University of Southampton

Faculty of Engineering and Physical Sciences

School of Electronics and Computer Science

A New Approach to Categorising Personal Data to Increase Transparency Under the Obligation to Inform

by

Emma Cradock

Thesis for the degree of Doctor of Philosophy

March 2022

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences
School of Electronics and Computer Science

Doctor of Philosophy

A New Approach to Categorising Personal Data Under the Obligation to Inform

by
Emma Cradock

This thesis contributes to the field of privacy and data protection law, within both Law and Computer Science, by helping to better understand how to increase the transparency of personal data processing and to categorise personal data.

To counter the threat to the privacy of individuals which increasing advancements in Information Technology have created, Data Protection laws have been introduced, which include the key principle of transparency. However, as the de facto method of compliance with the obligation to inform (which mandates the provision of certain information about personal data processing to individuals), Privacy Policies have continuously been criticised in their ability to make processing transparent. This problem makes the study of how to increase the transparency of personal data in the context of providing information to individuals about the processing of their personal data a key research area in both Law and Computer Science.

In researching this problem, this thesis begins by highlighting a gap in the current literature due to the assumption that the problem lies in how information about processing is presented, summarised or communicated, rather than questioning what information is required for processing to be transparent. The finding that Social Networking Sites provided information about the specific personal data they processed in their Privacy Policies, despite the UK data protection Regulator not making this a recommendation led to the next contribution, a critical analysis of the previous and current data protection law of the EU and the UK on when it is a requirement to inform individuals about the specific personal

data being processed. This analysis highlighted that despite its benefits in increasing transparency, organisations are not always required to provide information about the specific personal data they process under the obligation to inform and where they are, the term 'category' is used to differentiate between personal data, without a complete categorisation or sufficient guidance on how to do this beyond the categorisation of 'Special Categories' of personal data. This gap has led to various parties inferring categorisations from the law, or creating their own, without following a categorisation methodology or taking a consistent approach. The result is inconsistent approaches to categorisation of personal data, which fail to achieve the aims of the principle of transparency. The final contribution of this thesis is a proposed categorisation of personal data, based on categorisation methodology and the Data Information Knowledge Wisdom model in Computer Science, which aims to support organisations in increasing the transparency of their personal data processing and can be built upon in the future to support compliance with the Framework's wider compliance requirements.

Table of Contents

List of Tables

List of Figures

Research Thesis: Declaration of Authorship

Print name: Emma Rebecca Cradock

Title of thesis: A New Approach to Categorising Personal Data to Increase Transparency Under the Obligation to Inform

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:-

Cradock, E., Millard, D. And Stalla-Bourdillon, S. (2015). Investigating Similarity Between Privacy Policies of Social Networking Sites as a Precursor for Standardization. *Proceedings of the 24th International Conference on World Wide Web Companion* (pp. 283-289). International World Wide Web Conferences Steering Committee.

Cradock, E., Millard, D. And Stalla-Bourdillon, S. (2016) An Extended Investigation of the Similarity Between Privacy Policies of Social Networking Sites as a Precursor for Standardization *The Journal of Web Science 2(1)*

Cranor, L. F. (2003). P3P: Making privacy policies more useful. *IEEE Security & Privacy*, (6), pp.50-55.

Cradock, E., Stalla-Bourdillon, S., & Millard, D. (2017). Nobody puts data in a corner? Why a new approach to categorising personal data is required for the obligation to inform. *Computer law & security review*, *33*(2), 142-158.

Signature: .....Date:31/12/2020 .....

{Important note:
The completed signed and dated copy of this form should be included in your print thesis. A completed and dated but unsigned copy should be included in your e-thesis}

Acknowledgements

They say, *'it takes a village to raise a child'*, it feels that it has taken a town's worth of people to make the production of this thesis possible and I would like to extend my thanks to some of them in particular.

For their guidance, suggestions and support throughout this process, I would like to thank my supervisors: Dr David Millard and Professor Sophie Stalla-Bourdillon. Their motivation and direction have been invaluable to the production of this thesis and the publications which have supported it. To Dave in particular, thank you for metaphorically *'talking me off a ledge'* a good few times towards the end. I would also like to thank my examiners Professor Leslie Carr and Professor Paul Bernal for their time and a robust, yet enjoyable viva voce. To the lecturers and staff within the ECS and WAIS groups at the University of Southampton, I would like to thank you for supporting me through the process of producing this thesis. Furthermore, I would like to thank the EPSRC Centre for Doctoral Training in Web Science Innovation, University of Southampton [EP/G036926/1], for the PhD opportunity and funding, without which this research would not have been possible.

For her support, inspiration and for constantly creating a world for me in which I can thrive, I would like to thank my mum, Sue, without whom I would not have finished this thesis. For her support and humour, I would like to thank my sister who has always been there for me when I needed her throughout this process. To the rest of my family and friends, thank you for asking (and for not asking at times) how the PhD was going, the distractions and support you gave me helped me have the downtime and respite I needed to get to this point.

Thank you to Amy, Sophie, Doug and Anna for your friendship, support and encouragement, you are the cherries on the cake that was this opportunity. To Rob, thank you too for your constant support and encouragement towards the end, I couldn't have persevered without it. Additionally, to all the other peers and friends I have made on the journey to completing this thesis, in particular, Garf, Sami, Mark, Will, Maria and Johanna, you too have made my PhD experience so memorable.

Finally, to someone that may not be aware of the size of the impact they had, Helen Pottle. You were the first person to teach me the fundamentals of law, without which I would not have been able to progress and complete this research and thesis, thank you for your time, patience and ability to make even the dullest things fun.

Definitions and Abbreviations

Personal Data ..... (also known as personal information) is any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

Processing..... is any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction

Data Subject ..... is the natural person who is the subject of the personal data in question

Data Controller ..... is the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data

Data Processor..... is the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data

GDPR..... stands for General Data Protection Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC

DPD..... stands for Data Protection Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and of the free movement of such data

DPA 1998 ..... stands for Data Protection Act (1998) and is a United Kingdom Act of Parliament which enacted Data Protection Directive 95/46/EC in the UK

DPA 2018 ..... stands for Data Protection Act (2018) and is a United Kingdom Act of Parliament which complements the European Union's General Data Protection Regulation and replaces the DPA 1998

EU ..... stands for European Union, a political and economic union of 27 member states that are located primarily in Europe.

UK ..... stands for United Kingdom, a country made up of England, Scotland, Wales and Northern Ireland

Definitions and Abbreviations

The Web	stands for the World Wide Web, which is an information system where documents and other web resources are interlinked and accessible over the internet
Article 29 Working Party	is the independent European working party that dealt with issues relating to the protection of privacy and personal data until 25 May 2018 when the GDPR came into force, at which point the Article 29 Working Party was replaced by the European Data Protection Board (EDPB)
Member States	are the 27 states (countries) that make up the European Union
US	US stands for the United States of America, a country of 50 states covering a vast swath of North America.
SNS	this stands for Social Networking Sites, also known as social media sites, which are a type of website which allow users to communicate, and to upload and share content with others.
The Reasonable Person	The 'Reasonable Person', also known as the 'Reasonable Man' doctrine is a theory in which the behaviour of an individual (often the accused) is compared to how a hypothetical person or 'reasonable man' would respond in the same set of circumstances.
OECD	Stands for the Organisation for Economic Co-operation and Development, which is an intergovernmental economic organization with 37 member countries founded in 1961 to stimulate economic progress and world trade.
Council of Europe	The Council of Europe is an international organisation founded in the wake of World War II to uphold human rights, democracy and the rule of law in Europe.
Derogation	A derogation is a provision in EU legislative measure which allows all or part of the legal measure to be applied differently, or not at all, to individuals, groups or organisations.
European Commission	The European Commission is the executive branch of the European Union, responsible for proposing legislation, implementing decisions, upholding EU Treaties and managing the day to day business of the EU.
CJEU	The Court of Justice of the European Union (CJEU) is the judicial branch of the European Union and consists of two separate courts: The Court of Justice and the General Court.
Recitals	Recitals are the text at the start of an EU act that sets out the reasons for its operative provisions, while avoiding normative language and political argumentation.

ICO ..... The Information Commissioner's Office is a non-departmental public body which reports directly to the United Kingdom Parliament and is sponsored by the Department for Digital, Culture, Media and Sport.

Information Asymmetry..... Also known as 'information failure', occurs when one party to an economic transaction possesses greater material knowledge than the other party.

FTC ..... The Federal Trade Commission (FTC) is an independent agency of the United States government whose principal mission is the enforcement of civil US antitrust law and the promotion of consumer protection.

