematically reach a method to proof the answer (Francesconi, et al., 2010, p. 2). For automating statutes, the legal scholar would follow the legal sources as closely as possible and write logical propositions which paraphrase or reformulate the items in the source text. The items, then, are translated literally – “the sense ascribed to an item is the straightforward reading that would seem evident to ordinary persons, and this straightforward sense is converted into a logical proposition” (Poulin, et al., 1993, p. 92).

The most famous application of an expert system to the legal domain was that of Sergot, et al., (1986). They implemented a large portion of the 1986 British Nationality Act as a logic program written in the programming language Prolog. The program focused on “the limited objective of implementing rules and regulations with the purpose of applying them mechanically to individual cases” (Sergot, et al., 1986, p. 372). It ran approximately 150 rules, implemented as Horn clauses (variables are embedded into the predicates, for the sake of readability), in Prolog. The program would take questions in the form of propositions. The output would be proving (or not) the proposition and citing an explanation (Sergot, et al., 1986, p. 376):

```

Is Peter a British citizen on date (16 Jan 1984) by sect. Z?
Which X : Peter was born on date X ? (3 May 1983)
Is it true that Peter died before (16 Jan 1984) ? no
Is it true that Peter was born in the U.K. ? yes
Which X : X is father of Peter ? why

```

```

if X is father of Peter
then X is a parent of Peter

if X is a parent of Peter
and X is a British citizen on date (3 May 1983)
then Peter has a parent
    who qualifies under 1.1 on date (3 May 1983)

    Peter was born in the U.K.
    Peter was born on date (3 May 1983)
    (3 May 1983) is after or on commencement, so
if Peter has a parent
    who qualifies under 1.1 on date (3 May 1983)
then Peter acquires British citizenship
    on date (3 May 1983) by sect. 1.1

    Peter is alive on (16 Jan 1984), so
if Peter acquires British citizenship
    on date (3 May 1983) by sect. 1.1
and (16 Jan 1984) is after or on (3 May 1983)
and not[Peter ceases to be a British citizen on date Y
    and Y is between (3 May 1983) and (16 Jan 1984)]
then Peter is a British citizen on date (16 Jan 1984) by sect 1.1

```

```

Which X : X is father of Peter ? William
Which X : William was born on date X ? (1 March 1952)
Is it true that William died before (3 May 1983) ? no
Is it true that William was born in the U.K. ? yes
Is it true that William was found as a newborn infant
    abandoned in the U.K. ? no
Is it true that William was adopted ? no
Is it true that William was a citizen of the U.K. and
    Colonies on date (31 Dec 1982) ? yes

```

Figure 19: Sergot, et al., (1986, p.376) program output

The main idea behind Sergot, et al. (1986)'s expert system is that the content of a whole body of law can be captured using logic:

... through a set of logical axioms, to logically analyse the implications of that body of law for specific cases. For this purpose the rules directly expressing the content of a legal source may be supplemented with further rules specifying when the predicates in a legal rule are satisfied. Once it is agreed that a set of legal norms L provides an adequate representation of the law and that a set of factual statements F provides an adequate representation of the facts at issue, then determining whether a legal qualification holds in situation F can be done by checking whether it is logically entailed by $F \cup L$. (Prakken & Sartor, 2015, p. 5)

This idea was initially expressed by the legal scholars Alchourrón & Bulygin (1981, p. 98) and further developed by Sergot, et al. (1986). Focusing on the automation of the whole act, instead of a single rule, Sergot, et al. (1986, p. 384) also proposed the use of a tree to capture the overall logical structure, where further clauses determine the conditions under which the predicates in the body of higher level rules hold (Prakken & Sartor, 2015, p. 6). This proposition was by itself an innovation which led to the development of

machine learning algorithms using decision trees to analyse legal outcomes. Such cases will be discussed in detail at the next section.

Overall, the work of Sergot, et al., (1986) was highly influential for the “development of computational representations of legislation by showing how logic programming enables intuitively appealing representations that can be directly deployed to generate automatic inferences” (Prakken & Sartor, 2015, p. 6). Although still widely used, legal expert systems are no longer the paradigm for automating the legal domain. Ashley (2017, p. 8) presented two reasons for this. First, expert-systems deal with statutes in an ad hoc manner, hence uncertain and incomplete information needs to be made available by the human expert, which may be unreliable. Second, the process of acquiring rules is solely manual, hence cumbersome, time-consuming, and expensive (Ashley, 2017, p. 8). Another reason recognized by Waterman & Peterson (1981) is that expert systems do not always reach a correct result. Due to the dependency on manual ad hoc interpretation if the expert’s input is not detailed and explicit enough, the results provided by the program will most certainly be wrong.

Can expert systems automate Article 6(4)? In theory, they can. Expert systems can analyse logically any legal statute, rule or case law into a number of rules, on the basis of which an answer can be reached. However, in practice, if we compare the axioms of Sergot, et al., (1986) with the ones from the body of knowledge, defined in Chapter 2, they are distinctively different. Each statement from the British Nationality Act has been logically transformed into a self-evident truth. Moreover, if a single statement is not applicable that will clearly lead towards an answer. In contrast, at its current state, the body of knowledge regarding Article 6(4) contains statements which are vague and ambiguous. As already explained, expert systems cannot deal with such statements. They cannot observe the essential traits of the legal rule by itself. The human expert needs to draft the axioms in a logical way for the programming language to read it. There are other methods which instead of forming statements, break down legal rules into values which corresponds to some features (vectors). By relying on the features of a vector, machine learning algorithms can predict an answer, even if the vector/statement is not fully explicit yet. A more detailed explanation follows in the next section.

3.2.2 Machine learning algorithms

Machine learning is a sub-category of artificial intelligence (AI) which enables computers, with the use of statistical methods, to “progressively improve performance on a specific task” without being explicitly programmed to do so (Samuel, 2000). Machine learning methods can be of two types – supervised or unsupervised. The main difference between the two is that supervised algorithms need to be ‘fed’ with data which is labelled, in order for the algorithm to learn to predict the output from the input data; while for the unsupervised algorithms, the input data is unlabelled and the algorithm learns the inherent structure from it

(Brownlee, 2016). Simply put, supervised algorithms need to be taught what the features of an apple and a banana are, in order to be able to distinguish between the two; while unsupervised algorithms will detect ‘by themselves’ what distinguishes an apple from a banana. An unsupervised algorithms will not ‘understand’ that the two objects are an apple and a banana, but they will correctly determine the features which make those two distinctive. This work will not discuss any unsupervised machine learning methods, since they are not yet effective enough to extract from examples of further processing of personal data the facts indicating the (in)compatibility of the case.

In the legal domain, machine learning algorithms have great application. Supervised methods have been used to predict the answers of legal problems. Therefore, the next sub-sections will present two supervised (k-nn & decision tress) methods of classifying legal outcomes.

K-nearest-neighbours

The first supervised machine learning method to be discussed is k-nearest neighbours. The logic behind this algorithm is fairly simple. The model’s decisions are based on feature similarity – the unknown data-point gets assigned the label of the closest object from the training set within an N-dimensional feature space. Figure 20 illustrates this. The green dot, our unknown data-point, will be assigned either class 1 or class 2. The assignment will depend on the number of training examples to which we will compare the unknown data. If we compare it to the one closest neighbour, then the unknown data will be classified as class 1. However, if we expand our neighbouring scope to 3 or 5, the result will change (Bronshstein, 2017).

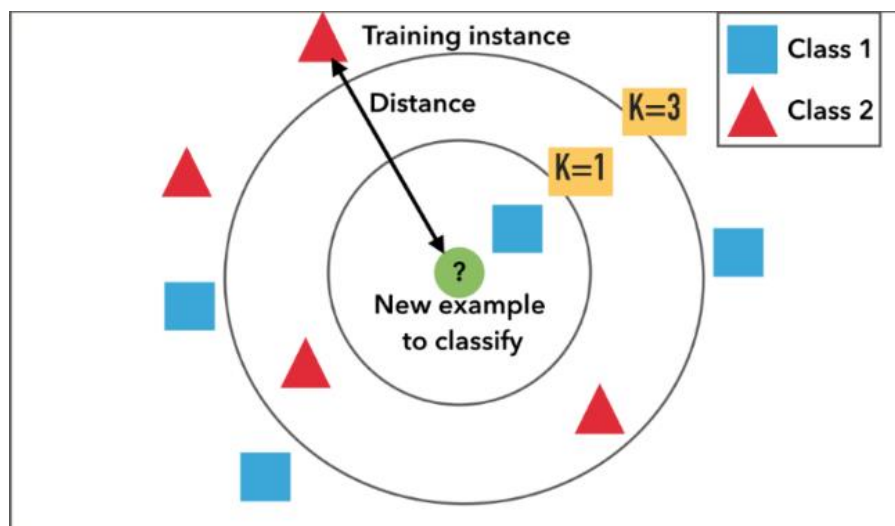


Figure 20: Example of a k-NN classification, (Bronshstein, 2017)

Nearest-neighbour methods are easy to implement. They also yield good results if the features are chosen carefully and are weighted carefully in the computation of the distance. Nevertheless, this model might be too simplistic, especially to capture complicated legal arguments and determine dispute outcomes. kNN

models do not simplify the distribution of objects in parameter space to a comprehensible set of parameters. Instead, the training set is retained in its entirety as a description of the object distribution. Thus, the algorithm can be rather slow if the training set is large. However, if the training set is not large enough, the model will not be able to determine the nearest neighbours. The most serious shortcoming of kNN models is that they are very sensitive to the presence of irrelevant parameters. Adding a single parameter (e.g. adding *class 3* at Figure 18) with random values for all objects can cause the results to be completely distorted (White, 1997, p. 2).

Despite the simplicity of the method, Mackaay & Robillard (1974, p. 10) trained a kNN algorithm to classify a new tax case compared to 60 Canadian tax legal cases (13 favouring the taxpayer and 47 against), over a 10-year span. Each tax case was represented by 46 binary features (true or false), such as the “private party is a company,” and the “private party had never engaged in real estate transactions” (Mackaay & Robillard, 1974, p. 10). When a new case is to be classified, the person who wants to know what the prediction category of the new case will be, has to answer all 46 features. On the basis of the similarities between other cases’ features, the program outputs a prediction based on the “nearest” existing cases (Mackaay & Robillard, 1974, p. 10). Although the research of Mackaay & Robillard (1974) was remarkable, it clearly demonstrates, again, great manual effort, both to define the relevant features and to input an answer to each one of the features, in order to teach the program to group the right neighbours together.

Can a kNN supervised machine learning model automate Article 6(4)? Yes, it can. A similar approach to Mackaay & Robillard (1974) can be adopted. If there are enough (in)compatibility cases which are decided in complete certainty (hence, if personal data is processed in a manner ‘a’ for a purpose ‘b’ the outcome will always be compatible), then the next step would have to be the (manual) extraction of features. Such features can be the sub-factors which we extracted for each of the compatibility factors at Chapter 2 (Table 3). Those sub-factors can be represented by a binary outcome (e.g. The context of collection is professional service (medical, legal), the further purpose is for commercial purposes.). However, our sub-factors are more than 80, which will make the model very disperse, hence not being able to group together all the right features in order to group the nearest neighbours correctly. With so many features, the accuracy of the model will be very low – the model will overfit (memorize the training set), but will not be able to judge correctly on unseen cases. Moreover, our sub-factors may not be a good representation of the features indicating (in)compatibility. As mentioned by WP29 and the authorities, cases of compatible further processing are constantly in flux – the examples are not set in stone and with just one factor being different, the outcome will be different as well. Overall, kNN model can be used to automate Article 6(4), but then the body of knowledge will have to be expanded to include more cases of (in)compatibility and the relevant features will have to be manually extracted and peer reviewed.

There are other, more simple, supervised machine learning models which can be used to automate legal rules, such as decision trees. The next section will focus on them.

Decision Trees

A decision tree is predicting the class (e.g. fruit) of an object (e.g. apple) from the values of its predictor variables (red, sweet, round), the result of which is similar to a tree structure (e.g. is it red – yes; is it round – yes; is it sweet – no; hence, not an apple). A decision tree is constructed by taking a learning sample of data in which the class label and predictor variables' values for each case are known, and apply it to unknown data. Each partition is represented by a node in the classification tree (Loh & Shih, 1997, p. 817'). Decision trees are designed for dependent variables that take a finite number of unordered values, with prediction error measured in terms of misclassification cost (Loh, 2011, p. 15).

Among the machine learning methods, trees are the most transparent and easy to interpret. They are based on the rule of separating observations into subgroups by creating splits on predictors. Those splits create logical rules that are transparent and easily understandable. The resulting subgroups should be more homogenous in terms of the outcome variable, thereby creating useful prediction or classification rules (Shmueli, et al., 2008). Figure 21 illustrates a simple decision tree.