EU Data Protection Framework..... This is the framework of laws and bodies within the EU that regulate the processing of personal data within the EU.

Chapter 1 Introduction

1.1 General Overview

Whilst there is no universal definition of Privacy, Information Privacy can generally be thought of as the right of individuals to control who knows what about them. Privacy as a concept has a very long history, however advances in Computing and Information Technology over the last century have increased the importance of Information Privacy as a concept, because these advances have reduced the control individuals have over their personal data.

The ability to collect, store, search and utilise large quantities of personal data about individuals has fundamentally changed the practices of information provisioning, increasing the need for careful consideration of the desirability of the effects of this. In particular, the fact that vast amounts of personal data are now not only provided by individuals themselves, but are also observed, derived and inferred about them without their knowledge (OECD, 2014). Of these technological developments, the World Wide Web ("the Web"), an information system where documents and other resources are accessible over the Internet, is one of the key developments of the information age and has created new challenges for individuals in controlling the access and use of their personal data.

To counter the threats of Computing and Information Technology to an individual's privacy, various jurisdictions have enacted Data Protection and Privacy Laws, which impose obligations on those who process personal data, and have created rights for individuals, which they can use to have control over their personal data. Of these, the European Union Data Protection Framework ("the Framework") is often viewed as the strictest globally, particularly since the introduction of the General Data Protection Regulation ("GDPR") in May 2018.

The Framework sets out key principles that should lie at the heart of any approach taken to the processing of personal data. Of these, the principle of 'transparency' is a long, established feature of the law of the European Union. Transparency is about engendering

trust in the processes which effect the individual, by enabling them to understand them, if necessary, challenge those processes (Article 29 Working Party, 2018).

Under the Framework, transparency is intrinsically linked to fairness, and the new principle of accountability which was introduced under the GDPR. Transparency can be seen to mean that those processing personal data must enable individuals to understand how their personal data is being processed. This empowers individuals to hold Data Controllers (the organisations that decide the purpose and means of processing of their personal data) to account and to exercise control over their personal data, by exercising their rights, or highlighting where data controllers are not fulfilling their obligations. However, the extent to which these obligations and rights extend is often the subject of debate, in particular the extent to which Data Subjects should be given information about and have access to the personal data that is inferred and derived about them (Wachter and Mittelstadt, 2019).

Certain aspects of how the principle of transparency should be applied in practice are prescribed by the GDPR. For example, Article 15 provides individuals with the Right of Access to their personal data. In addition, Articles 13 and 14 specify the minimum information that should be made available to individuals so that they are legally 'informed' about how their personal data is being processed (the 'what'). Whereas, other aspects are left to be decided by the data controller, such as the form and manner in which the information that must be provided under Articles 13 and 14 (the 'how'). However, Article 12 GDPR does suggest that standardised icons that are machine readable could be used to support the 'how' of this information.

Despite this lack of vast direct guidance on the 'how' of transparency, the most common method of doing so is to provide a Privacy Policy (also referred to as privacy notices, privacy statements, data use policies or fair processing notices). This approach has been heavily criticised in practice and is an area where extensive research has been conducted in order to try and improve the 'how' of transparency.

1.2 Research Questions

The gap of this research field is addressed in this thesis, which focuses on what information needs to be provided to individuals to increase the transparency of personal data

processing, and in particular to increase their understanding of the effect of personal data processing on their privacy. An interdisciplinary approach is taken which incorporates research methods, techniques and subject matters from both Computer Science and Law including Legal Doctrinal Research Method, Requirements Analysis and Thematic Analysis. Based on the literature examined, which is covered across the four proceeding Chapters, an overall research goal of the thesis was established:

Research Goal: To understand deficiencies in the current approaches to the transparency of personal data processing in the context of the obligation to inform in practice and to propose an improvement to the way organisations can be transparent about their personal data processing.

Being such a broad topic, it was necessary to devise research questions (RQs) that would produce the insights required to build arguments on this topic. The evolution of this research has resulted in four distinct phases, with the outputs from the first, second and third phases feeding into and shaping the subsequent phases of the research. The overall research questions for each area are:

RQ1: Are the privacy policies of Social Networking Sites (SNS) similar enough in the information they provide about their personal data processing for the standardization of Privacy Policies to be possible?

RQ2: When is there a legal requirement in the EU and UK to provide information about the specific personal data being processed under the obligation to inform and what is the requirement for this?

RQ3: Do various current approaches to categorising personal data and informing individuals of these achieve the aims of transparency under the European Union Data Protection Framework?

RQ4: Could the DIKW model be adapted and used to increase the transparency of personal data processing in relation to the obligation to inform under the European Union Data Protection Framework?

RQ1 was designed to understand whether privacy policies of SNS shared enough information attributes for the standardisation of privacy policies to be possible. It also investigated whether there was any information that privacy policies included that went beyond the UK Regulator's recommendations for transparency, and vice versa. **RQ2** was devised to understand whether there is a legal requirement in the EU and the UK to provide information about the specific personal data that is being processed under the obligation to inform and if so, what this requirement is. **RQ3** was devised to investigate whether some of the current approaches to categorising personal data and providing this information to individuals in privacy policies provides a model of personal data that achieves the aims of transparency of personal data processing. **RQ4** was developed to understand whether a new model of personal data, based on the review of the current approaches and concepts proven in other disciplines had the potential to increase the transparency of personal data processing.

1.3 Scope

The ways in which an appropriate categorisation of personal data could support Accountability under the Framework are many. The ways in which the transparency of personal data processing can be improved are also numerous, thus in order to answer the research questions set previously it was important to limit the scope of this thesis.

Firstly, this work focuses solely on information privacy (opposed to physical privacy etc.) this is because this is the area of privacy that has increasingly come under threat with advances in computing and information technology.

Secondly, the research questions will be analysed in relation to the European Union Data Protection Framework and in particular the General Data Protection Regulation ('the GDPR') as the lens through which information privacy is regulated by law opposed to other jurisdictions. This is because It is the largest omnibus law aimed at the protection of information privacy in the world. Although alternative approaches in other jurisdictions are

discussed at points when examining the literature, such as the notion of Personal Identifiable Information (PII) in the United States of America (US), the ultimate research goal is to increase transparency within the EU and UK legal framework in the context of the obligation to inform. Whilst the contributions of this thesis may also have promise for other jurisdictions, this will not be argued for expressly.

The EU also proves interesting because of its single data protection law, aimed at harmonizing the data protection laws of Member States throughout the EU. However, despite this aim, implementations sometimes differ slightly. In an ideal research world, one would consider the implementations of the GDPR in all EU countries. However, truly understanding the legal dimensions of all of these would be a gargantuan task, far beyond the scope of a single PhD. Therefore, one Member State's implementation was considered alongside discussion of EU law for context. Whist any Member State could have been chosen, the United Kingdom (UK) seemed appropriate, given the researchers familiarity with its legislation. Interestingly, this geographical/jurisdictional scoping is not inherently a problem for computer science as a discipline in the same way that it is for law. Therefore, discussion of work in the area of improving transparency, and work on the DIKW hierarchy Is not geographically limited to the EU.

During the time that this research was undertaken the UK voted to leave the EU. Whilst the UK officially left the European Union on the 31st January 2020, it has been in a transition period which will end on the 1st January 2021. During the transition period, the GDPR continued to apply to the UK and therefore the claims made in relation to the GDPR in this thesis still applied to the UK. It is envisaged that following the end of the transition period, the GDPR will be brought into law as the 'UK GDPR' and therefore it is expected that the findings of this thesis will continue to be applicable to the UK even after the GDPR ceases to directly apply to it.

Thirdly, there are points throughout the research where it is necessary to examine Privacy Policies which Data Controllers are using in practice. In each of these cases Social Networking Sites (SNS)(also known as Social Media Sites) have been chosen as a case study for this research. Whilst, it is vital that all online services collecting and processing

information about individuals are transparent about their activities, one 'type' of Data Controller was chosen for investigation, to reduce the possible effect of confounding variables. SNS proved appropriate because they were the second most frequently visited 'type' of website globally (after search engines) (IPSOS, 2013) at the time of the preliminary study discussed in Chapter 3 and have (and continue to) attract criticism and concerns for their collection and use of personal data (Anderson, 2009; Beigi & Liu (2018)). Therefore, any improvements to their policies, and the way that they are transparent about their personal data processing will have a wide reach. Whilst this limitation in scope means that some conclusions can only be drawn about SNS, in other ways the results can aid understanding of online privacy policies and approaches to the obligation to inform as a whole. In this sense, the privacy policies of SNS are being studied to provide information on a broad range of phenomena, and to develop principles and models that apply to similar settings.

Finally, under the current Framework, compliance in general, and in particular, transparency is thought to be secured through a combination various obligations and rights under the GDPR. It could have focused on the Data Controllers' obligation to inform the data subject of certain information (Articles 13 and 14 GDPR) and the right of access data subjects have to the data which organisations are processing about them (Article 15 GDPR). Whilst the contributions of this thesis could be used to increase transparency in each of these areas, it focuses on the ability of the contributions to increase transparency in relation to the 'obligation to inform' (or the 'right to information' as it has more commonly become known'). Any further scoping of the thesis is discussed in the individual Chapters.

1.4 Outline of the thesis

This introduction has provided the context of this research and the motivation behind it. It has described the goal of the research, which was to understand a deficiency in the current approaches to transparency in the context of the obligation to inform in practice within the EU Data Protection Framework and to propose an improvement to the way Data Controllers can be transparent about their personal data processing. It also outlines the four research questions, the answers to which have guided the contributions made in this thesis towards solving of this problem. The remainder of this thesis is organised as follows:

Chapter two acknowledges the literature relevant to this field of study, which is fundamental to answering the Research Questions of this thesis. Specifically, it begins by explaining why information Privacy has come under threat due to technological advancements and how Data Protection and Privacy Laws have been introduced to counter this threat. In particular, it discusses the concept of transparency in European Data Protection law and specifically, the obligation to inform, which has prompted the adoption Privacy Policies by Data Controllers as the de facto means of compliance with the obligation. The review of this research highlights the deficiencies of privacy policies in achieving the transparency of personal data processing in practice, a problem which this thesis aims to address. Whilst other literature is also reviewed in Chapters Three, Four and Five, it is specific to the research questions covered in them, which is why it is dealt with in those chapters along with the specific methodologies used.

Chapter three reports on the first set of key findings for this thesis. The results of a preliminary study into the similarity of the Privacy Policies of Social Networking Sites (SNS) and the potential for the standardization of privacy policies in general are presented in answer to **RQ1** of this thesis. The findings indicated that similarity in terms of specific clauses used was low, but that similarity in terms of the themes of information the privacy policies covered was much higher, and recommendations for standardisation of privacy policies to be achieved were made. However, the findings also highlighted an assumption of the current research in the field, which is that the information that the law requires to be provided to individuals (the what) will actually lead to transparency in practice. This means that much of the research in this area focuses on how to present this information so that individuals or computers can interpret it (the how). Yet, as discussed above, the point of informational privacy is for individuals to control 'who knows **what** about them'. In particular, the study found that all of the privacy policies studied included information about the specific personal data that they processed (the 'what') and yet this was not a recommendation made by the UK Data Protection Regulator (the ICO) for a compliant Privacy Policy to include this information. Yet, the literature review indicated that providing information on the personal data being processed is fundamental to the transparency of personal data processing, especially given that a large amount of personal data processed by organisations that is not provided by individuals, but instead observed, derived and

inferred about them (OECD, 2014). Thus, if this problem of investigating 'what' information individuals should be provided with was left unresolved, it would stand in the way of the 'how' of transparency including the creation of a transparent, standardized, and legally compliant privacy policy. It would also stand in the way of the goal of this research, which is to improve the transparency of personal data processing in the context of the obligation to inform.

Chapter four then discusses the second study undertaken as part of this thesis in response to the findings of the preliminary study. The aim of this study was to investigate the assumptions highlighted by the preliminary study. In particular, it looked at the 'what' of personal data processing and whether it is a legal requirement for organisations to provide information on the specific personal data they process in the context of the obligation to inform, and if so how this should be done. The study used critical analysis and descriptive, legal doctrinal and evaluative research methods to describe the current and previous state of affairs in EU and UK law and to evaluate whether the current rules work in practice in order to answer **RQ2**. It found that the requirement to provide this information on the specific personal data being processed is not consistent, and that there is limited guidance on how to provide this information in practice which means that it is unclear how to do this in practice.

Chapter five then discusses the next phase of the research, which sought to understand whether it is important for the principle of transparency that individuals have information on the specific personal data being processed, as if it is not then this would be a fruitless area of research. This chapter allows for an understanding of whether this is an important deficiency in the law that needs to be rectified to increase the transparency of personal data processing and merited further research and a proposal of a categorisation of personal data. It also looked to understand whether, despite the lack of guidance on how to categorise personal data under the obligation to inform, a consistent approach to categorising personal data has emerged in practice or at least an appropriate one that could be adopted for the requirement under the obligation to inform.

Chapter Six presents a new approach to categorising personal data to support transparency under the obligation to inform, using a model from Computer Science. This chapter then

uses understanding from Psychology and Computer Science to validate this approach and to discuss how it has the potential to improve the transparency of personal data processing in the context of the obligation to inform.

Finally, Chapter Seven summarises the key findings identified in this thesis and the contributions made in this thesis are listed. Proposals for future work are presented and the final concluding statement is then made.

Chapter 2 Literature Review: Privacy, Transparency and Privacy Policies

This chapter reports on the threat to the information privacy of individuals that increasing advancements in communication and information technology have created. It synthesises significant literature which is fundamental in setting the context of this thesis and in answering its Research Questions. The Chapter provides an overview of the concept of Privacy and the laws in the European Union that have been introduced to counter the threat to privacy, which technological advancements have created. It also introduces the concept of 'personal data', which is the data that is protected under these laws and the principle of transparency, which requires organisations to be transparent about how they process such data. It then discusses a specific obligation Data Controllers have aimed at increasing the transparency of processing, which is to provide certain information to individuals about how their personal data is processed. Finally, it discusses the phenomenon of Privacy Policies, which have been adopted as the de facto form of compliance with the specific obligation Data Controllers have to provide certain information to individuals about how their personal data is processed. In particular it discusses the criticism Privacy Policies have received in their ability to make the processing of personal data transparent, a gap in this field which this thesis intends to address. It is worth noting that some additional literature is reviewed in Chapters Three, Four, Five and Six where it is more relevant to the specific research question the chapter is answering.

2.1 Privacy

The collection of personal data about individuals is not a new phenomenon, indeed '*The collection of personal data is as old as society itself. It may not be the oldest profession but it is one of the oldest habits*' (Earl Ferrers (HL Deb (1993-1994 549 col. 37). However, technological and computational advancements over the last two centuries, including the creation of the Web (Gervais et al., 2017), in addition to increases in the abilities of Machine Learning and the use of Big Data (Wachter and Mittelstadt, 2019) have facilitated the collection and processing of personal data on a scale like never before. This dramatic increase in the potential of personal data has benefitted both society and the individual, from the identification of cures for diseases (Gostin, Levit and Nass, 2009) to the provision of personalised services (Teltzrow, and Kobsa, 2004). Yet, tis dramatic increase in the potential of personal data has also led to concerns over its effect on the privacy of

individuals (Rowland et al., 2012).

Despite being a fundamental human right (Article 12, United Nations Declaration of Human Rights, 1948; Article 8, European Convention on Human Rights, 1953) and an age-old concept, 'privacy' has proved notoriously difficult to define (Martin and Murphy, 2017).

Moore (2008) suggested that this difficulty is because rituals of association and disassociation are cultural and species relative, what would be considered a violation of privacy in one culture may be permitted, or culturally accepted in another. This makes reaching a consensus on exactly what privacy is, and the point at which it is violated, difficult. Perri 6 argued that the reason for the lack of consensus over a definition is because *'as a society we do not and cannot agree on what it is about life and privacy that we value'* (1998:21). Indeed, Stewart (2017) asserts that individuals have clear perceptions about when they experience a loss of privacy but may differ in what they believe constitutes privacy.

Despite this difficulty, many scholars have attempted to provide definitions. Perhaps one of the earliest definitions was that of Warren and Brandeis (1890) who defined privacy as the 'right to be let alone', which suggests that individuals have a private self, in addition to a public one. Gavison (1980) believes that the ability to control personal information is a determinant in the definition of privacy, which makes both a scope of the concept and the provision of legal protection over it problematic, because the definition relies on subjective choice. Solove (2002) contends that the reason suggested definitions have not prevailed as universal definitions of privacy is because their scope was either too broad or too narrow and proposed six categories for these definitions of what privacy is (1) the right to be let alone, (2) limited access to the self, (3) secrecy, (4) control of personal information, (5) personhood and (6) intimacy. Indeed, Martin and Murphy (2017) assert that there has been little research published which focuses on privacy as a construct, and that instead research has been focused on the relationship of privacy to other things. For example, how important it is and the consequences of losing it.