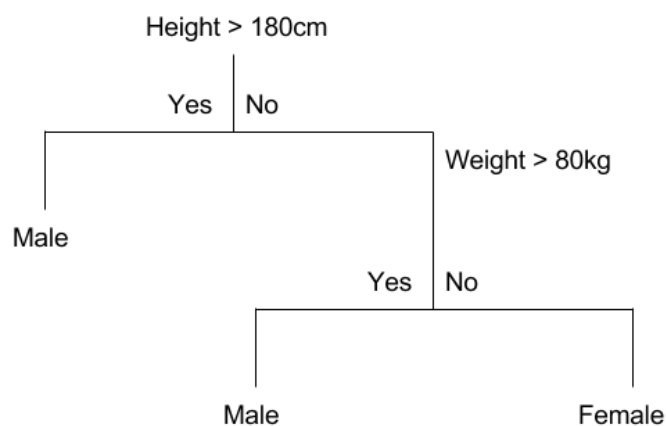


Figure 21: Example classification tree (Brownlee, 2016)

The decision on when to make a distinction between yes and no (path 1/path 2) for decision trees is made by the algorithm on the basis of the rules inferred from the training set. Decision tree's predictions are data-driven. Similarly to the kNN models, if the algorithm is trained with biased data, the prediction accuracy may be distorted or be based on spurious correlation, instead of actual causation. The algorithm would learn rules based on statistical regularities that may not be what a human expert is expecting. Therefore, the strength of machine learning models, not to rely on human logic and to be able to identify features not-

obvious to a human eye, could also be a drawback if the expert cannot identify the reason for the prediction. For example, for the simple decision tree in Figure 21, a woman which is higher than 180cm will be always classified as a man. In order to avoid such misclassification, the expert would have to adjust the features better – e.g. add the feature of country.

This characteristic of the decision trees is especially relevant to their legal domain application. “Since the rules [that the] [...] algorithm infers do not necessarily reflect explicit legal knowledge or expertise, they may not correspond to a human expert’s criteria of reasonableness” (Ashley, 2017, p. 111). That does not mean that the decision tree’s predictions to legal problems would always be wrong. It does mean, however, that there must be safeguards to ensure the model will decide (make a split) in a logical way. How to guarantee this without relying on human expertise again? The following example shall explain.

Katz, et al. (2014) predicted the outcome of United States’ Supreme Court cases, using a decision tree algorithm. The prediction, or the task was to: “either affirm or reverse the judgment of a lower court” (Katz, et al., 2014, p. 1). In order to achieve this task, a decision tree method, developed by Breiman, et al., (1984) and later improved by Ruger, et al., (2004) to “forecast the respective votes of Supreme Court justices for the October 2002 Term” (Katz, et al., 2014, p. 2), was used. The model was fed with data from the US Supreme Court Database (SCDB). Each case was labelled with up to 247 variables, including chronological, background, outcome, voting and opinion variables. The labelling of each case was done manually and was performed by Katz, et al. (2014) who relied on previous research on the US Supreme Court decision-making to determine which features would be most meaningful. Examples of some of the features are “court level and justice-level variables such as party of appointing president, segal-cover nomination score, year of birth and natural court”; and case variables such as “issue, law Type, respondent, petitioner, case Origin” (Katz, et al., 2014, p. 2).

The algorithm had the task “to explore the space and identify the optimal configuration that best predicts the Court’s behaviour based on the large number of features” (Katz, et al., 2014, p. 6). Hence, Katz, et al. (2014) were explicitly looking for a machine-induced rules which will show them whether the manually-selected features are performing well. As a result, they “correctly forecast[ed] 69.7% of the Court’s overall affirm / reverse decisions and 70.9% of the votes of individual justices across the 7,700 cases and more than 68,000 justice votes” (Katz, et al., 2014, p. 10). Those results mean that the decision tree model, once it has been trained to recognize cases using the multiple features described earlier, would correctly predict the outcome in 7 out of 10 Supreme Court pending cases. For the other 3 cases, predicted wrongly, the model has interpreted that a case’ features (e.g. appointing president) would lead to an answer affirm instead of reverse the judgment of a lower court. Although, this accuracy is quite good, the legal expert would then trace back the logic of the algorithm and evaluate where the decision tree made a wrong split and learn from it. Then the

features could be changed or limited to the ones which are most informative. However, this analysis will also have to be performed manually and is quite tedious.

Can a decision tree algorithm be used to automate Article 6(4)? Yes, it can. The task of the Article 6(4) decision tree would be predict whether a certain further purpose will be compatible or incompatible. Similarly to the k-NN methods, and as demonstrated by Katz, et al. (2014), the decision tree model needs features. Those features would need to be manually extracted and ‘fed’ into the model. The greatest challenge to the creation of those features is that, in comparison to Katz, et al. (2014) who relied on previous research on the US Supreme Court decision-making, there has been no previous research on the body of knowledge about Article 6(4). The sub-factors identified in Chapter 3 would optimally be a subject of academic critique and will be improved so that they can be used to train a machine learning model.

In conclusion, without the manual extraction of features no machine learning model will be able to predict the outcome of further processing purposes. There are however, machine learning models which can extract features from the text directly.

3.2.3 Ontologies and taxonomies for legal text analysis

Legal knowledge is expressed through domain-specific terminology which is not directly machine-readable. Hence, legal texts need to be converted, using extracting and mining methods, to enable the formation of a domain-representation-model, such as an ontology. An ontology is an “explicit, formal, and general specification of a conceptualization of the properties of and relations between objects in a given domain” (Wyner, 2008, p. 361). An ontology would transform a vague legal domain into a human understandable, machine readable format that consists of entities, attributes, relationships and axioms (Santos, et al., 2016). In particular, ontologies reflect the “semantic relationships between terms” (Dietrich, et al., 2007; Ashley, 2017).

This definition requires that any terms and the relations among them should be explicitly expressed and represented using a formal language. The most commonly used formal language to create ontologies is the Web Ontology Language (OWL). It is a Semantic Web language designed to represent rich and complex knowledge about terms, how they are grouped and the relations between them. OWL is a computational logic-based language, hence any knowledge expressed through it can be exploited by computer programs. In order to transform legal terms into logic, legal ontologies translate the relations among terms by representing them as labelled links in one of the following ways:

- is-a: class membership expression
- has-as-parts: indicating a part-whole relationship
- has-function: indicating a functional role of the parent
- has-parent, has-child: indicating relative position in a hierarchy

Figure 22 illustrates an ontology on the rights which emerge whenever an incident occurs and the links between them and other legal terms. This ontology was developed by Santos, et al., (2016, p. 9) and it represents the relevant legal knowledge for consumer disputes. Knowledge about consumer disputes is usually found in text excerpts from many heterogeneous sources and it is very difficult for a non-specialist to identify those sources. Therefore, the Ontology of Relevant Legal Information in Consumer Disputes (RIC), from Figure 22, is “the domain-independent ontology modelling this relevant legal information comprising rights, their requisites, exceptions, constraints, enforcement procedures, [and] legal sources” (Santos, et al., 2016, p. 1). The labels indicate what is the relationship between the legal sources, thus making it logical not only for a machine to understand it but and for non-experts (Santos, et al., 2016, p. 9). For example, the bundle of rights are depicted in a legal source, while the entitlement of rights will depend on a requisite.

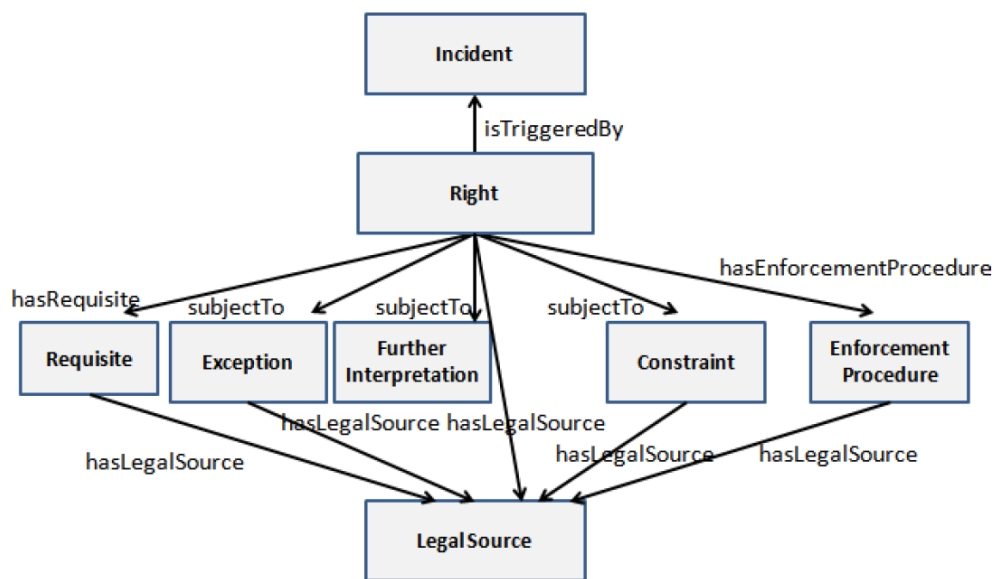


Figure 22: Relevant Legal Information for Consumer Disputes ontology. Arrows denote object properties, domain and range (Santos, et al., 2016)

This ontology focuses on a specific legal problem and its relation to other legal concepts and terms, hence it is a lower-level domain ontology. It offers “a specification of the objects, predicates, and relations for a given domain” (Ashley, 2017, p. 173). However, ontologies can also be of high-level - ontological frameworks (Breukers & Hoekstra, 2004; Breuker, et al., 2004; Ashley, 2017, p. 173). An ontological framework specifies the fundamental concepts for a knowledge engineering. Such an ontology would be positioning the consumer disputes topic among all other types of disputes and show the connections between each, thus connecting dispute resolutions within broader concepts and parts of law.

Ontological frameworks are very useful to visualize large bodies of knowledge. As a result, they have been used extensively by academics to represent the relations between the data protection concepts and principles within the DPD or the GDPR (Casellas, et al., 2010; Cappelli, et al., 2007). For the GDPR, Bartolini, et al.,

(2015, p. 6) developed a bottom-up ontology describing the elements of the Regulation and their relations. Figure 23 highlights the obligations of the data controller, derived and defined from the GDPR, while contrasting them to the data subjects' rights. The result is a set of ontology classes, their attributes and the relations between them fostering the transition of IT-based systems, services and businesses to comply with the GDPR (Bartolini, et al., 2015, p. 1).

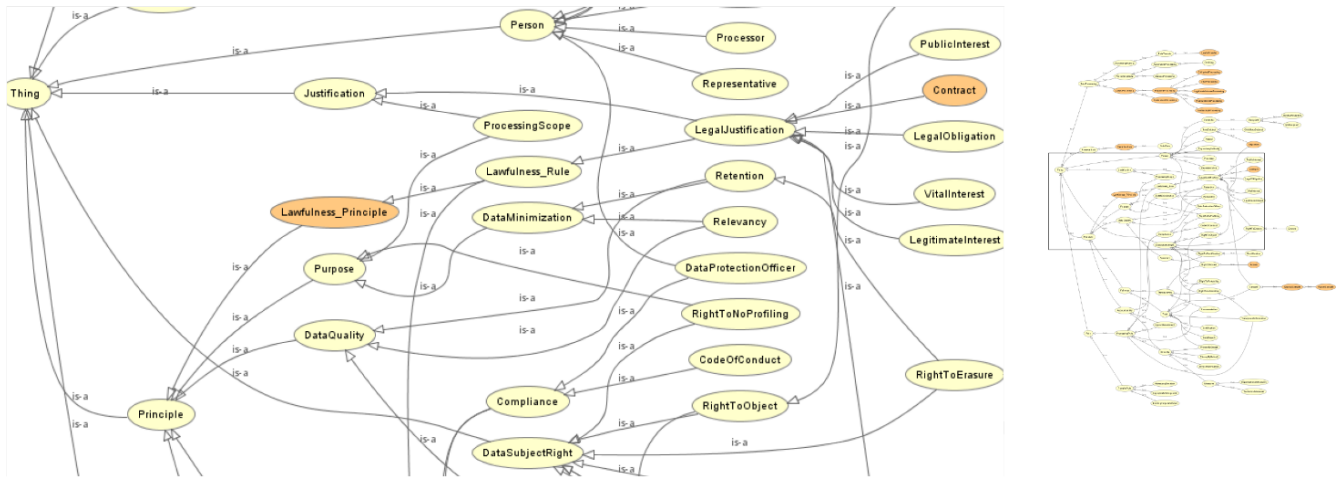


Figure 23: Part of the GDPR ontology (Bartolini, et al., 2015)

This ontology ultimately constitutes a knowledge base from which the concepts to annotate a workflow model can be extracted. This model will allow “data controllers [...] [to] have a clearer view of their duties with respect to data protection in the context of their business; auditors [...] [to] have a first-look model to assess the GDPR compliance; [and] DPAs [...] [to] have a structured approach to detect potential violations” (Bartolini, et al., 2015, p. 3). Although of substantial value for the data protection domain, building and maintaining an ontology, such as Bartolini et al.’s, manually is a resource-intensive, time consuming and costly task. This difficulty in capturing knowledge is the knowledge-acquisition-bottleneck which is also a major obstacle to expert systems (El Ghosh, et al., 2017, p. 473).

An alternative to ontologies are taxonomies. Ontologies comprise of five main modelling primitives: concepts, taxonomical relations (sub-class relations), non-taxonomical relations, axioms and instances (individuals) (El Ghosh, et al., 2017, p. 477). The taxonomical relations (sub-class relations) are part of ontologies but they can be used on their own. A taxonomy's purpose is knowledge classification. In comparison, an ontology goes beyond and creates a knowledge representation. In the legal domain, a taxonomy “is sorting and classifying rules of law”, in order to “make law easier to access and use” (Sherwin, 2009). The ultimate benefit is providing a common vocabulary of general legal terms which help legal practitioners to discuss a subject in a consistent manner and “understand it at a higher level of abstraction” (Sherwin, 2009, p. 42).