Despite the lack of consensus on the concept of privacy, this thesis focuses on what was referred to in Soloves's (2002) category (4), the 'control of personal information', which has become more broadly referred to as 'information Privacy' (Solove, 2016; Smith, Dinev and Xu, 2011).

The reason for this focus is because of the dramatic impacts computational progress have had on the ability of an individual to control how their own personal data is shared and used. Conversations about technology and privacy have been occurring since the late 1890s, with Warren and Brandeis seminal paper 'The Right to Privacy' (1890). Their paper discussed the availability of portable photography equipment and the challenges to privacy this may cause. Even fears surrounding personal data and computers have been long voiced, with the Younger Committee on Privacy in the UK (1972) identifying three characteristics, which distinguished the ability of computers to store information from more traditional methods. This included the use of computers to compile personal profiles, their capacity to correlate information and the ease with which unauthorised access to data could be obtained, often from remote sites.

Instead of looking for a consensus on a definition of privacy to address concerns over Information Privacy, an alternative approach has been suggested as to look for a consensus of when it has been violated. McArthur (2001) looked at what 'reasonable expectations' of privacy are required in order to assess whether concerns that privacy is eroding are justified. Applying the mischance and voluntary principles, McArthur (2001) held that the reasonable person should not have expectations of privacy regarding his web and internet use, because in knowing they can be tracked, one is negatively volunteering the information through lack of care. However, it is questionable how much users 'know' about the technology involved with online personal data processing due to breadth and pace of various technological developments, including the prevalence of cookies and the power of search engines. This begs the question of how much even 'technically savvy' users are aware of how their online activity is being monitored and therefore what standard for the 'reasonable person' should be. Indeed, the OECD (2014) acknowledged that it is not only personal data that is provided by individuals which is collected and processed online, but also personal data that observed, derived and inferred about them without their knowledge.

Westin (1967) proposed that there were three types of individuals in relation to privacy: fundamentalists who believe in privacy rather than openness, pragmatists who weigh up the opportunities which services provide when It comes to compromising their privacy and the unconcerned. However, when it comes to the processing of personal data online this classification is problematic. Norberg, Horne and Horne (2007) identified a 'privacy paradox' because of the differences between what people say and do regarding their

privacy online. This paradox means that what someone might say may class them as one type of individual in Westin's definitions, whereas their actions may class them as another type, meaning that there is more nuance than just three types of individuals. This compliments Richards and Solove's (2007) view that privacy cannot be reduced to a single essence, because it is a multiplicity of different, yet related things. Kokolakis (2017) found that despite extensive research on this paradox there is still much more future work to do to understand it. As discussed further in Section 2.2.5, it may be that users do not have the controls or knowledge to allow their behaviour online to reflect their views on how they want to control their information.

2.2 Privacy and Data Protection Laws

2.2.1 Privacy and data protection laws

Due to this threat to privacy, and the reduced ability to control information that advancements in technology and computation have created, Data Protection and Privacy Laws have been used to provide some protection and attempt to strike the balance between the rights of individuals to privacy and the ability of organizations to collect and use personal data (Rowland, Kohl and Charlesworth., 2012). in the UK these laws were traditionally viewed as technical legislation, relating to information management practices (Rowland, Kohl and Charlesworth, 2012:51). However, the terms 'data protection' and 'privacy' have come to often be used interchangeably (Gellman, 1998) leading to criticism from some that data protection law is 'back door' privacy law in the UK (HL Deb (1997-1998 585 col. 445), 1998).

Before 1995, action from the Organisation for Economic Co-operation and Development (OECD) (1980) and Council of Europe (1981) had influenced the legislative action on data protection and privacy in many EU Member State countries. Yet, despite this movement to legally protect individuals, great fragmentation had developed between Member State regulatory regimes, and a great number remained without legislation (Charlesworth, 2000). This fragmentation led to the creation and adoption of the Data Protection Directive 95/46/EC (DPD) in the EU in 1995. Two supplementing Directives also formed the EU Data Protection Framework, Directive 2002/58/EC and Directive 2009/136/EC on privacy and electronic communications were also adopted with the aim of protecting an individual's information privacy in relation to electronic communications. Although. In-depth

discussion of this particular legislation is outside the scope of this thesis.

As a Directive, Member States were required to transpose the DPD's requirements into their national legislation and whilst the DPD was legally binding on the result to be achieved by these transpositions, Member States were free to choose the form and methods of doing so. In the UK the DPD was transposed via the Data Protection Act 1998 (DPA 1998).

Whilst the introduction of the DPD was a great step forward in providing protection and control to individuals, in practice, the ability of Member States to create slightly different implementations meant that despite the aim of having a single law for the whole of the EU, differing implementations lead to fragmentation of the DPD throughout the EU (Korff, 2003). The DPD's provisions also became increasingly criticized, for failing to effectively regulate the modern technological climate (European Commission, 2015). In light of these issues, in January 2012, the European Commission proposed a comprehensive reform of the data protection rules in the EU, the result of which was the enactment of the General Data Protection Regulation (GDPR). This Regulation replaced the DPD and is directly applicable in Member States due to its status as a Regulation rather than a Directive, without the need for implementing national legislation (although various derogations on specific sections allow for some degree of flexibility to how the requirements apply in Member States). The GDPR was given final approval by the European Council and Parliament in April 2016 and came into force on 25 May 2018. In the UK, the Data Protection Act 1998 was replaced with the Data Protection Act 2018 (DPA 2018) which compliments the GDPR and implements the UK's derogations in relation to the GDPR.

As discussed previously, the Framework is principles-based, and so at the heart of it are a number of Data Protection Principles (Article 5 GDPR; Section 34 DPA 1998) which are set out at the start of the legislation and inform everything that follows (ICO, 2020). The Framework then goes on to provide a number of obligations for Data Controllers (organisations that determine the purposes and means of the processing of personal data); and Data Processors (organisations that process personal data on behalf of Data Controllers). The Framework also provides rights to Data Subjects, the individuals who are the subject of the personal data being processed by Data Controllers and/or Data Processors.

To oversee the Framework, there are bodies in both the EU and the UK that are responsible for overseeing compliance with the Framework. The Article 29 Working Party (WP), full name '*The Working Party on the Protection of Individuals with regard to the Processing of Personal Data'* was the independent European working party and advisory body made up of a representative from the data protection authority of each EU Member State, the European Data Protection Supervisor and the European Commission. The role of the WP was to provide advice to Member States regarding Data Protection and to promote consistent application of the DPD. The WP was replaced by the European Data Protection Board (EDPB) 25 May 2018 with the introduction of the GDPR, who has the responsibility to ensure that the GDPR is consistently applied in all countries and to provide general guidance (including guidelines, recommendations and best practice) to clarify the GDPR (Europe, 2020). During its first plenary meeting the EDPB endorsed several GDPR-related WP Guidelines, and therefore some guidance produced by the WP is still relevant today. In the UK, the Supervisory Authority which oversees compliance with the Framework is the Information Commissioner's Office (ICO) who held this role under both the previous and now the current Framework.

2.2.2 Personal data

A key concept which dictates the material scope and application of the European Data Protection Framework and the control that it provides to individuals is that of 'personal data', which is defined by the GDPR as:

> *'any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person'*
>
> **Article 4(1) of the General Data Protection Regulation**

Whilst Personal Data has generally been interpreted to have a broad definition under the EU Data Protection Framework, a landmark case in the UK in 2003 threatened to limit this.

In *Durant v Financial Services Authority*[1] an individual making a Subject Access Request for personal data from the Financial Services Authority (FSA) relating to a complaint made about Barclays Bank was not provided with all documents he believed to be considered his 'personal data'.

In deciding whether the FSA was obliged to disclose further information in relation to its investigation of the complaint, the Court of Appeal considered the definition of 'personal data'. In particular, the Court considered two definitions of the 'relating to' element of the definition of personal data (Chalton, 2004). The first required data to have a direct connection with an individual, whereas the second simply required that data must have some connection with, or be connected to the individual, in order for it to 'relate to' them.

The Court of Appeal favoured the former definition and rejected the appeal, holding that the data the individual sought was not 'personal data' and related to the complaint made by him, rather than to him. The Court held that "*the data must go beyond the mere mention of an individual's involvement in a matter that has no personal connotations… the data should have the individual as its focus, rather than some other person with whom they may have been involved or some transaction or event in which the individual may have had an interest."*.

15

Whist the judgement was welcomed to some extent by the Information Commissioner's Office (ICO), for providing clarity on the definition of personal data, it was controversial because it effectively narrowed the definition of personal data, and therefore the scope of the data protection framework in the UK at this point.

Subsequently in 2007, the Article 29 Working Party issued an opinion on the definition of personal data (WP, 2007) which endorsed a broad interpretation of the definition in clear contrast to the restrictive interpretation in the Durant case. They opined that there were three central concepts of how data may relate to an individual – through purpose, content and result. Whilst this opinion was not binding, it provides basis for the interpretation of the definition of personal data in line with the Data Protection Directive at this time.

[1] Durant v Financial Services Authority [2003] EWCA Civ 1746

The UK Information Commissioner's Office (ICO) also issued guidance which followed the WP's opinion and advised that Durant should only be considered where data is not 'obviously about' an individual or clearly 'linked to' them. This guidance was also endorsed by the UK Court of Appeal in Edem[2], a case which also involved an individual requesting information from the FSA about the handling of a complaint. The request was to have access the names of three employees who had worked on the complaint. The Court of Appeal held that names were personal data on the basis that a name is always personal data when the context in which it appears is sufficient to identify the named individual. This decision confirmed the approach in the ICO Guidance that the Durant test should be confined to scenarios where personal data is not 'obviously about' an individual.

With the introduction of the General Data Protection Regulation (GDPR) in 2018, the narrow definition of personal data in Durant became redundant to some extent. Indeed, Purtova (2018) asserts that both Article 29 Working Party guidelines and the caselaw of the Court of Justice of the European Union (CJEU) suggest that in the near future, everything will be or will contain personal data, leading to the application of data protection laws to everything.

In understanding what personal data is not, the opposite of personal data is 'anonymous
16 ' which has no formal definition under the GDPR. Whilst the concept of 'anonymous data' is outside the scope of regulation, it is referred to in the Recitals where it is stated that:

> *The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable'.*
>
> **Recital 26 of the General Data Protection Regulation**

Various academics have argued that advancements in data processing technologies and the abundance of data that is now available mean that absolute and irreversible anonymity is now no longer possible (Ohm, 2010; Sweeney, 2000; and Schwartz and Solove, 2011).

[2] Efifiom Edem v Information Commissioner and Financial Services Authority [2014] EWCA Civ 92

However, it has also been argued that a risk-based approach should be able to produce data that is sufficiently anonymous to be protected in a legal sense under the Framework (Arbuckle and Mian, 2020; Cavoukian and El Emam, 2011; Aldhouse, 2013).

The European Court of Justice (ECJ) has ruled on the meaning of personal data in relation to both the Data Protection Directive and the GDPR, and in particular, took a wide view of identifiability in the Breyer (2016) case[3], which held that dynamic IP Addresses may constitute personal data even where only a third party (in this case an Internet Service Provider) has the additional data necessary to identify the individual.

Although the Court often confirms that the scope of 'personal data' is very wide, it does not often discuss the complete definition in further detail but instead confirms whether a particular type of data involved constitutes 'personal data' and the circumstances in which it does (Purtova, 2017). In the case of YS and others (2016), examples of types of data which the Court had previously explicitly pronounced as personal data were listed, including 'the name of a person in conjunction with his telephone co-ordinates or information about his working conditions or hobbies (2003)[4], his address (2009)[5] his daily work periods, rest periods and corresponding breaks and intervals (2013)[6].

Due to the fact that the definition of personal data under the Framework is broad and that even in 2014 the OECD found that much of the personal data that is processed about individuals is observed, derived and inferred without their knowledge (OECD), the principle of transparency has a key role to play here.

2.2.3 Transparency

Both the DPD and the GDPR provided key principles which should lie at the heart of a Data Controller or Processor's handling of personal data (Article 6 DPD, Article 5 GDPR). Transparency has always been a key concept within the EU data protection framework, but in the DPD whilst it underpinned all of the principles, it was not called out specifically, but

[3] Case C-582/14, Patrick Breyer v. Bundesrepublik Deutschland [2016] ECLI:EU:C:2016:779.
[4] Case C-101/01 *Bodil Lindqvist* [2003] ECR I-12992
[5] *College van burgemeester en wethouders van Rotterdam v M.E.E. Rijkeboer* [2009] ECR I-3889
[6] Case C-342/12 *Worten Equipamentos para o Lar SA v Autoridade para as Condições de Trabalho (ACT)* [2013] OJ C225/37

was particularly key to the first principle, that data must be processed 'fairly and lawfully'. However, the introduction of the GDPR reinforced this importance of transparency by amending the principle to state that data must be '*processed lawfully, fairly and **in a transparent manner***' (emphasis added) under Article 5(1)(a) GDPR.

Whilst there is no specific definition of transparency, it is generally the idea that individuals should be aware of processing involving their personal data. Data controllers must be clear, open and honest with people about who they are and how and why they use a data subject's personal data (ICO, 2019). Recital 39 does state that:

> *'It should be transparent to natural persons that personal data concerning them are collected, used, consulted or otherwise processed and to what extent the personal data are or will be processed. The principle of transparency requires that any information and communication relating to the processing of those personal data be easily accessible and easy to understand, and that clear and plain language be used. That principle concerns, in particular, information to the data subjects on the identity of the controller and the purposes of the processing and further information to ensure fair and transparent processing in respect of the natural persons concerned and their right to obtain confirmation and communication of personal data concerning them which are being processed. Natural persons should be made aware of risks, rules, safeguards and rights in relation to the processing of personal data and how to exercise their rights in relation to such processing. In particular, the specific purposes for which personal data are processed should be explicit and legitimate and determined at the time of the collection of the personal data'*
>
> **Recital 39 of the General Data Protection Regulation**

Transparency helps individuals to understand what is going on with their information; to make informed and subjective decisions about processing (6, 1998); and (when necessary) enforce their data protection rights (BEUC, 2011). Importantly, transparency is not simply something that is simply desirable within the EU data protection framework, it is vital to the efficacy of the framework as a whole. As a principle and rights-based, complaint-driven system (Taylor, 2015), where data protection authorities do not have the budget, powers, or staff to completely police the processing of personal data by organisations, the principle of transparency allows data subjects to enforce their rights and keep a check on data

controllers.

Because of the elusiveness of the definition of 'privacy' and the link between data protection laws and the protection of information privacy (Rowland, Kohl and Charlesworth, 2012), transparency becomes a very import concept. It allows people to make subjective decisions regarding the processing of their information in place of complete paternalistic regulation. Some proponents argue that data protection laws restrict the free market for data (Rowland, Kohl and Charlesworth, 2012) and so transparency and choice are used to strike this balance between and allow individuals to make choices about what data they are willing to provide. Stewart (2017) argues that where the consumer can control what, and how they share information consumer this actually leads to greater willingness to share personal data which evidences the importance of transparency and control. Even where individuals do not have a choice over the processing of their personal data, transparency is still essential, to enable individuals and the regulator to detect any unlawful practices to report these to a Supervisory Authority. Weber (2015) also suggests several steps to increase transparency:

> *Provide information about the intended collection, storage and/or data processing; provide an overview of what personal data have been disclosed to what data controller under which policies; provide online access to the personal data and how they have been processed; and provide counter profiling capabilities helping the user to anticipate how their data match relevant group pro-files, which may affect future opportunities or risks*
> **Weber (2015)**

Under the EU Data Protection Framework, 'transparency' is also intrinsically linked to fairness, and the principle of accountability that was introduced under the GDPR. In general, transparency is thought to be ensured in the Framework via a combination of obligations that are placed on data controllers and/or data processors and the rights that are given to data subjects. The specific obligations and rights that directly support transparency in the Framework are:

- The obligation of data controllers to inform data subjects about certain specific aspects of personal data processing (referred to here as the 'obligation to inform') (Articles 13 and 14 GDPR); and

- The right data subjects have to access the personal data which organisations are processing about them (Article 15 GDPR)

A certain amount of transparency was also created under the DPD with the data controller obligation to notify their data protection supervisory authority about certain aspects of their processing. However, this only concerned the data controller's processing activities as a whole and was replaced with an obligation to internally document this information instead under the GDPR (Article 30 GDPR). Whilst this can create a certain amount of transparency for Supervisory Authorities who have the power to consult them (Article 30(4)), this information is not usually readily available to individuals.