According to Sherwin (2009, p. 28) there are three main methods for classifying law using taxonomies. One method is to use a formal taxonomy to sort legal rules in such a way as to maintain the logical relationships between the categories of law. This translates into the structure of a foundation of legal categories, drawn from tradition or from the general functions of the rules, on the basis of which the body of legal materials is sorted into a logically coherent classificatory scheme. This method is fairly simple - it facilitates legal analysis and communication, however it does not track nor establish “normative grounds for legal decision-making” (Sherwin, 2009, p. 43).

Another method to classify law is using a function-based taxonomy which classifies legal rules according to “the roles they perform within a legal system or society at large” (Sherwin, 2009, p. 34). Distinctively, for this method, the relations between legal categories will not refer to any rationales for or against particular legal rules, nor will they explain whether those rules are sound solutions to the problems they address. A functional taxonomy simply provides a purposive overview of the field - it accommodates “a critical evaluation of law as a social institution by providing a comprehensive overview of the field” (Sherwin, 2009, p. 43). This method provides an analytical tool that may, potentially, shape legal reasoning, however it cannot answer legal questions. A function-based taxonomy would be too passive for researchers aiming at contributing directly to the improvement of legal outcomes.

Last but not least, is the reason-based taxonomy method. It “classifies legal rules and decisions according to the moral principles or ‘legal principles’ thought to justify them” (Sherwin, 2009, p. 1). Such a taxonomy offers courts a set of high-level decisional rules drawn from legal data, hence making the law clearer and more complete by guiding the courts in deciding new cases and evaluating precedents. Reason-based taxonomy may be useful to lawmakers but is unhelpful when offered as a guide to adjudication of disputes.

The taxonomy method to be used will depend on the task at hand. For example, Young (2013) created a formal personal data taxonomy, Figure 24, in order to solely illustrate the different sources of personal data, without focusing on how the data is being used and the implications of any data use.

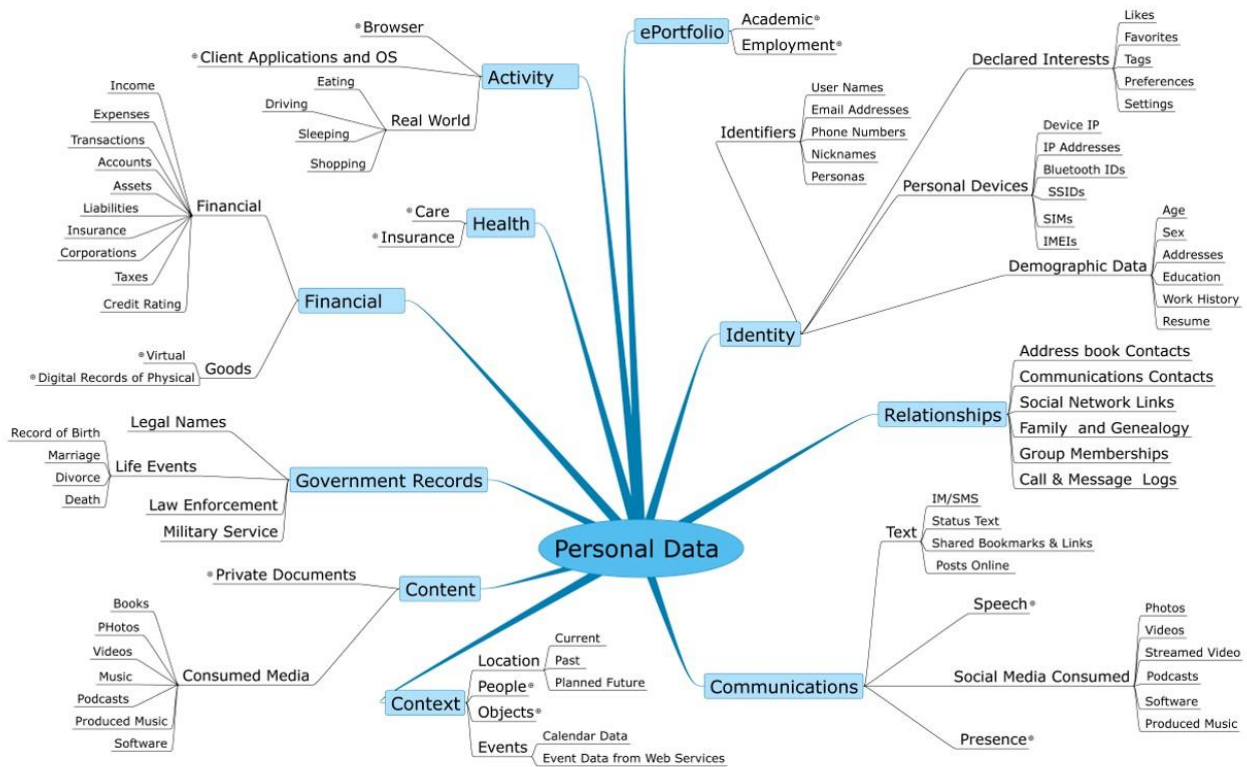


Figure 24: Personal data taxonomy (Young, 2013)

This taxonomy can help data controllers to evaluate whether their data processing involves personal data, hence confirm that they need to comply with the GDPR’s requirements. However, this ontology does not help a data controller to determine whether certain categories of personal data are sensitive or what types of technical and organizational measures are appropriate for which type of personal data processed. Hence, Young’s taxonomy, a formal type, would not be appropriate to domain of processing personal data, neither Article 6(4) GDPR. Instead, the described function and reason-based taxonomies can be used to help compliance with the GDPR. For example, one can use reason-based taxonomies to classify the purposes of data processing. This is a very challenging task since a processing purpose can be anything – any string of text determined by the data controller. Hence, one needs an abundant source of purposes which are recognized as being compliant with the purpose specification as proposed by Koops & Leenes (2014).

One such unique source of data controllers’ purposes was collected by the Dutch Data Protection Authority. The Dutch DPA used to, under the former Dutch Personal Data Protection Act (Wet bescherming persoonsgegevens), oblige data controllers to report the purposes of any personal data processing. All reported purposes were recorded in a public register – Meldingenregister Autoriteit Persoonsgegevens. Although, this obligation, under the GDPR, is no longer applicable and since november 2017 the Dutch DPA does not maintain that public register any more, “53,423 purpose notifications from 32,632 responsible

organizations, institutions, companies and governments” have been recorded and made available as a dataset²⁰ (OpenState, 2017).

This dataset has great potential. It contains a substantially large amount of purposes written in multiple possible ways, from diverse types of data controllers, aiming to realize processing activities ranging from website cookies, public obligations to marketing profiling and academic research. However, this dataset has three major flaws.

First, the fact that a purpose is part of this public register does not automatically entail lawfulness and compliance with the purpose specification. The Dutch DPA did not check the purposes in terms of content but merely recorded what was reported. The ultimate responsibility that a processing purpose is legitimate and compliant with the principles of data processing lies within the data controller. This entails that any scholar who would like to use this data would have to first address the lawfulness and compliance with the purpose specification of each purpose recorded, which would be a very complex task.

Second, not every possible purpose can be found within the 53,423 recorded purposes. Under the Dutch implementation of the DPD, data controllers had a choice – they could report their purposes to the company DPO, which on theory should make those purposes publicly available. Irrespectively of whether the DPO made those purposes publicly available or not, they are not part of the dataset in question. Moreover, some processing activities are exempt from reporting. This would include ‘obvious’ processing operations such as payroll or membership administration. Processing of personal data by the police and judiciary is also not part of this database because those purposes, although notified to the DPA, are not publicly disclosed.

Third, Hollebeek (2017) conducted an analysis of the quality of the dataset and concluded that this dataset has a very low quality which does not allow for direct analytics. The data was collected through a form which the data controller(s) filled-out on their own and was not checked afterwards for intelligibility or completeness. The fields of the form are represented as features in the database. Hence, for each record within the database, there are always the same features: name of the data controller, name of the processing activity, personal data elements, a general purpose description with possibility to name several related purposes (description and goal), whether any sensitive personal data is transferred, who receives the personal data and whether the personal data is shared with 3rd parties outside of the EEA. The populated answers, however, are not standardized – each answer is a different string of words. For example, the category of personal data ‘email address’ has been written down in multiple variations: ‘e-mail addresses’, ‘email’, ‘email address’, ‘e-mail address’, ‘email adres’. Recognizing that those terms relate to the same concept is a trivial task for a human, however for a machine this is not an easy task. Moreover, a great number of the

²⁰ The dataset can be accessed here: <https://data.openstate.eu/dataset/meldingenregister-autoriteit-persoonsgegevens-ap> and here: <https://openstate.eu/nl/2016/07/meldingenregister-autoriteit-persoonsgegevens-ontsloten-als-open-data/>

features are made out of unrecognizable words due to typing errors. Examples include ‘Orgaaan’, ‘ddor’, ‘administant’. Therefore, the dataset would have to be a subject to extensive and careful pre-processing. That would include spelling correction, tokenization (splits longer strings of text into smaller pieces, or tokens), normalization (eliminating affixes (suffixed, prefixes, infixes, circumfixes) from a word in order to obtain a word stem), lowercase all characters, remove numbers, punctuation and default stop words (the, as, a, an, and, to). If all of those methods are performed, the dataset could be suitable for an analysis. Such a pre-processing can be achieved automatically with the use of regular expressions and/or other text mining methods. Nevertheless, another more fundamental issue with this dataset puts its use at jeopardy - some instances are incorrectly completed. Since the data controllers themselves had to fill-in the information, there is has a great number of empty fields or strings which are non-meaningful (for example, a purpose of ‘working together on scientific’). In some instances, when an answer is empty or incomplete there could be additional information provided at another field of the form. In order to identify the value of each purpose record a manual check, on what kind of information is missing, it is essential for the complete understanding of the purpose and is it provided within the other answers, would be required. Such a manual effort would require a time investment by a domain specialist to determine when a missing information does not affect the record and when it would have to be removed because it would introduce noise and potential bias to the data.

In conclusion, a dataset containing a large record of purposes of processing personal data can be formalized as a taxonomy. It would bring for a classification of the different types of processing of personal data and ultimately help for the better formation of a body of knowledge regarding Article 6(4) GDPR. Nevertheless, there is no dataset, available yet, to realize this aim. The dataset of the purposes’ public register from the Dutch DPA is a step in the right direction. It translates a legal obligation into a machine readable format which, upon a substantial effort to pre-process the data, can have some major implications to the automation of the purpose limitation principle.

Overview of the chapter

Chapter 3 pointed out the classical challenges which each legal knowledge engineering needs to meet – legal texts are vague and ambiguous, their application and analysis is ex post and their meaning may change over time. Article 6(4) GDPR adheres to those challenges and adds specific challenges which further complicate the aim of automating it. Nevertheless, there are several methods which can (partially) meet those challenges. Although not every method would be applicable for the automation of Article 6(4) GDPR, there is a great potential to realize the task at hand. Thus, the next section will discuss a method to automate Article 6(4).

Chapter 4

4.1 The method to automate Article 6(4) GDPR

Can the legal reasoning behind the notion of compatible use be automated? From the methods of knowledge engineering presented and discussed in Chapter 3, it became clear that expert systems cannot deal with vague and ambiguous statements, which are typical for Article 6(4) GDPR. Only if a domain expert overcomes the challenge of drafting axioms which elucidate how (in)compatibility is to be address, expert systems can automate Article 6(4). A similar answer can be given to the question if ontologies and taxonomies can be used. Ontologies enable a broader representation of principles' connections, hence are not applicable in general. Taxonomies, although they cannot help with answering legal questions can classify the purposes of data processing. However, this would be a very challenging task since a processing purpose can be anything – any string of text - determined by the data controller, unless a domain expert can overcome the challenges identified with the purposes dataset formed by the Dutch DPA.

What other methods of legal engineering are left? Supervised machine learning algorithms. They can be used to automate Article 6(4) if there are enough cases (instances) and good for prediction features. Although supervised machine learning algorithms do not specifically take into account the knowledge constrains and the specific difficulties of Article 6(4) they can nevertheless produce an answer (prediction) on the outcome of a further processing of personal data. This entails that the automation realized does not provide a solution towards the openness or ambiguity of Article 6(4) GDPR. Instead, it takes an alternative path – a dataset, using 60 cases or instances to extract 13 features, is created. With this dataset, a number of supervised machine learning algorithms are training to predict the outcome of a further purpose (either compatible or incompatible). The algorithms selected are a baseline model, a single attribute model, a decision tree and a k-nearest neighbour model. Upon comparing the results from each model test it can be stated that the results are promising, however, additional research would be needed to confirm the effectiveness of the dataset and the methods used.

4.1.1 The dataset

The automation of Article 6(4) GDPR consists of the creation of a dataset and training machine learning classification methods to predict the outcome of a further processing of personal data. The prediction is using supervised classification models. Such models are used to predict the class a data point is part of (*discrete value*). In classification, the model induced from the data defines a decision boundary that separates the data described by its features into 2 classes or more.