There are other areas of the Framework that support transparency indirectly. Data controllers are required to carry out a Data Protection impact Assessments (DPIAs) (under Article 35 GD[R) if they process personal data in a new way that will result in a high risk to the rights and freedoms of individuals. Wachter (2018) argues that DPIAs can be useful to increase user trust and transparency, especially if they are published publicly, as previously recommended by the Article 29 Working Party (WP, 2017).

Thus, of these compliance requirements that are aimed to ensure transparency in the Framework, the obligation to inform is especially key to making data processing transparent. It provides the only information data subjects are guaranteed to receive about processing with no further effort on their part (unlike enforcing their right of access). Furthermore, information is generally given at the time the personal data is obtained (or within a reasonable period of no longer than a month if the personal data is obtained from elsewhere, Article 14(3)). This ensures that in theory, Data Subjects have the information that would allow them to make required, and informed and appropriate decisions where they have a choice over processing. Where there is no choice, the information helps data subjects: understand what is happening with their personal data; enforce their data protection rights (when necessary); and to detect any unlawful, or questionable practices. This thesis therefore focuses on the obligation to inform and its role in making the processing of personal data transparent, particularly as researchers in the field of Human Computer Interaction (HCI) of Computer Science, which studies the design and use of Computers by humans, still struggle to understand how the current legal Framework should

be interpreted and applied in practice despite several initiatives to improve transparency and accountability (Krebs,. et. al, 2019).

2.2.4 The Obligation to Inform

As discussed, the data controller's 'obligation to inform' (also referred to as the right to information) requires that data controller's provide individuals with certain information prior to processing their personal data. Article 13 GDPR governs cases of collection of personal data from the data subject and Article 14 GDPR governs cases where the personal data has not been obtained from the data subject. Each Article lists the information that must always be provided in any circumstance and the point in time at which this must be done. There is also a list of information that should be provided where the specific circumstances require it, to ensure fair and transparent processing.

There are exemptions to the obligation to provide this information and both Articles provide one in the circumstances that the data subject already has the information (Article 13(4) and Article 14(5)(a)). Article 14(5) GDPR also provides exemptions from providing this information when personal data is not obtained from the data subject and:

- The provision of such information proves impossible or would involve a disproportionate effort (Article 14(5)(b));
- The processing is laid down by Union or Member State law, which provides appropriate measures to protect the data subject's legitimate interests (Article 14(5)(c)); or
- Where the personal data must remain confidential subject to an obligation of professional secrecy regulated by Union or Member State law, including a statutory obligation of secrecy (Article 14(5)(d));

The obligation to inform is a stand-alone obligation. However, it is also a prerequisite and first step of processing personal data 'lawfully, fairly and in a transparent manner in relation to the data subject', the first principle under the GDPR (Article 5(1(a))). The second aspect of satisfying this principle is to legitimize the processing. This is done by satisfying one (or more) of the specified preconditions under Article 6 GDPR which provide a lawful basis which justifies why the processing can occur e.g. because it is in the legitimate interests of the organisation (Article 6(1)(f)) or because the processing is necessary for the performance

of a contract to which the data subject is party (Article 6(1)(b)).

Because of this link between providing information and legitimizing the processing, the obligation to inform is often linked to gaining a data subject's consent (the first of the list of possible legitimizing preconditions). Yet, as the Article 29 Working Party (2011:19) noted, under the previous Framework, consent does not always follow the provision of information (as another ground for legitimizing the processing can be used), but there must always be information before there can be consent.

The importance of transparency of personal data processing has grown with technological progress over the last twenty years. In 1995, at the inception of the DPD, the majority of the personal data processed by controllers was 'provided' by individuals, with their full awareness that their personal data was being obtained. As the active source, the presumption was that both data subjects and controllers had equal information on exactly what **'**personal data' was being collected and processed.

However, technological progress has seen a substantial growth in the amount of personal data that is observed, derived and inferred about individuals, without their awareness (OECD, 2014). At the same time, the definition 'personal data' has broadened, to account for new technologies as discussed in Section 2.2.2. Data processing tools have become increasingly powerful, sophisticated, ubiquitous, and inexpensive, making information easily searchable, linkable and traceable (OECD, 2020). The result of this progress is that it can no longer be presumed that data subjects are aware of what specific 'personal data' is being collected or how it can be generated and processed by data controllers.

In its role of redressing the balance of information between Data Subjects and Data Controllers, it is the task of the 'obligation to inform' to reverse this presumption, and require that data controllers provide information to individuals that makes what specific personal data is being collected and processed transparent. If not, data processing will be less transparent now than it was twenty years ago, as data subjects will have access to less information about the processing of their data and will be less capable of assessing the compliance of controllers.

2.2.5 Privacy Policies

In its role in promoting transparency, the obligation to inform has previously led to the

adoption of privacy policies (also referred to as Privacy Notices) as the de facto method of complying with this requirement. Privacy policies are the explanations individuals are given when information is collected about them (Information Commissioner's Office, 2010). There is no strict definition of these, but policies/notices can be considered to be accessible texts, aiming to inform the average data subject and contain specific legal information delineating the data subject's rights and the data controller's obligations (ICO, 2009:28).

However, despite their common usage, the importance of transparency and the instructive provisions of the GDPR in specifying the information that a Privacy Policy needs to include, the role of privacy policies in informing users has proven unsatisfactory. They have long been criticized for being too long (McDonald and Cranor, 2008), legalistic, complex (ICO, 2010) and ineffective in helping users understand their rights (Scribbins, 2001). An early study in the US showed that users would need to have at least US college-level education to understand their complexity and sentence structure (Anton et al., 2004). Evidence also suggests that their use it targeted at meeting applicable legal requirements rather than serving a real transparency benefit towards the consumer (ICO, 2009:29). The result is that individuals do not read them, with even recent studies finding that 74% of participants continue to skip reading the Privacy Policy (Obar and Oeldorf, 2020) in using a service.

Arguably, this invalidates the role Privacy Policies have in making processing transparent, (ICO, 2009:30) but yet they continued to be used as the de facto approach to the obligation to inform in the EU and the UK and years on these problems with Privacy Policies continue to be an issue for effective transparency (Sarne, et. al., 2019)

Data Controllers could be seen as complicit in this lack of transparency, as with individuals still using their services, they lack strong incentives to improve their Privacy Policies and the information that they provide under the obligation to inform. Even if incentivized, creating a concise and compliant policy is not easy, given the supranational nature of the web, where data is processed in numerous jurisdictions, each with differing legal requirements (Rowland et al., 2012). To attempt to include all informational requirements for transparency around the world can sometimes mean that Data Controllers have no choice but to have long and complicated Privacy Policies or Notices. Additionally, there is a lack of guidance within the legislation on some aspects of the requirements of executing this obligation. For example, information about personal data processing is required to be

'easily accessible' and 'easy to understand' under the GDPR (Recitals 39, 58 and Article 12(1)) but there is no further guidance within the legislation on what this means in practice other than that icons could be used to support the information (Article 12(7) GDPR). The Article 29 Working Party has made some related suggestions of improving the accessibility of Privacy Policies by making them easier to understand, however these were labelled as naïve and contradictory by members of the commercial sector. This was because some Member State national laws required full descriptions of data processing activities, which prove difficult to describe in a form the consumer can understand (ICO, 2009:29) whilst also trying to keep Privacy Policies short.

Despite these issues, the benefits of providing this information are numerous and if executed well, privacy policies can promote transparency and reduce information asymmetry (Tsai et al., 2011). Information Asymmetry is an economic term which is used to describe transactions where one party (usually the supplier) has more or better information than the other (Akerlof, 1970). By communicating information that enables users to make effective privacy choices, privacy policies or the effective communication of privacy information has the potential to reduce this information asymmetry and support the aims of data protection and privacy laws to increase the transparency of personal data processing. Indeed, evidence suggests users are privacy aware and active (Hargittai et al., 2010), just that they do not view privacy policies as a means of expressing consent (Robinson et al., 2009). Grossklags and Acquisti (2007) suggested that Privacy is a risk trade-off, thus effective privacy policies can enable users to decide on the risks of not providing personal data. Well executed Privacy Policies can also give organisations a competitive advantage, by reassuring potential and existing customers that you take their privacy seriously (ICO, 2009:7). Indeed, Tsai et al. (2011) found that in the context of online shopping, participants were likely to pay a premium for privacy when privacy information was made more accessible. It is worth noting that this was a laboratory study, which may have affected participant's choices because they were being monitored. It was also a type of transaction that involved money already, thus the results may not extend to free-to-use websites. Other benefits of a well-executed Privacy Policy have been cited as: higher levels of trust and better relationships with the people you collect information about (ICO, 2009:7).