Classifiers need a set of features to characterize each object. Therefore, a dataset was created on the basis of all cases used in the formation of the body of knowledge regarding Article 6(4) GDPR. Each case discussed in Chapter 2, for each of the layers of knowledge, were translated into rows of instances²¹. An overview of all cases used can be found in Table 4 (Annex). For each case several features could be extracted. Each feature represented the categories into which a case can be labelled and was represented in the dataset as the columns. The logic behind the feature selection was fairly simple. We started with the first case - WP29, Example 1: Chatty receptionist caught on CCTV – and manually extracted a number of distinctive facts which also corresponded to the sub-factors presented at Table 3. Each feature could be answered with a binary output – yes/no. For example, the feature of ‘initial_purposes_different’ when answered with a ‘yes’ it meant that the initial purpose is different from the further purpose. The feature selection started by defining several labels for the first case, as presented in table below.

Table 4: Initial features selected

Feature	Full description
initial_purposes_different	The initial purposes is different from the further purpose
ds_informed	The data subject is informed about the further processing purposes and is given an opportunity to object.
power_imbalance	There is a power imbalance between the data controller and data subject.
negative_impact	The further processing will have a negative impact on the individual.
positive_impact	The further processing will have a positive impact on the individual.
new_lg_compensate	New legal ground to compensate for the change of purpose.

²¹ For example, the first instance is based on WP29’s Example 1: (page 56):

Example 1: Chatty receptionist caught on CCTV

A company installs a CCTV camera to monitor the main entrance to its building. A sign informs people that CCTV is in operation for security purposes. CCTV recordings show that the receptionist is frequently away from her desk and engages in lengthy conversations while smoking near the entrance area covered by the CCTV cameras. The recordings, combined with other evidence (such as complaints), show that she often fails to take telephone calls, which is one of her duties.

Apart from any other CCTV concerns that may be raised by this case, in terms of the compatibility assessment it can be accepted that a reasonable data subject would assume from the notice that the cameras are there for security purposes only. Monitoring whether or not an employee is appropriately carrying out her duties, such as answering phone calls, is an unrelated purpose that would not be reasonably expected by the data subject. This gives a strong indication that the further use is incompatible. Other factors, such as the potential negative impact on the employee (for example, possible disciplinary action), the nature of the data (video-footage), the nature of the relationship (employment context, suggesting imbalance in power and limited choice), and the lack of safeguards (such as, for example, notice about further purposes beyond security) may also contribute to and confirm this assessment.

Gradually, for each next case, if found fit, a new feature would be added. For example, the first case from WP29 does not include any information on anonymization, however other cases do and this is a very important feature to detect (in)compatibility as the GDPR does not apply to non-personal data²². The full list of features, for all cases, is presented at Table 4 of the Annex.

The features selected is a subjective task. On the basis of the gut-feeling of the author, as explained in Chapter 1, developed from the formation of the body of knowledge regarding Article 6(4) GDPR, the 13 features were selected. This method is in contrast to Katz, et al. (2014) who predicted the outcome of United States' Supreme Court cases by relying on previous research on the US Supreme Court decision-making to determine which features would be most meaningful. This translates into a certain degree of uncertainty on whether the features are the optimal type and number. In particular, several challenges were identified.

Initially, the data set had both positive and negative features, but then the features were adjusted to include only positively phrased feature names. Whether the label is positively or negatively framed may have influence on the researcher who will have to label any additional cases of (in)compatibility. Hence, if the feature is framed negatively, the researcher will have a tendency to classify it with a negative value – 'no', which will ultimately will have an impact on the performance of the model (e.g. ds_not_informed vs ds_informed) (Heck & Krueger, 2016, p. 337).

Moreover, it was difficult to describe some features in a binary way – 'yes' or 'no'. For example, the feature of 'power imbalance' in the case of Case 5 WP29 "The data subjects are not informed of the initiative prior to the supermarket sending out the leaflets, and the initiative itself is not defined in law", can be a subject of discussion. The supermarket could be perceived as having a power imbalance because if the client unsubscribes s/he will lose her loyalty program. The same problem occurred when deciding for cases on the basis of the feature 'sensitive_data'. WP29's example 7, states that "[t]he photos are inoffensive but somewhat intimate as they artistically capture private moments and emotions while trekking at high altitudes" (WP29 203, 2013, p. 60). The decision on whether this feature is to be represented by a 'yes' for this specific case may vary among experts. On the one side, photos may reveal racial information, hence the data can be sensitive. On the other side, on the basis of the case description there is not enough information to determine whether the personal data is sensitive in accordance with Article 9 GDPR²³.

²² With some exceptions for data analytics

²³ **Article 9 Processing of special categories of personal data** 1.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.

Another peculiarity of this dataset is that, although, WP29 provided only 22 examples of applying on practice the compatibility assessment for further processing of personal data, some of the examples could be broken down to two or more data inputs. For example, were WP29 has clearly indicated that if for a certain incompatible further processing of personal data can be compensated by a valid consent, then this will be represented as two cases – one as the incompatible case and one as the compatible case where a valid consent compensates for the change of purpose. Accordingly the ‘yes’ and ‘no’ labels for each feature will be corresponding to the outcome.

Last but not least, the features presented in Table 4 Annex do not include all possible features which may be strongly correlated to the output variable (label). Correlated features can consistently predict the value of the label. This means that for the 60 cases within the dataset there are *only* 13 features, but there potentially could be additional ones or different features which may be strongly correlated to the label, hence always predicting that label. For example, the dataset takes into account that for personal data which has been anonymized, even if it is to be further used for analytics purposes, as long as it is no longer in an identifiable format, the further processing would be compatible. However, the features, as they currently are, do not take into account that even anonymized data, if to be used afterwards to affect other individuals based on results inferred from data analytics, would be considered an incompatible further processing.

Considering that the dataset can be a subject to change, with the cases from the body of knowledge, and the features similar to the sub-factors presented in Table 3, a training set was developed. The training set is used by the classification programs to learn how to classify objects. There are two phases to constructing a classifier. In the training phase, the training set is used to decide how the features ought to be weighted and combined in order to separate the various classes of objects. In the application phase, the weights determined in the training set are applied to a set of objects that do not have known classes in order to determine what their classes are likely to be. A closer look into those two phases follows at the next section.

4.1.2 The classification and evaluation of its performance

The data set described in the previous section (Table 6, Annex) is used for the training of several supervised classification machine learning algorithms. Machine learning research tends to focus on improved predictive accuracy. Accuracy is of primary concern for all applications of machine learning and is easily measured by having training, test and validation datasets (Quinlan, 1996).

Hence, in addition to our test dataset, optimally, there should also be a validation set, which will be used to evaluate the model’s accuracy. Having a training and validation datasets helps to ensure that the model is optimized and it does not underfit or overfit the data. A model is *under-fitting* when it performs poorly on the training data (model is too simple for the data). That would be the case when the model is unable to capture the relationship between the input examples and the target labels. In contrast, a model is *over-fitting* when it

performs well on the training data but does not perform well on the test data (model is too complex for the data). That would be the case when the model is memorizing the data it has seen and is unable to generalize on unseen examples. Therefore, it is of specific importance first that the model is optimized and evaluated, using validation techniques, and second that a test set is used to estimate the prediction performance on unseen instances.

Our dataset serves as the training set and it is also used to optimize the dataset (using 10fold validation method), however there are not enough instances to create a test set and determine how well the classifiers perform on unseen instances. Therefore, there is a substantial chance that our models are overfitted. Figure 25 shows that the classifiers selected performs quite well although they have not been tested for performance on unseen instances.

The classifiers selected were trained using Weka. In Weka, the data passed through two phases, namely pre-processing and classification. The pre-processing consisted of choosing a filter which appointed the label attribute in the dataset as the label to be used for classification. Namely, we chose as an unsupervised filter the attribute ClassAssigner, which saved the label attribute as the classifying instance. Since we have only 60 instances in our dataset, there is great danger of overfitting. Taking this into account, as part of the pre-processing, the model evaluation method of a cross-validation was applied. Cross-validation can be of different types, however for our purposes we used a 10-fold cross validation check in order to determine how well each classifier generalizes to new data. When the classifier is trained, instead of using the entire dataset, a part of the data is removed before the training begins. Once the training is completed, the part of the dataset which was removed is used to test the performance of the learned model on *new* data.

The data set is divided into 10 subsets, and the holdout method is repeated [...] [10] times. Each time, one of the [...] [10] subsets is used as the test set and the other [...] [9] subsets are put together to form a training set. Then the average error across all [...] [10] trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set [...] [9] times (Schneider, 1997).

The classification phase consisted of selecting several classifiers and adjusting their parameters in order to find the best accuracy without overfitting. Typically, classification rules induced by machine learning systems are judged on the basis of two criteria: their classification accuracy on an independent test set and their complexity (Holte, 1993). For our purposes, we are mainly interested in the accuracy achieved by each classifier. When it comes to complexity, the models selected are fairly ‘simple’ – hence they are very transparent on how they make their prediction. The classifiers selected were: a baseline model, a single attribute model, a decision tree and a k-nearest neighbour model. For each one of those classifiers (except the

baseline) there are specific attributes which were tested in order to determine what is their impact on the ultimate accuracy.

A must for classification tasks is to, first, run a base-line (ZeroR) model which will show the simplicity (or complexity) of the dataset. ZeroR is the simplest classification method which relies only on the label and ignores all predictors (features) - it simply predicts the majority category (class). In our data set there is a majority of incompatible labels, hence when the ZeroR model constructs a frequency table for our dataset label (incompatibility/compatibility) it will select its most frequent value – incompatible. Because most of the instances are labelled as incompatible, by simply putting this label to the whole dataset, the model ZeroR has an accuracy of 63% percent. Such an accuracy is marginally high and translates into the need for the dataset to have not only more examples but and specifically more compatibility examples. Moreover, those 63% are our baseline – this classifier has no predictability power but it is useful to determine the baseline performance which will be the benchmark for the other classifiers trained. Therefore, any other classifier must perform better than the baseline (accuracy higher than 63%), otherwise it would be clear that there is underfitting²⁴.

Figure 25 presents the accuracy of seven classifiers. Each one of them achieved better accuracy in comparison to the baseline, hence there is no underfitting. Nevertheless, the consistent accuracy of above 80% for each model indicates a potential overfitting to the dataset even though there was a 10-cross-validation applied. This, however, is a typical feature for dataset with an insufficient number of instances.

Dataset	(1) rules.ZeroR ''	(2) rules.OneR '	(3) lazy.IBk '-K	(4) lazy.IBk '-K	(5) trees.J48 '-	(6) lazy.IBk '-K	(7) lazy.IBk '-K	(8) lazy.IBk '-K	
DataSet_Compatibility	(100)	63.33(6.70)	86.67(12.76) v	83.67(12.07) v	81.33(11.91) v	86.50(12.69) v	82.17(12.59) v	82.17(12.59) v	82.00(12.69) v
		(v/ /*)	(1/0/0)	(1/0/0)	(1/0/0)	(1/0/0)	(1/0/0)	(1/0/0)	(1/0/0)

Key:

(1) rules.ZeroR '' 40055541465867954
(2) rules.OneR '-B 6' -3459427003147861443
(3) lazy.IBk '-K 3 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\\\" \" -3080186098777067172
(4) lazy.IBk '-K 5 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\\\" \" -3080186098777067172
(5) trees.J48 '-C 0.25 -M 2' -217733168393644444
(6) lazy.IBk '-K 3 -W 0 -X -A \"weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\\\" \" -3080186098777067172
(7) lazy.IBk '-K 5 -W 0 -X -A \"weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\\\" \" -3080186098777067172
(8) lazy.IBk '-K 11 -W 0 -X -A \"weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\\\" \" -3080186098777067172

Figure 25: An initial selection of classifiers and their results

²⁴ <http://chem-eng.utoronto.ca/~datamining/dmc/zeror.htm>

At the bottom of Figure 25 each of the classifiers run is described, including their parameters. For some classifiers (OneR) no parameters needed to be adjusted, however for other (such a k-nn and decision trees) the right parameters' settings can be key for the best accuracy score.

OneR is a single attribute model or one-level decision tree, generating one rule for each predictor in the data, then selecting the rule with the smallest total error as its "one rule". OneR have been recognized as producing rules (predicting) only slightly less accurate than state-of-the-art classification algorithms while producing rules that are simple for humans to interpret (saedsayad, sd; Holte, 1993). For this classifier no parameter adjustments were conducted. This was in contrast to the other classification models – decision trees and k-nn.

Among the data-driven methods, trees are the most transparent and easy to interpret. Trees are based on separating observations into subgroups by creating splits on predictors. Those splits create logical rules which are transparent and easily understandable (as already described in Chapter 3). The resulting subgroups should be more homogenous in terms of the outcome variable, thereby creating useful prediction or classification rules (Shmueli, et al., 2008). Decision trees can predict better when pruning is applied. Pruning reduces the size of the decision tree by removing sections which provide little power to classify instances. Overall, pruning reduces the complexity of the final classifier, hence improving the accuracy by reduction of overfitting. The default parameter of the decision trees model, in Weka, is 0.25. Although smaller value incur more pruning, any pruning below 0.25 did not improve the accuracy of the model. Moreover, tuning other parameters, such as the minimum number of instances per leaf (set to 1) and the basic technique for smoothing probability estimates – Laplace (set to True). However, those adjustments did not yield higher accuracy either. Hence, none of the parameters tuning helped to achieve a better accuracy in comparison to the originally set parameters and their performance. An similar trait was observed for the last type of classifier trained.

The last simple classifier trained was the k-nearest-neighbours method. Nearest neighbour methods have the advantage of being easy to implement. They also give good results if the features are chosen carefully and are weighted carefully in the computation of the distance. This is specifically important for value of k . If k is a small number (1 or 3) it may be able to classify very accurately, especially if there are more than two labels. Instead, it would be better that k equals 5 or 7, so that the model has enough neighbours to correctly classify the unknown task. Nevertheless, for our case since there are only two labels (incompatible or compatible) the k-NN model did not perform better (even performed worse) when k would be set to a number higher than 3. Moreover, selecting cross validation for the second time also did not add any additional predictive power to the model.

Overall, it was observed that adjusting some of the parameters did not lead to better performance. After selecting several different options for the parameters and comparing the results, it was concluded that the lowest performing models are not of value, hence were removed. A complete overview of the experiments run can be found in the Annex, and the final results are presented in Figure 26.

Dataset	(1) rules.ZeroR ''	(2) rules.OneR '	(3) trees.J48 '-	(4) lazy.IBk '-K	(5) lazy.IBk '-K
DataSet_Compatibility	(100) 63.33(6.70)	86.67(12.76) v	86.50(12.69) v	84.33(12.49) v	83.67(12.07) v
	(v/ /*)	(1/0/0)	(1/0/0)	(1/0/0)	(1/0/0)

Key:

```
(1) rules.ZeroR '' 48055541465867954
(2) rules.OneR '-B 6' -3459427003147861443
(3) trees.J48 '-C 0.25 -M 2' -217733168393644444
(4) lazy.IBk '-K 1 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.EuclideanDistance -R first-la:
(5) lazy.IBk '-K 3 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.EuclideanDistance -R first-la:
```

Figure 26: Results predicting (in)compatibility of further processing of personal data

The results are very promising. Even though the data set has a very limited set of instances – 60, the classifiers trained performed significantly better than the baseline. Although there is a slight sign of overfitting, the accuracy is not 100% which indicates that if more data instances are to be provided the danger of overfitting will be avoided.

The most accurate classifiers are the OneR and the simple decision tree (C4) using their default parameters²⁵. The difference between the performance of the two is marginal. This is particularly interesting, since the two classifiers reach their predictions in a different way. While OneR ranks attributes according to their error rate on the training set, decision trees (C4) calculate the entropy for all measures and select the most informative ones (Holte, 1993). OneRules are usually a little less accurate than C4's pruned decision trees, although this is not the case for our dataset. C4's trees are not larger in terms of the number of attributes measured to classify the average example, which means that if the OneRule is performing better there is a chance for overfitting the data. Thus, OneRule can be used as a benchmark - giving a reasonable estimate of how one learning system would compare with others. “If a complex rule is induced, its additional complexity must be justified by its being correspondingly more accurate than a simple rule” (Holte, 1993). This indicates that if none of our more complex algorithms are outperforming the OneR model then the dataset and the features should be a subject of review.

Is the outcome useful? This outcome is both very useful and in the same time not practically useful. It does not help the millions of data controllers to identify whether they are compliant with the one of the most

²⁵ confidenceFactor is 0.25; Minimal number of Observations is 2 and Laplace smoothing is set to False.

essential data protection principles – purpose limitation. It also does not help authorities into creating a uniform method on how to interpret (in)compatibility across the EU. Most of all, the outcome does not help the data subjects to easily test whether the new purpose for which that company asks their consent fits within the previous purposes of processing their personal data. Nevertheless, this outcome is very useful as it, first, created a dataset of (in)compatible cases and second, it defined features on the basis of which machine learning models can be trained. Moreover, the classification results indicate the potential benefit of simply having to answer some questions and get a prediction on whether a further processing is compatible or not. Despite that, the feature selection and the classifiers parameters' tuning can and should be improved.

4.2 Discussion and Conclusion

When I made a purchase account at Zalando I was not expecting to learn, several months later, that Zalando had shared my email address with Facebook. My personal data was collected for a set of initial purposes which were specifically and explicitly provided by Zalando and based on the legal ground of performance of a contract. However, the sharing of my personal data with a third party, especially matching my Zalando details with a potential social media account of mine, is further processing of my personal data. According to the purpose limitation principle, as laid down in Article 5(1) GDPR, my personal data should not be further processed in a manner that is incompatible with the collection purposes.

In order to address the issues of what exactly is an incompatible further processing of personal data and how it is to be detected, a comprehensive and extensive literature review was conducted. The ultimate goal of the literature review was to enable the automation of Article 6(4) GDPR. During the literature review, a number of challenges, both in general for legal concepts and specifically for Article 6(4) were identified. Despite the presence of many methods of legal knowledge engineering, meeting all challenges of Article 6(4) GDPR with one method is not possible at this point in time. Instead of focusing on solving each of the challenges presented, a more simplistic approach was adopted – from the body of knowledge a dataset with (in)compatibility cases was created. This dataset was then used to train several supervised machine learning classifiers.

The prediction results from the classifiers were very promising and indicate the benefit of simply having to answer a limited number of questions and get a prediction on whether a further processing is compatible or not. Nevertheless, it must be acknowledged that the domain is unfortunate – it is highly complicated and there is not enough information available to build a legal knowledge engineering tool which can meet both the needs and requirements of Article 6(4) GDPR. One source of information which can accelerate the creation of such a tool is academic research. Despite the abundant source of scholarly criticism on the purpose limitation principle, no research has so far focused on analysing the formal and/or substantive assessments for further processing of personal data, as demonstrated in Chapter 2.

The need for a unitary method is essential to the proper functioning of the purpose limitation principle as acknowledged by WP29: "If the assessment were to be made case by case without any further guidance, this would risk inconsistent application and lack of predictability, as it has been the case in the past" (WP29, 2018). Although some guidance has been provided so far (an overview is available in Chapter 2), the core of this research (creating a comprehensive body of knowledge about and automating Article 6(4) GDPR) should be a subject of peer review.

It could be that in the near future, when the currently pending cases in front of the CJEU are judged upon, there will be a better understanding on how to frame Article 6(4)'s body and knowledge and which methods

of knowledge engineering are most appropriate to meet its needs. Nevertheless, this work can serve as the foundation to lay out additional machine learning methods, to create better features and enable the ultimate compliance with the purpose limitation principle.

Based on the challenges identified so far, it can be stated that computer programs cannot fully automate Article 6(4). Papanikolaou, et al. (2011) even pointed out that “[i]t is unreasonable to expect” computer programs to do so. This observation, however, has not prevented researchers from both legal and other science domains to explore a variety of techniques to analyse, interpret and extract information from legal texts, as there have been various attempts at applying such techniques in the context of privacy, in spite of the challenges and limitations of the legal knowledge engineering (Papanikolaou, et al., 2011, p. 167). Instead on aiming to fully automate Article 6(4) GDPR, an alternative method – building a dataset on the basis of which a classifier can predict the outcome of a further processing of personal data, is the ultimate result from the creation of a body of knowledge regarding Article 6(4). The dataset and the prediction accuracy achieved (86%) are a good start, which will be a subject of discussion in future research.

Annex

Table 5: Cases used for the (in)compatibility dataset

Case #	Description
Case 1	WP29, Example 1: Chatty receptionist caught on CCTV
Case 2	WP20, Example 2: Breathalyser checks working hours
Case 3	WP29, Example 3: Security clearance certificates stored to evidence and audit departmental compliance
Case 4	WP29, Example 4; 'Get Well Quick' breaks
Case 5	WP29, Example 5: A public-private partnership: lovers of fatty food told to eat less - part 1, analyse customer's data
Case 6	WP29, Example 5: A public-private partnership: lovers of fatty food told to eat less - 3rd party sends out leaflets
Case 7	WP29, Example 5: A public-private partnership: lovers of fatty food told to eat less - valid consent obtained to analyse the data
Case 8	WP29, Example 5: A public-private partnership: lovers of fatty food told to eat less - valid consent obtained to transfer the data to a 3rd party.
Case 9	WP29, Example 6: Safe internet training for children
Case 10	WP29, Example 7: Consent for use of holidays photographs to promote a website - no valid consent
Case 11	WP29, Example 7: Consent for use of holidays photographs to promote a website - valid consent
Case 12	WP29, Example 8: Photo-sharing website changes privacy policy
Case 13	WP29, Example 9: Secret algorithms predict pregnancy of customers from purchasing habits
Case 14	WP29, Example 10: Special offer for a lawnmower- not sensitive data
Case 15	WP29, Example 10: Special offer for a lawnmower - if to be sensitive data
Case 16	WP29, Example 11: Car manufacturer uses public vehicles registry data to notify car owners of malfunction and recall the cars
Case 17	WP29, Example 12: Transfer of results of pre-employment medical examination- incompatible
Case 18	WP29, Example 12: Transfer of results of pre-employment medical examination - informed consent and only positive medical results
Case 19	WP 29, Example 13: Housing Department needs access to data for fire protection - no clear communication to data subject
Case 20	WP 29, Example 13: Housing Department needs access to data for fire protection - clear communication to the data subjects, give them reasonable time to act

Case 21	WP29, Example 14: Victims of rape - no anonymization
Case 22	WP29, Example 14: Victims of rape - with irreversible anonymization
Case 23	WP29, Example 15: Mobile phone locations help inform traffic calming measures
Case 24	WP29, Example 16: Patients vouching for an alternative medical practitioner
Case 25	WP29, Example 17: Data Retention Directive
Case 26	WP29, Example 18: Fingerprints of asylum seekers used for law enforcement purposes
Case 27	WP29, Example 19: passenger name records ('PNR')
Case 28	WP29, Example 20: Smart metering data used for tax purposes and to detect indoor cannabis factories
Case 29	WP29, Example 21: Smart metering data mined to detect fraudulent energy use
Case 30	WP29, Example 22: Transactions in EU climate change registry used to detect VAT fraud
Case 31	CNIL, Facebook violation - systematic collection of personal data on 3rd party websites
Case 32	CNIL, Facebook violation - combining all personal data of customers to display targeted ads. - sensitive data
Case 33	CNIL, Facebook violation - combining all personal data of customers to display targeted ads. - no sensitive data
Case 34	Dutch DPA, Microsoft - processing purposes
Case 35	German state office for Data Protection, General data warehouses - no sensitive data, no anonymization
Case 36	German state office for Data Protection, General data warehouses - yes sensitive data, no anonymization
Case 37	German state office for Data Protection, General data warehouses - no personal data, yes anonymization
Case 38	Data protection commissioner of Ireland, bank card information collected for a specific transaction. - further processing without consent
Case 39	Data protection commissioner of Ireland, bank card information collected for a specific transaction. - further processing with consent
Case 40	Data protection commissioner of Ireland, telephone providers continue processing personal data of the data subject, without their consent
Case 41	Data protection commissioner of Ireland, telephone providers continue processing personal data of the data subject, with their consent
Case 42	Data protection commissioner of Ireland, telephone providers continue processing personal data of the data subject, but anonymized
Case 43	Data protection commissioner of Ireland, telephone providers continue processing personal data

	of the data subject, but anonymized with negative consequences
Case 44	EDPS advice, creditors personal data processing - adheres to the purpose specification
Case 45	EDPS advice, creditors personal data processing - does not adhere to the purpose specification
Case 46	EDPS, use of personal data originating from an access security system or a time management system for investigate purposes - data subject informed
Case 47	EDPS, use of personal data originating from an access security system or a time management system for investigate purposes - data subject not informed
Case 48	Supervisory Body of Europol, access and use of VIS data - for a specific task
Case 49	Supervisory Body of Europol, access and use of VIS data - without a specific task
Case 50	ECtHR, copying of documents containing banking data and their subsequent storage by local authorities - for a purpose, which is specific and adhered to, without collecting more than what is necessary
Case 51	ECtHR, copying of documents containing banking data and their subsequent storage by local authorities - without a purpose
Case 52	ECtHR, one's personal life is intruded upon by a systematic surveillance or transfer of personal data with the intention to realize negative actions against that individual
Case 53	ECtHR, unrestricted monitoring of one's correspondence, although permitted by local bankruptcy law
Case 54	ECtHR, injuries to an applicant's reputation if they were caused by a systematic collection and storing of 'false' personal data
Case 55	ECtHR, obligation towards private companies to provide tax auditors with access to individuals' personal data without a concrete and specific reason
Case 56	CJEU, have search engine results about a data subject altered even though the information was true and lawfully published by third parties - the interference with the subject's fundamental rights was not be justified by the overriding interest of the general public; and the search engine had made the data ubiquitous
Case 57	CJEU, have search engine results about a data subject altered even though the information was true and lawfully published by third parties - an interference with the subject's fundamental rights would be justified by the overriding interest of the general public; or if the search engine had not made the results accessible to anyone
Case 58	CJEU, C-201/14 (Bara case) a processing of personal data when the data subject was not informed about it & not legal ground
Case 59	CJEU Article 7(f) DPD allows for a strictly necessary (further) processing of personal data in order to realize a third party's legitimate interests (Case C-13/16, 2016)
Case 60	CJEU, store citizens' telecommunications data in order for police and security agencies to be able

	to request such data under the general interest of fight against serious crime and public security
--	--