Thus, understanding how to make data processing transparent is an important research area and much work has been done in this space. Two broad approaches taken to this

research are: firstly, making the information privacy policies provide more accessible to the human reader; or secondly, to incorporate the use of technology to aid in comprehension.

In relation to the first approach, a study on improving the public's understanding of legal documents generally, Wogalter et al., (1999) suggested that technicality should be decreased and explanations and definitions should be provided within documents. Becker et al. (2014) also found that providing visualisations as a means for communicating data privacy and security measures had a positive effect on trust and that participants perceived a higher level of ability in the visualisation condition which may have led to the suggestion Article 12 GDPR that individuals should include standardised icons with their information. This suggests that visualisations may aid users to comprehend the text of a privacy policy and could create transparency, by reducing the complexity, which may encourage users to read them. Robinson, et al., (2009) has also recommended that Data Protection Authorities, with guidance from the European Data Protection Supervisor, should be encouraged to develop privacy policies comparable to the model for intellectual property right licences. This Creative Commons model has established certain types of licences, which can be communicated to users through short, easy to understand descriptions (e.g. "attribution", "non-commercial", "no derivative works"). It was suggested that a comparable approach could be adopted with regard to privacy policies, by providing summary notices based on standardised descriptions which could be relatively easy for interested consumers to understand (ICO, 2009).

In relation to the second approach to research, much work has been done within the Human Computer Interaction (HCI) Community to improve the transparency of data processing. The Usable Privacy Research Project (2016) has focused on developing methods and techniques to semi-automatically analyse privacy policies with crowdsourcing, natural language processing and machine learning to make different types of data practices more transparent. Because of the difficulties of expressing privacy Policies in natural language, in 2002 the Wide Web Consortium (W3C) developed and officially recommended the Platform for Privacy Preferences Project (P3P). P3P is a protocol, which allows websites to express their privacy policies in machine-readable XML format (Olurin, Adams and Logrippo, 2012). However, despite remaining one of the most widely used structured privacy policy languages (Dong et al., 2011) it achieved limited adoption due to its complexity (Schwartz, 2009) and lack of industry participation (Cranor, 2012). P3P has since remained stagnant (Cranor et al., 2006) although Olurin, Adams and Logrippo, (2012)

believe that P3P-based techniques have considerable potential with the challenge being to design formalized privacy policy languages. In particular, Hogben (2003) and Anton et al. (2004) have identified a number of shortcomings of the P3P specification, with the absence of formal semantics being one of the most crucial. If these issues could be overcome, and such a language could be achieved, an intelligent recommender system could also be used to help users make decisions about their online privacy, combining user data and privacy policies to provide recommendations for privacy management to the user (Rasmussen and Dara, 2014). Such a system does so by providing recommendations and warnings on privacy management to individuals based on their privacy preferences.

Work on understanding Privacy Policies within the Computer Science community has continued. Wilson, et. al (2016) looked to create and analyses a 'Website Privacy Policy Corpus' to distil Privacy Policies into ten data practice that Privacy Policies should include. Sarne., et. al. (2019) then looked at unsupervised topic extraction from privacy policies based on the topics suggested by Wilson et. al., (2016). Maurel and Pardo (2020) have explored what they referred to as the three facets of Privacy Policies: natural language, graphical and machine-readable privacy policies, and Guo and Birrell (2020) have further investigated tools for visualising Privacy Policies. This indicates that the research into Privacy Policies and their ability to support the aims of the obligation to inform continue to be an important area of research.

2.3 Chapter Summary

In this chapter I have discussed the elusive concept of privacy and in particular how technological advancements have increased the threat to the information privacy of individuals because of the effect it has had on their ability to control information about themselves. I have also discussed how data protection and privacy laws have been introduced in the European Union in an attempt to neutralise this threat, and that a key principle of these laws that of 'transparency' of personal data processing. This principle requires organisations processing personal data to be transparent about how they are doing this. One of the key methods of ensuring transparency in the EU data protection Framework is by requiring data controllers to provide individuals with information about how their personal data is being processed, which is known as the 'obligation to inform'. I have also discussed that in practice, the de facto approach to satisfying this obligation has

been for organisations to provide a publicly available privacy policy, which includes the information the obligation to inform requires them to provide. However, privacy policies have been heavily criticised and proven to be ineffective in practice. This begs the question of how these can be improved, to increase the transparency of personal data processing, which allows individuals to have control of their personal data and to protect their information privacy. Whilst various suggestions have been made for improvements to privacy policies, these have not led to universal adoption and have often been abandoned. The gap this research looks to fill is to investigate deficiencies in the current approach to transparency in practice and to propose an improvement to the way that organisations can be transparent about their personal data processing.

The next chapter looks further at **RQ1**, which investigates the phenomena of privacy policies in practice and in particular, the extent to which privacy policies are similar enough to be standardised.

Chapter 3 The Phenomena of Privacy Policies

3.1 Introduction

As discussed in Chapter 2, technological and computational advancements have facilitated the ability to process personal data on a scale like never before. This dramatic increase in the ability to harness the potential of personal data (Rowland et al., 2012) has led to concerns over the effect of this on the privacy of individuals. To some extent, data protection laws have been created in an effort to strike the balance between the rights of individuals to privacy and the ability of organizations to utilise the personal data of individuals (Rowland et al., 2012), with the obligation to inform requiring organisations to provide individuals with information about how they process personal data. This obligation has led to the adoption of privacy policies by organisations as the de facto means of compliance with it. However, despite the many benefits of a well-executed privacy policy in theory, their current role in increasing the transparency of data processing has been heavily criticized (McDonald and Cranor, 2008). Additionally, whilst many improvements have been proposed and demonstrated to have an effect, these have not been adopted widely by organisations in practice.

It is because of these issues that the ultimate Research Goal of this thesis is to investigate the deficiencies in the current approach to transparency in the context of the obligation to inform in practice and to propose an improvement to the way data controllers can be transparent about their personal data processing. Given this background and goal, this thesis begins with an investigation into privacy policies in their current form. To achieve the goal of increasing the transparency of personal data processing, it was important to undertake a preliminary study to understand more about these phenomena, to understand more about the policies and where the friction in making the proposed improvements to privacy policies lies.

This chapter describes the preliminary study that undertaken as part of this thesis. The study was presented as part of a research paper at the 24th International Conference on the World Wide Web (Cradock, Millard and Stalla-Bourdillon, 2015) and then subsequently extended and published in a Journal paper with The Journal of Web Science (Cradock, Millard and Stalla-Bourdillon, 2016). The overall Research Question for this study was:

RQ1: Are the privacy policies of Social Networking Sites (SNS) similar enough in the information they provide about their personal data processing for the standardization of privacy policies to be possible?

This Research Question was designed to understand whether privacy policies shared enough information attributes for standardization of the information they contain to be possible. Given the overall Research Goal of this thesis, standardization was chosen as the potential improvement to investigate at this point. This was because not only would it be a standalone improvement that could be made to privacy policies, it is also one that could begin the groundwork for other improvements as discussed in Section 3.2.2.

In case the study found that standardisation was not possible, the study also investigated whether there was any information that SNS included in their privacy policy that went beyond the recommendations of the UK Regulator (The Information Commissioner's Office), or whether there was any information that the UK Regulator recommended should be included in a Privacy Policy, but which the SNS had omitted. This was to support the findings on standardisation but also to understand more about the privacy policies and the information they provide in comparison to the recommendations of the UK Regulator as a check on the guidance that is provided by the Regulator.

The rest of this chapter is structured as follows: it begins by explaining why SNS were selected as the provider of the privacy policies for comparison and then explains why the standardization of privacy policies was chosen initially as the potential improvement to be investigated. It then goes on to give a summary of the literature on the comparison of privacy policies that was available at the time of the study and then following this, it discusses the specific sub-research questions and the methodology used. It then finishes with the results and findings of the study, including the similarity of the privacy policies under investigation, the potential for standardization of privacy policies based on this and the differences in the information provided by SNS and the recommendations of the UK Regulator. Finally, it highlights a key finding of this study which directed the research for the rest of this thesis.

3.2 Background

3.2.1 Social Networking Sites

Whilst all websites, regardless of type, from search engines to e-commerce sites, would benefit from the improved transparency of their personal data processing in their privacy policies, for this investigation one 'type' of website was chosen. The reason for this was to make sure that similarity was measured fairly by reducing the possible effect of confounding variables. Such variables could be caused because different types of websites may process personal data in different ways.

Social Networking Sites (SNS) were chosen as the type of website from which privacy policies were selected because they were the second most frequently visited 'type' of website globally at the time of the study (after search engines) (Ipsos, 2013; Alexa, 2014).