Table 6: Names features and their corresponding values

Feature	Full description
collect_purpose_different	The collection purpose is different from the further purpose
ds_informed	The data subject is informed about the further processing purposes and is given an opportunity to object.
power_imbalance	There is a power imbalance between the data controller and data subject.
negative_impact	The further processing will have a negative impact on the individual.
positive_impact	The further processing will have a positive impact on the individual.
new_lg_compensate	New legal ground to compensate for the change of purpose.
conseq_foreseable_com	The consequences of the further processing are foreseeable and are communicated clearly to the data subject.
collect_purp_lawful	The processing of the personal data is based on a valid legal ground
sensitive_data	The further processing will involve the processing of sensitive personal data
coll_purp_legal_oblig	The collection purpose is based on the legal ground of compliance with a legal obligation
same_dcontroller	The data controller is the same for the collection and every other further purpose
anonymized	The personal data to be further processed is irreversibly anonymized
ds_reasonable_expect	The data subject can reasonably expect the further purpose

Table 7: Data set compatibility

collect_purpose_different	collect_purp_lawful	coll_purp_legal_oblig	ds_reasonable_expect	ds_informed	conseq_foreseable_com	power_imbalance	negative_impact	positive_impact	new_lg_compensate	sensitive_data	same_dcontroller	anonymization	outcome
yes	yes	no	no	no	no	yes	yes	no	no	no	yes	no	incompatible
yes	yes	yes	no	no	no	yes	yes	no	no	yes	yes	no	incompatible
no	yes	yes	yes	yes	yes	yes	yes	yes	no	yes	yes	no	compatible

yes	yes	no	no	no	no	yes	no	yes	no	yes	no	no	incompatible
yes	yes	no	no	no	no	no	yes	no	no	yes	yes	no	incompatible
yes	yes	no	no	no	no	no	yes	no	no	yes	no	no	incompatible
yes	yes	no	yes	no	no	no	yes	no	yes	yes	yes	no	compatible
yes	yes	no	yes	no	no	no	yes	no	yes	yes	no	no	compatible
yes	yes	no	yes	no	no	no	no	yes	no	yes	no	no	incompatible
yes	yes	no	no	no	no	no	yes	no	no	yes	yes	no	incompatible
yes	yes	no	no	no	no	yes	yes	no	yes	yes	yes	no	compatible
yes	no	no	no	yes	no	yes	yes	no	no	yes	yes	no	incompatible
yes	yes	no	no	no	no	no	yes	no	no	yes	yes	no	incompatible
no	yes	no	yes	yes	yes	no	no	yes	yes	no	yes	no	compatible

no	yes	no	yes	yes	yes	no	no	yes	yes	yes	yes	no	com pati ble
yes	yes	yes	yes	yes	yes	no	no	yes	no	no	no	no	com pati ble
yes	no	no	no	no	no	yes	yes	no	no	yes	no	no	inco mpa tible
yes	no	no	yes	yes	yes	yes	no	yes	yes	yes	no	no	com pati ble
yes	yes	no	no	no	no	yes	yes	yes	no	yes	no	no	inco mpa tible
yes	yes	no	yes	yes	yes	yes	yes	yes	yes	yes	no	no	com pati ble
yes	yes	no	yes	no	no	yes	yes	no	no	yes	no	no	inco mpa tible
yes	yes	no	yes	no	no	no	no	no	no	no	no	yes	com pati ble
yes	yes	no	no	no	no	yes	yes	no	no	yes	no	yes	com pati ble
yes	yes	no	no	no	no	no	yes	no	no	yes	no	no	inco mpa tible
yes	yes	no	no	no	no	yes	yes	no	no	yes	no	no	inco mpa tible

yes	yes	yes	no	no	no	yes	yes	no	no	yes	no	no	incompatible
yes	yes	no	no	no	no	yes	yes	no	no	yes	no	no	incompatible
yes	yes	no	no	no	no	yes	yes	no	no	yes	no	no	incompatible
yes	yes	no	yes	yes	yes	no	yes	no	no	yes	yes	no	compatible
yes	yes	no	no	no	no	yes	yes	no	no	no	no	no	incompatible
yes	yes	no	no	no	no	no	no	no	no	yes	yes	no	incompatible
yes	yes	no	no	no	no	no	yes	no	no	yes	yes	no	incompatible
yes	yes	no	no	no	no	no	yes	no	no	no	yes	no	incompatible
yes	no	no	no	no	no	yes	yes	no	no	no	yes	no	incompatible
yes	yes	no	no	no	no	no	no	no	no	no	no	no	incompatible
yes	yes	no	no	no	no	no	no	no	no	yes	no	no	incompatible

yes	yes	no	no	no	no	no	no	no	no	no	no	yes	com pati ble
yes	yes	no	no	yes	yes	no	no	no	no	yes	yes	no	inco mpa tible
yes	yes	no	yes	yes	yes	yes	no	yes	yes	yes	yes	no	com pati ble
yes	yes	no	no	no	no	no	no	no	no	no	yes	no	inco mpa tible
yes	yes	no	yes	yes	yes	no	no	no	yes	no	yes	no	com pati ble
yes	yes	no	no	no	no	no	no	no	no	no	yes	yes	com pati ble
yes	yes	no	no	no	no	no	yes	no	no	no	yes	yes	inco mpa tible
yes	yes	yes	yes	yes	yes	no	no	no	yes	yes	no	no	com pati ble
yes	no	yes	no	yes	yes	no	no	no	yes	yes	no	no	inco mpa tible
yes	yes	no	yes	yes	yes	yes	yes	no	no	no	yes	no	com pati ble
yes	no	no	no	no	no	yes	yes	no	no	no	yes	no	inco mpa tible

yes	yes	no	no	no	no	yes	yes	no	yes	yes	no	no	com pati ble
yes	yes	no	no	no	no	yes	yes	no	no	yes	no	no	inco mpa tible
yes	yes	no	yes	yes	yes	yes	yes	no	yes	yes	no	no	com pati ble
yes	yes	no	no	no	no	yes	yes	no	no	yes	no	no	inco mpa tible
yes	yes	no	no	no	no	yes	yes	no	no	no	yes	no	inco mpa tible
yes	yes	no	no	no	no	yes	yes	no	yes	yes	no	no	inco mpa tible
yes	no	no	no	no	no	yes	yes	no	yes	yes	no	no	inco mpa tible
yes	yes	no	no	no	no	yes	yes	no	yes	yes	no	no	inco mpa tible
yes	yes	yes	no	yes	no	yes	yes	no	yes	yes	no	no	inco mpa tible
yes	no	yes	no	yes	no	yes	yes	no	yes	yes	no	no	com pati ble
yes	yes	yes	no	no	no	yes	yes	no	no	no	no	no	inco mpa tible

yes	yes	yes	yes	no	yes	no	no	no	yes	no	no	no	compatible
yes	yes	no	no	no	no	yes	yes	no	no	yes	no	no	incompatible

Figure 27: Test one

```

Dataset          (1) rules.ZeroR '' | (2) rules.OneR ' (3) trees.J48 '- (4) trees.J48 '- (5) lazy.IBk '-K (6) lazy.IBk '-K (7) lazy.IBk '-K (8) lazy.IBk '-K (9) lazy.IBk '-K (10) lazy.IBk '- (11) lazy.IBk '- (12) lazy.IBk '- (13) lazy.IBk '- (14) lazy.IBk '-
-----
DataSet_Compatibility (100) 63.33(6.70) | 86.67(12.76) v 86.50(12.69) v 80.50(15.18) v 84.33(12.49) v 83.67(12.07) v 81.33(11.91) v 80.00(12.31) v 81.17(12.68) v 84.33(12.49) v 82.17(12.59) v 82.17(12.59) v 82.17(12.59) v 82.00(12.69) v
-----
(v/ /*) | (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0)

Key:
(1) rules.ZeroR '' 48055541465867954
(2) rules.OneR '-B 6' -3459427003147861443
(3) trees.J48 '-C 0.25 -M 2' -217733168393644444
(4) trees.J48 '-C 1.0 -M 1 -A' -217733168393644444
(5) lazy.IBk '-K 1 -W 0 -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172
(6) lazy.IBk '-K 3 -W 0 -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172
(7) lazy.IBk '-K 5 -W 0 -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172
(8) lazy.IBk '-K 7 -W 0 -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172
(9) lazy.IBk '-K 11 -W 0 -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172
(10) lazy.IBk '-K 1 -W 0 -X -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172
(11) lazy.IBk '-K 3 -W 0 -X -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172
(12) lazy.IBk '-K 5 -W 0 -X -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172
(13) lazy.IBk '-K 7 -W 0 -X -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172
(14) lazy.IBk '-K 11 -W 0 -X -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172

```

Figure 28: Test two

```

Dataset          (1) rules.ZeroR '' | (2) rules.OneR ' (3) trees.J48 '- (4) lazy.IBk '-K (5) lazy.IBk '-K (6) lazy.IBk '-K (7) lazy.IBk '-K (8) lazy.IBk '-K (9) lazy.IBk '-K
-----
DataSet_Compatibility (100) 63.33(6.70) | 86.67(12.76) v 86.50(12.69) v 84.33(12.49) v 83.67(12.07) v 81.33(11.91) v 84.33(12.49) v 82.17(12.59) v 82.17(12.59) v
-----
(v/ /*) | (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0)

Key:
(1) rules.ZeroR '' 48055541465867954
(2) rules.OneR '-B 6' -3459427003147861443
(3) trees.J48 '-C 0.25 -M 2' -217733168393644444
(4) lazy.IBk '-K 1 -W 0 -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172
(5) lazy.IBk '-K 3 -W 0 -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172
(6) lazy.IBk '-K 5 -W 0 -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172
(7) lazy.IBk '-K 7 -W 0 -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172
(8) lazy.IBk '-K 9 -W 0 -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172
(9) lazy.IBk '-K 11 -W 0 -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172

```

Figure 29: Test three

```

Dataset          (1) rules.ZeroR '' | (2) rules.OneR ' (3) trees.J48 '- (4) lazy.IBk '-K (5) lazy.IBk '-K (6) trees.J48 '- (7) trees.J48 '- (8) trees.J48 '-
-----
DataSet_Compatibility (100) 63.33(6.70) | 86.67(12.76) v 86.50(12.69) v 84.33(12.49) v 83.67(12.07) v 83.67(12.97) v 84.17(12.62) v 83.67(12.97) v
-----
(v/ /*) | (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0) (1/0/0)

Key:
(1) rules.ZeroR '' 48055541465867954
(2) rules.OneR '-B 6' -3459427003147861443
(3) trees.J48 '-C 0.25 -M 2' -217733168393644444
(4) lazy.IBk '-K 1 -W 0 -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172
(5) lazy.IBk '-K 3 -W 0 -A '\weka.core.neighboursearch.LinearNNSearch -A \\\weka.core.EuclideanDistance -R first-last\\\'\'\' -3080186098777067172
(6) trees.J48 '-C 0.1 -M 2' -217733168393644444
(7) trees.J48 '-C 0.1 -M 1 -A' -217733168393644444
(8) trees.J48 '-C 0.1 -M 2 -A' -217733168393644444

```

Figure 30: Test four

Dataset	(1) rules.ZeroR ''	(2) rules.OneR '	(3) trees.J48 '-	(4) lazy.IBk '-K	(5) lazy.IBk '-K	(6) lazy.IBk '-K	(7) lazy.IBk '-K	
DataSet_Compatibility	(100)	63.33(6.70)	86.67(12.76) v	86.50(12.69) v	84.33(12.49) v	83.67(12.07) v	82.17(12.59) v	82.17(12.59) v
		(v/ /*)	(1/0/0)	(1/0/0)	(1/0/0)	(1/0/0)	(1/0/0)	(1/0/0)

Key:

- (1) rules.ZeroR '' 48055541465867954
- (2) rules.OneR '-B 6' -3459427003147861443
- (3) trees.J48 '-C 0.25 -M 2' -217733168393644444
- (4) lazy.IBk '-K 1 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\" \"\" -3080186098777067172
- (5) lazy.IBk '-K 3 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\" \"\" -3080186098777067172
- (6) lazy.IBk '-K 3 -W 0 -X -A \"weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\" \"\" -3080186098777067172
- (7) lazy.IBk '-K 5 -W 0 -X -A \"weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\" \"\" -3080186098777067172

Bibliography

Alchourrón, C. E. & Bulygin, E., 1981. The expressive conception of norms. *New studies in deontic logic*, Springer, Dordrecht, pp. 95-124.

Allen, L. E. & Engholm, C. R., 1977. Normalized legal drafting and the query method. *J. Legal Educ.*, Volume 29, p. 380.

Alvarez, S. A., 2017. *Decision Tree Pruning based on Confidence Intervals (as in C4.5)*. [Online] Available at: <http://www.cs.bc.edu/~alvarez/ML/statPruning.html> [Accessed 27 July 2017].

Anon., 1998. *Linear Regression, Statistics 101, Yale*. [Online] Available at: <http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm> [Accessed 11 July 2018].

Arthurs, H. W., 1983. Law and learning: report to the Social Sciences and Humanities Research Council of Canada by the Consultative Group on Research and Education in Law. *Social Sciences and Humanities Research Council of Canada*.

Ashley, K. D., 2017. *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. 1st ed. Cambridge: Cambridge University Press.

Autoriteit Persoonsgegevens, 2017. *Dutch DPA: Microsoft breaches data protection law with Windows 10*. [Online] Available at: <https://autoriteitpersoonsgegevens.nl/en/news/dutch-dpa-microsoft-breaches-data-protection-law-windows-10> [Accessed 01 June 2018].

Bartolini, C., Giurgiu, A., Lenzini, G. & Robaldo, L., 2016. Towards legal compliance by correlating Standards and Laws with a semi-automated methodology. *Springer, In Benelux Conference on Artificial Intelligence*, pp. 47-62.

Bartolini, C., Muthuri, R. & Santos, C., 2015. *Using Ontologies to Model Data Protection Requirements in Workflows*. [Online] Available at: https://link.springer.com/chapter/10.1007/978-3-319-50953-2_17 [Accessed 13 6 2018].

Bendiek, A. & Schmieg, E., 2016. *Bendiek, Annegret, and Evita Schmie European Union data protection and external trade: having the best of both worlds?*. [Online] Available at: <http://nbn-resolving.de/urn:nbn:de:0168-ssoar-464328> [Accessed 01 February 2018].

Bird&Bird, 2016. *Guide to the General Data*. [Online] Available at: <https://www.twobirds.com/~media/pdfs/gdpr-pdfs/bird--bird--guide-to-the-general-data-protection-regulation.pdf?la=en> [Accessed 1 November 2017].

- Boyd, D. & Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(no. 5), pp. 662-679.
- Branting, K. L., 2017. Data-centric and logic-based models for automated legal problem solving. *Artificial Intelligence and Law*, 25(no. 1), pp. 5-27.
- Breaux, T. D., Vail, M. W. & Anton, A. I., 2006. Towards regulatory compliance: Extracting rights and obligations to align requirements with regulations.. In *Requirements Engineering, 14th IEEE International Conference*, pp. 49-58. IEEE.
- Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A., 1984. *Classification and regression trees*. s.l.:CRC press.
- Breuker, J., Valente, A. & Winkels, R., 2004. Legal ontologies in knowledge engineering and information management. *Artificial intelligence and law*, Volume 12(4), pp. 241-277.
- Breukers, J. & Hoekstra, R., 2004. *Epistemology and ontology in core ontologies: FOLaw and LRI-Core, two core ontologies for law*. [Online]
Available at: <http://ceur-ws.org/vol-118/paper2.pdf>
[Accessed 17 7 2018].
- Bronshtein, A., 2017. *A Quick Introduction to K-Nearest Neighbors Algorithm*. [Online]
Available at: <https://medium.com/@adi.bronshtein/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7>
[Accessed 12 June 2018].
- Brownlee, J., 2016. *Supervised and Unsupervised Machine Learning Algorithms*. [Online]
Available at: <https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/>
[Accessed 11 July 2018].
- Buttarelli, G., 2016. *Ethics at the Root of Privacy*. [Online]
Available at:
https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/EDPS/Publications/Speeches/2016/16-04-19_MIT_Ethics1_EN.pdf
[Accessed 11 01 2017].
- Butti, E., 2013. *The Roles and Relationship between the Two European Courts in Post-Lisbon EU Human Rights Protection*. [Online]
Available at: <http://www.jurist.org/dateline/2013/09/elena-butti-lisbon-treaty.php>
[Accessed 13 Februari 2017].
- C-131/12 (2014) CJEU.
- C-201/14, *Bara and Others* (2015) EU:C:2015:638.
- C-207/16 (2016) CJEU.
- C-682/15 (*Berlioz Investment Fund*) (2017) CJEU.

- Cappelli, A., Lenzi, V. B., Sprugnoli, R. & Biagioli, C., 2007. Modelization of domain concepts extracted from the Italian privacy legislation. *In Proceedings of the Workshop on Computational Semantics (IWCS-7)*.
- Case C-13/16 (2016) CJEU.
- Casellas, N. et al., 2010. Ontological Semantics for Data Privacy Compliance: The NEURONA Project. *In AAAI Spring Symposium: Intelligent Information Privacy Management*.
- Cate, F. H., 1994. The EU data protection directive, information privacy, and the public interest. *Iowa L. Rev.*, 80(431).
- Cate, F. H., 2006. The failure of fair information practice principles. *SSRN*.
- Chander, A., 2017. The Racist algorithm?. *115 Mich. L. Rev.* 1023 .
- Chynoweth, P., 2008. Legal research. *Advanced Research Methods in the Built Environment*, Wiley-Blackwell, Oxford, pp. 28-38.
- CJEU, 2014. *The Court of Justice declares the Data Retention Directive to be invalid*. [Online] Available at: <https://curia.europa.eu/jcms/upload/docs/application/pdf/2014-04/cp140054en.pdf> [Accessed 27 May 2018].
- Claire, D., 2011. *J.Mc.B v. L.E.: The Intersection of European Union Law and Private International Law in Intra-European Union Child Abduction*. [Online] Available at: <http://ir.lawnet.fordham.edu/ilj/vol34/iss5/11> [Accessed 13 February 2017].
- CNIL, 2017. *Common Statement by the Contact Group of the Data Protection Authorities of The Netherlands, France, Spain, Hamburg and Belgium*. [Online] Available at: <https://www.cnil.fr/en/common-statement-contact-group-data-protection-authorities-netherlands-france-spain-hamburg-and> [Accessed 01 June 2018].
- Cuijpers, C., Purtova, N. & Kosta, E., 2014. Data Protection Reform and the Internet: The Draft Data Protection Regulation.
- Custers, B. & Uršič, H., 2016. Big data and data reuse: a taxonomy of data reuse for balancing big data benefits and personal data protection. *International Data Privacy Law*, Volume 6.1, pp. 4-15.
- Daniel, P., Bratley, P., Frémont, J. & Mackaay, E., 1993. Legal interpretation in expert systems. *In Proceedings of the 4th international conference on Artificial intelligence and law, ACM*, pp. 90-99.
- Data Protection Commissioner, I., 2008. *Guidance Note for Data Controllers on Purpose Limitation and Retention in relation to Credit/Debit/Charge card transactions*. [Online] Available at: <https://www.dataprotection.ie/docs/Guidance-Note-for-Data-Controllers-on-Purpose-Limitation-and-Retention/859.htm> [Accessed 27 May 2018].
- De Hert, P. & Gutwirth, S., 2009. Data protection in the case law of Strasbourg and Luxemburg: Constitutionalisation in action. *Springer Netherlands In Reinventing data protection?*, pp. 3-44.

De Hert, P. & Papakonstantinou, V., 2016. The new General Data Protection Regulation: Still a sound system for the protection of individuals?. *Computer Law & Security Review*, Volume 32.2, pp. 179-194.

Dietrich, A., Lockemann, P. C. & Raabe, O., 2007. Agent approach to online legal trade. In *Conceptual Modelling in Information Systems Engineering*. Springer, Berlin, Heidelberg, pp. 177-194.

EC [4], 2016. *Reform of EU data protection rules*. [Online]
Available at: http://ec.europa.eu/justice/data-protection/reform/index_en.htm
[Accessed 14 February 2017].

EC [6], 2015. *How will the EU's data protection reform strengthen the internal market?*. [Online]
Available at: http://ec.europa.eu/justice/data-protection/files/4_strengthen_2016_en.pdf
[Accessed 13 June 2017].

EC [7], 2017. *Protection of personal data*. [Online]
Available at: <http://ec.europa.eu/justice/data-protection/>
[Accessed 18 January 2017].

EC[9], 2016. *Opinions and recommendations*. [Online]
Available at: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/index_en.htm
[Accessed 13 September 2017].

EC, 2016. *The EU Data Protection Reform and Big Data: Factsheet, March 2016*. [Online]
Available at: http://ec.europa.eu/justice/data-protection/files/data-protection-big-data_factsheet_web_en.pdf
[Accessed 30 September 2016].

EC, 2017. *Protection of personal data*. [Online]
Available at: <http://ec.europa.eu/justice/data-protection/>
[Accessed 22 September 2017].

EC, 2018. *What are Data Protection Authorities (DPAs)?*. [Online]
Available at: https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-are-data-protection-authorities-dpas_en
[Accessed 03 June 2018].

ECtHR, 2017. *Guide on Article 8 of the European Convention on Human Rights*. [Online]
Available at: https://www.echr.coe.int/Documents/Guide_Art_8_ENG.pdf
[Accessed 05 June 2018].

EDPS, 2013. *Answer to a consultation under Article 46(d) on the use of data collected for a specific purpose to a different purpose (Case 2013-0279)*. [Online]
Available at: https://edps.europa.eu/sites/edp/files/publication/13-04-17_eib_use_of_data_en.pdf
[Accessed 25 May 2018].

EDPS, 2014. *Privacy and competitiveness in the age of big data: The interplay between data protection, competition law and consumer protection in the Digital Economy*. [Online]
Available at: https://edps.europa.eu/sites/edp/files/publication/14-03-26_competition_law_big_data_en.pdf
[Accessed 14 June 2017].

- EDPS, n.d. *Data Protection Legislation*. [Online]
Available at: <https://secure.edps.europa.eu/EDPSWEB/edps/EDPS/Dataprotection/QA/QA2>
[Accessed 6 October 2016].
- El Ghosh, M., Naja, H., Abdulrab, H. & Khalil, M., 2017. Ontology Learning Process as a Bottom-up Strategy for Building Domain-specific Ontology from Legal Texts. *ICAART*, Volume 2, pp. 473-480.
- El Ghosh, M., Naja, H., Abdulrab, H. & Khalil, M., 2017. Towards a Legal Rule-Based System Grounded on the Integration of Criminal Domain Ontology and Rules. *Procedia Computer Science*, Issue 112, pp. 632-642.
- europa.eu, 2018. *European Data Protection Supervisor (EDPS)*. [Online]
Available at: https://europa.eu/european-union/about-eu/institutions-bodies/european-data-protection-supervisor_en
[Accessed 03 June 2018].
- European Commission, 2016. *The EU Data Protection Reform and Big Data: Factsheet, March 2016*. [Online]
Available at: http://ec.europa.eu/justice/data-protection/files/data-protection-big-data_factsheet_web_en.pdf
[Accessed 30 September 2016].
- Fatema, K. et al., 2016. A Semi-Automated Methodology for Extracting access control rules from the European Data Protection Directive. In *Security and Privacy Workshops (SPW)*, pp. 25-32.
- Francesconi, E., Montemagni, S., Peters, W. & Tiscornia, D., 2010. *Semantic processing of legal texts: Where the language of law meets the law of language*. Vol. 6036 ed. s.l.:Springer.
- Gaur, S., Nguyen, H. H. V., Kashihara, K. & Baral, C., 2014. Translating simple legal text to formal representations. In *JSAI International Symposium on Artificial Intelligence; Springer Berlin Heidelberg*, pp. 259-273.
- GDPR, 2016. *General Data Protection Directive* [Interview] 2016.
- Gellman, R., 2017. Fair information practices: A basic history. *SSRN*.
- Giltrow, J. & Stein, D., 2017. *The Pragmatic Turn in Law: Inference and Interpretation in Legal Discourse*. s.l.:s.n.
- Han, J., Pei, J. & Kamber, M., 2011. *Data mining: concepts and techniques*. Elsevier.
- Hart, H. L. A., 1961. *The Concept of Law*. s.l.:Oxford University Press.
- Heck, P. R. & Krueger, J. I., 2016. Social Perception of Self-Enhancement Bias and Error. *Social Psychology*, , 47(6), pp. 327-339.
- Hevner, A. R., Ram, S. & March, S. T., 2004. Design science in information systems research. *Management Information systems quarterly*, 28(1), pp. 75-105.
- Hijmans, H., 2016. *The European Union as Guardian of Internet Privacy: The Story of Art 16 TFEU*. s.l.:Springer Vol. 31..

- Hollebeek, N. J., 2017. *Persoonsgegevens in Databanken: WBP Meldingenregister*. [Online] Available at: <https://www.hollebeek.nl/databanken/jurdaty.html> [Accessed 18 July 2018].
- Holte, R. C., 1993. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Machine Learning*, , 11(1), pp. 63-90.
- Hustinx, P., 2013. EU data protection law: The review of directive 95/46/EC and the proposed general data protection regulation. *Collected courses of the European University Institute's Academy of European Law*, Volume 24th Session on European Union Law, pp. 1-12.
- ICO, 2017. *Accountability and governance*. [Online] Available at: <https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/> [Accessed 20 January 2028].
- ICO, 2017. *Big data, artificial intelligence, machine learning and data protection*. [Online] Available at: <https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf> [Accessed 06 June 2018].
- Jasmontaite, L., 2016. *IMPLEMENTATION OF THE GDPR: THE EUROPEAN DATA PROTECTION BOARD*. [Online] Available at: <http://brusselsprivacyhub.eu/publications/ws02.html> [Accessed 11 November 2017].
- Joined Cases C-468/10 and C-469/10* (2011) CJEU, Reference for a preliminary ruling from the Tribunal Supremo.
- Karagiannis, ed., D., 1994. *Database and Expert Systems Applications: 5th International Conference, DEXA'94, Athens, Greece, September 7-9*,. Vol. 856 ed. Athens: Springer Science & Business Media.
- Katz, D. M., Bommarito, I. I., Michael, J. & Blackman, J., 2014. Predicting the behavior of the supreme court of the united states: A general approach. *arXiv preprint arXiv:1407.6333*..
- Kelleher, D., 2016. *In Breyer decision today, Europe's highest court rules on definition of personal data*. [Online] Available at: <https://iapp.org/news/a/in-breyer-decision-today-europes-highest-court-rules-on-definition-of-personal-data/> [Accessed 16 June 2017].
- Kiang, M. Y., 2003. A comparative assessment of classification methods. *Decision Support Systems*, 35(no. 4), pp. 441-454.
- Kim, M.-Y., Xu, Y. & Goebel, R., 2014. Legal question answering using ranking svm and syntactic/semantic similarity. *In JSAI International Symposium on Artificial Intelligence, Springer, Berlin, Heidelberg*, pp. 244-258.
- Kiriinya, R. K. M., 2015. "The Place of Legal Ontologies in Regulatory Compliance".
- Koning, M. E., 2015. Purpose Limitation.

- Koops, B.-J., 2014. The trouble with European data protection law. *International Data Privacy Law* 4.4, Volume 4.4, pp. 250-261.
- Koops, B.-J. & Leenes, R., 2014. Privacy regulation cannot be hardcoded. A critical comment on the 'privacy by design' provision in data-protection law. *International Review of Law, Computers & Technology*, 28(2), pp. 159-171.
- Korff, D., 2002. EC study on implementation of data protection directive - Comparative summary of national laws. *Online*.
- Kuchinke, W. et al., 2016. Legal assessment tool (LAT): an interactive tool to address privacy and data protection issues for data sharing. *BMC medical informatics and decision making*, 16(1)(81).
- Kuner, C. et al., 2016. The language of data privacy law (and how it differs from reality). *International Data Privacy Law*, 6(4), pp. 259-260.
- LDI, 2000. *Data Warehouse, Data Mining und Datenschutz*. [Online]
Available at:
https://www.ldi.nrw.de/mainmenu_Service/submenu_Entschliessungsarchiv/Inhalt/Entschliessungen_Datenschutzkonferenz/Inhalt/59_Konferenz/20000314_Data_Warehouse_Data_Mining_und_Datenschutz/Data_Warehouse_Data_Mining_und_Datenschutz.php
[Accessed 01 June 2018].
- Leith, P., 2016. The rise and fall of the legal expert system. *International Review of Law, Computers & Technology*, 30(no. 3), pp. 94-106.
- Libaque-Saenz, C. F. et al., 2016. The role of perceived information practices on consumers' intention to authorise secondary use of personal data.. *Behaviour & Information Technology* 35, no. 5, pp. 339-356.
- Licklider, J. C., 1960. Man-computer symbiosis. *IRE transactions on human factors in electronics*, pp. 4-11.
- Loh, W., 2011. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(no. 1), pp. 14-23.
- Loh, W.-Y. & Shih, Y.-S., 1997. Split selection methods for classification trees. *Statistica sinica*, pp. 815-840.
- Lynskey, O., 2016. *In the Shadows of the Data Protection Juggernaut: Bara and Weltimmo*. [Online]
Available at: <https://europeanlawblog.eu/2016/01/28/in-the-shadows-of-the-data-protection-juggernaut-bara-and-weltimmo/>
[Accessed 01 June 2018].
- Mackaay, E. & Robillard, P., 1974. Predicting judicial decisions: The nearest neighbour rule and visual representation of case patterns.
- Maras, M.-H., 2015. Internet of Things: security and privacy implications. *International Data Privacy Law*, 5.2(99).
- Micklitz, H.-W., 2014. On the politics of legal methodology. *Maastricht journal of European and comparative law*, 21(no. 4), pp. 589-595.

- Mohammed, A. F. & Humbe, V. T., 2016. A review of big data environment and its related technologies. *In Information Communication and Embedded Systems (ICICES), 2016 International Conference on* (pp. 1-5). IEEE..
- Mont, M. C. et al., 2010. EnCoRe: Towards a conceptual model for privacy policies. *Primelife*.
- Mortier, R. et al., 2014. Human-data interaction: the human face of the data-driven society.
- Muthuri, R. et al., 2017. Compliance patterns: harnessing value modeling and legal interpretation to manage regulatory conversations.. *In 16th International Conference on Artificial Intelligence and Law*, Volume ACM, pp. 1-10.
- OpenState, 2017. *Meldingenregister Autoriteit Persoonsgegevens ontsloten als open data*. [Online] Available at: <https://openstate.eu/nl/2016/07/meldingenregister-autoriteit-persoonsgegevens-ontsloten-als-open-data/> [Accessed 18 July 2018].
- Oxford Dictionaries, n.d. *Data*. [Online] Available at: <https://en.oxforddictionaries.com/definition/data> [Accessed 19 January 2017].
- Pagallo, U. & Durante, M., 2016. The pros and cons of legal automation and its governance. *European Journal of Risk Regulation*, Volume 7.2, pp. 323-334.
- Papanikolaou, N., Pearson, S. & Mont, M. C., 2011. Towards natural-language understanding and automated enforcement of privacy rules and regulations in the cloud: survey and bibliography. *Secure and Trust Computing, Data Management, and Applications*, pp. 166-173.
- Pasquale, F. A., 2018. A Rule of Persons, Not Machines: The Limits of Legal Automation. *SSRN*.
- Pasquale, F. A. & Cashwell, G., 2015. *Four Futures of Legal Automation*. [Online] Available at: http://digitalcommons.law.umaryland.edu/fac_pubs/1539 [Accessed 10 7 2018].
- Pentland, A. S., 2013. The data-driven society. *Scientific American*, 309(no. 4), pp. 78-83.
- Pertierra, M., Lawsky, S., Hemberg, E. & O'Reilly, U.-M., 2017. Towards Formalizing Statute Law as Default Logic through Automatic Semantic Parsing. *MIT*.
- Pipino, L. L., Lee, Y. W. & Wang, R. Y., 2002. Data quality assessment. *Communications of the ACM*, 45(no. 4), pp. 211-218.
- Poulin, D., Bratley, P., Frémont, J. & Mackaay, E., 1993. Legal interpretation in expert systems. *Proceedings of the 4th international conference on Artificial intelligence and law*. ACM, Volume ACM., pp. 90-99.
- Prakken, H. & Sartor, G., 2015. Law and logic: a review from an argumentation perspective. *Artificial Intelligence*, Issue 227, pp. 214-245.
- Prins, C. & Moerel, L., 2016. Privacy for the Homo digitalis: Proposal for a new regulatory framework for data protection in the light of big data and the internet of things. *SSRN Repository abstract= 2784123*, pp. 1-98.

- Quinlan, J. R., 1996. Bagging, boosting, and C4. 5. *In AAAI/IAAI*, Volume Vol. 1, pp. 725-730.
- Rais, M., 2015. *Data analysis: Benefits and challenges for small and medium businesses*. [Online] Available at: <https://www.linkedin.com/pulse/data-analysis-benefits-challenges-small-medium-businesses-minhaj-rais> [Accessed 19 January 2017].
- Raul, A. C., 2017. *The Privacy, Data Protection and Cybersecurity Law Review*. 3rd ed. London: The Privacy, Data Protection and Cybersecurity Law Review.
- Robinson, N., Graux, H., Botterman, M. & Valeri, L., 2009. Review of the European data protection directive. *Cambridge: RAND*.
- Robinson, N., Graux, H., Botterman, M. & Valeri, L., 2009. *Review of the European Data Protection Directive*. [Online] Available at: <https://ico.org.uk/media/about-the-ico/documents/1042349/review-of-eu-dp-directive.pdf> [Accessed 22 September 2017].
- Rozinat, A. & van der Aalst, W. M., 2008. Conformance checking of processes based on monitoring real behaviour. *Information Systems*, 33(1), pp. 64-95.
- Ruger, T. W., Kim, P. T., Martin, A. D. & Quinn, K. M., 2004. The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking. *Columbia Law Review*, , 104(4), pp. 1150-1209.
- saedsayad, n.d. *OneR*. [Online] Available at: <http://www.saedsayad.com/oner.htm> [Accessed 18 July 2018].
- Salzberg, S., 1991. A nearest hyperrectangle learning method. *Machine learning*, 6(no. 3), pp. 251-276.
- Samuel, A. L., 2000. Some studies in machine learning using the game of checkers. *Ibm Journal of Research and Development*, , 3(3), pp. 210-229.
- Santos, C., Rodriguez-Doncel, V., Casanovas, P. & van der Torre, L., 2016. Modeling relevant legal information for consumer disputes. *Springer, Cham*, Issue In International Conference on Electronic Government and the Information Systems Perspective, pp. 150-165.
- Schafer, B., 2017. Formal Models of Statutory Interpretation in Multilingual Legal Systems. *Statute Law Review*, 38(3), pp. 310-328.
- Schneider, J., 1997. *Cross Validation*. [Online] Available at: <https://www.cs.cmu.edu/~schneide/tut5/node42.html> [Accessed 23 07 2018].
- Schwartz, P. M., 1999. Privacy and Democracy in Cyberspace. *Vanderbilt Law Review*, 1607(1614), p. 52.
- Sergot, M. J. et al., 1986. The British Nationality Act as a logic program. *Communications of the ACM* 29, Volume no. 5, pp. 370-386.

- Shaikh, A. A. & Karjaluo, H., 2015. Making the most of information technology & systems usage: A literature review, framework and future research agenda. *Computers in Human Behavior*, Volume 49, pp. 541-566.
- Shavlik, J. W., Mooney, R. J. & Towell, G. G., 1991. Symbolic and neural learning algorithms: An experimental comparison. *Machine learning*, 6(no. 2), pp. 111-143.
- Sherwin, E., 2009. Legal Taxonomy. *Legal Theory*, 15(1), pp. 25-54.
- Shmueli, G., Patel, N. R. & Bruce, P. C., 2008. *Data mining for business intelligence: concepts, techniques, and applications in Microsoft Office Excel with XLMiner*. s.l.:John Wiley & Sons.
- Siems, M. M. & Síthigh, D. M., 2012. Mapping legal research. *The Cambridge Law Journal*, 71(no. 03), pp. 651-676.
- Slocum, B. G., 2017. *The Nature of Legal Interpretation: What Jurists Can Learn about Legal Interpretation from Linguistics and Philosophy*. Chicago: University of Chicago Press.
- Smith, T., 1994. *Legal expert systems: discussion of theoretical assumptions*. Utrecht: Onderwijs Media Instituut, Utrecht University.
- Solove, D. J., 2001. Privacy and power: Computer databases and metaphors for information privacy. *Stanford Law Review*, pp. 1393-1462..
- Solove, D. J., 2004. *The digital person: Technology and privacy in the information age*. New York: NYU Press.
- Spaeth, H. et al., 2014. *Supreme Court Database Code Book*. s.l.:s.n.
- Verheij, B., 2017. Formalizing Arguments, Rules and Cases..
- Voss, J. F. & Bisanz, G. L., 2017. Sources of Knowledge in Reading Comprehension: Cognitive Development and Expertise in a Content Domain. *In Interactive processes in reading*, Volume Routledge, pp. 215-239.
- Waismann, F., 1968. Verifiability. *How I See Philosophy*. Palgrave Macmillan, London, pp. 39-66..
- Waldron, J., 2016. Desanctification of Law and the Problem of Absolutes. *For Hebrew University Workshop: Law as Religion, Religion as Law*.
- Walker, R. F., 1992. *An expert system architecture for heterogeneous domains: a case-study in the legal field*. s.l.:Doctoral dissertation, AD Druk BV.
- Warren, S. D. & Brandeis, L. D., 1890. The right to privacy. *Harvard law review*, pp. 193-220.
- Waterman, D. A. & Peterson, M. A., 1981. *Models of legal decisionmaking*. s.l.:Rand Corporation.
- Watkins, D. & Burton, M., 2013. *Research methods in law*. s.l.:Routledge.
- Weber, M., 1967. *Rechtssoziologie. Aus dem Manuskript herausgegeben und eingeleitet von Johannes Winckelmann*. Auflage: Neuwied.

- White, R. L., 1996. *Methods for Classification*. [Online]
Available at: <http://sundog.stsci.edu/rick/SCMA/node2.html>
[Accessed 23 July 2018].
- White, R. L., 1997. *Object classification in astronomical images*. New York, NY: Statistical Challenges in Modern Astronomy II. Springer.
- WP260 rev.01, 2018. *Guidelines on transparency under Regulation 2016/79*. Brussels, European Commission.
- WP29 203, 2013. *On Purpose Limitation*, s.l.: Article 29 Data Protection Working Party.
- WP29, 2014. *Statement on Statement of the WP29 on the impact of the development of big data on the protection of individuals with regard to the processing of their personal data in the EU*. [Online]
Available at: http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp221_en.pdf
[Accessed 06 June 2018].
- WP29, 2018. *Guidelines on Transparency under Regulation 2016/679*. [Online]
Available at: http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=622227
[Accessed April 2018].
- Wyner, A. Z., 2008. An ontology in OWL for legal case-based reasoning. *Artificial Intelligence and Law*, , 16(4), pp. 361-387.
- Xu, L. et al., 2014. Information security in big data: privacy and data mining. *IEEE Access*, Volume 2, pp. 1149-1176.
- Young, K., 2013. *Unlocking the Value of Personal Data*, s.l.: World Economic Forum.
- Zarsky, T. Z., 2016. Incompatible: The GDPR in the Age of Big Data. *Seton Hall L. Rev.* 47, Issue 995.
- Zarsky, T. Z., 2017. Incompatible: The GDPR in the Age of Big Data. *The Seton Hall Law Review*, , 47(4), p. 2.