They were also chosen because they have a reputation for abuse of user privacy. SNS are a product of Web 2.0, which changed the relationship users had with the web, from a passive one where they could only view information, to one where users could upload and download information. SNS (which are also referred so as Social Media Sites) were at the forefront of this move to Web 2.0 (Rogers, 2011: 213) and allow users to communicate with each other and to upload and share content with other users.

SNS are often free for individuals to use, which is also an influential factor in their popularity (Sellars, 2011). However, because SNS are still businesses, the trade-off for the free use of their services is the personal data that they harvest from users. This user data can then be monetized via various processes (such as targeted advertising) to support the provision of the service. However, despite this trade-off for free use, a 2011 survey (Special Eurobarometer 359, 2011.) found that 72% of SNS users worry that they are giving away too much data online. The European Commission (2014) cite the example of an Austrian law student who requested all the information held about him from a SNS, which returned 1224 pages.

Indeed, it is not just the data that users knowingly share with the SNS, or the data that is observed about them without their awareness that SNS can access, but also the information that SNS can derive and infer about users from seemingly innocuous data. A 2013 study found that Facebook 'likes' (which were publicly available by default at the time, through Facebook's API) could be used to accurately predict a variety of attributes about

an individual, including some attributes which individuals would consider to be sensitive. Indeed, the attributes that could be predicted included ethnic origin, religious beliefs and sexual orientation (Kosinski et al., 2013). This issue of being able to use innocuous information to generate other information about individuals, and in particular special category or sensitive data, was also recognized by an OECD roundtable of 65 privacy experts from governments, privacy enforcement authorities, academia, businesses and the Internet technical community from around the world in 2014. They acknowledged that '*increasing amounts of data are not collected from the individuals concerned, but are instead observed, derived and inferred*' (OECD, 2014). These findings highlight the issue of information asymmetry, where organisations know much more about the personal data they are processing than the individuals whose personal data it is.

Thus, although SNS rely on personal data, both as a by-product of their service and because of their business model, it is questionable how much personal data this entitles them to. This is especially so given the criticisms they have received regarding their collection and use of personal data (Anderson, 2009). The growing concern SNS have attracted regarding their effect on user privacy typifies the need for them to have more informative and transparent privacy policies. Thus, improving the ability of SNS to communicate information to individuals about their collection and use of their personal data has the ability to support users in making better judgments on how much information they are willing to surrender to gain free use of these services.

3.2.2 Standardisation

As discussed in Chapter 2, there are various potential approaches to improving the transparency of personal data processing in practice. These include both technical solutions and organizational ones. In particular, there are various suggested improvements that could be made to privacy policies, to increase the role they play in the transparency of personal data processing. As a suggestion for improvement, the standardization of privacy policies offers a lot of potential. It offers benefits to various stakeholders, including both the consumers and producers of privacy policies, as well as beginning the groundwork for other suggested improvements to transparency. Indeed, a United States of America (US) Federal Trade Commission (FTC) report (Commission, 2010) called for privacy policies to be clearer, shorter and more standardized.

In relation to users of websites, the benefits of the standardization of privacy policies are also various and include to support the user in making comparisons between the policies of different websites. Standardisation also supports increasing familiarity with the terminology used within privacy policies, by making sure websites use the same terminology in consistent ways. It can also allow for the easy location of particular information within a privacy policy, which could allow an individual to quickly locate information about an aspect of personal data processing that they have a particular concern or question about (Cranor, 2012). Studies have also shown that standardized presentations of privacy policies can have significant positive effects on a reader's enjoyment of reading policies compared to non-standardized presentations of them (Kelley et al., 2010).

There are also various benefits for organisations of a standardized privacy policy, including allowing them to verify their compliance with the law (Cranor, 2012) but also a reduction in the hassle of creating policies completely from scratch as it would give them a clear standard to follow in the production of these. Whilst these benefits may not be attractive enough on their own to incentivise organizations to improve their policies in the same way consumer demand might, they do reduce the deterrents or roadblocks to them doing so.

Standardization also allows for large-scale analysis of privacy policies (Cranor et al., 2013) which supports regulators and researchers. This supports regulators in various aspects of their role, by enabling them to assess policies for compliance more quickly, to gain a better understanding of them in general and also to move away from the human annotation that is currently required to understand and compare them.

Standardizing elements of privacy policies also begins the groundwork for other suggested improvements which can be made to privacy policies. For example, it begins the process of information reduction and refinement, which is required when developing formalized privacy policy languages (Olurin et al., 2012). It is also the first step towards standardizing descriptions within policies, which could allow for a creative commons model approach to the transparency of information about personal data processing to be utilized (Robinson et al., 2009). However, despite all these potential benefits, to succeed, standardization requires policies to share attributes, upon which standards can be built. Given the fragmented evolution of the privacy policies of SNS, in their creation by different organizations, potentially governed by differing jurisdictional legal requirements, it cannot

be assumed that the shared attributes required for standardization are present. Thus, prior to any attempt at standardization, it is important to examine the data in question as assess for similarity. This allows a researcher to ascertain whether standardization is immediately possible, and if so at what level. It also allows the researcher to see where the similarities and differences between the privacy policies lie and understand more about the information they include in general. This is important as much of the research has taken a top-down approach to examining privacy policies, looking at them through the lens of legal requirements or other frameworks rather than a bottom up approach of looking at what they include.

3.2.3 Comparative Analysis of Privacy Policies

In looking to compare the Privacy Policies of SNS, it was important to investigate the current literature on comparative analysis of Privacy Policies. This enables understanding of the different approaches that have been taken to assess this in practice. However, at the time of the study there was not a vast amount of literature on the comparison of privacy policies (of SNS or otherwise) although the studies of McRobb and Stahl (2007) and Yee and Korba, (2005) do look into this. One study that did look at this was and was of interest is Wu et al (2010). Here, the researchers adapted a privacy taxonomy which had previously been applied to data storage policies, and then extended it to the privacy policies SNS. Wu et al (2010) applied the privacy taxonomy of Barker et al. (2009) to the policies of six SNS (Facebook, LinkedIn, MySpace, Orkut, Twitter, and YouTube) to compare how the published policies protected user privacy in reality. Based on the taxonomy, Wu et al (2010) asserted that privacy policies are formed by four elements (purpose, visibility, granularity and retention), all of which centre around the personal data involved. However, despite adaptation, this taxonomy was still primarily aimed at providing a means for thinking about data privacy technologically (and specifically for data repositories) (Barker et al., 2009), opposed to thinking about Privacy Policies from a legal perspective. The taxonomy elements were created from principles of handling data from various sources, rather than from concrete legal requirements. Given that one of the main benefits of a standardized policy is to help organizations comply with the law more easily, this factor is also likely to also be a huge incentive for the adoption of a standardized Privacy Policy. Therefore, finding similarity with a complete set of elements indicating legal compliance is more appropriate for supporting conclusions about the potential for a legally compliant, standardized privacy policy and is more likely to lead to adoption in practice. Thus, for this

research a similar approach to Wu et al (2010) of comparing against a set of elements was followed, but a set of elements that would indicate legal compliance were used instead, which is discussed further in Section 3.3.3.

3.3 Research Questions, Sample and Methodology

3.3.1 Research Questions

As discussed previously, the overall Research Question (RQ1) for this study was:

Are the privacy policies of Social Networking Sites (SNS) similar enough in the information they provide about their personal data processing for standardization of Privacy Policies to be possible?

Because of the breadth of this question, it was broken down into five Research Sub-Questions (RSQ's):

RSQ 1. What is the similarity between the privacy policies of the top six SNS globally, in the clauses that they use?

RSQ 2. What is the similarity between the privacy policies of the top six SNS globally, in their coverage of forty recommendations of information to include in a privacy policy made by the UK Information Commissioners Office (ICO) in their Code of Practice (ICO Code)?

RSQ 3. Are there any recommendations of the ICO Code, which all privacy policies do not address?

RSQ 4. Are there any themes of information addressed in all of the privacy policies that were not included in the forty recommendations from the ICO Code?

RSQ 5. To what extent is standardization possible between the privacy policies of SNS?

3.3.2 Sample

As discussed in Section 3.2.1, SNS were chosen as the 'type' of website to investigate. To limit the data under study, it was decided that the most frequently visited SNS would be chosen, as ranked by Alexa.com (Alexa, 2014), a web analytics website that publishes a global traffic rank for major websites. As these SNS are used the most, any improvement

to their Privacy Policies has the potential to benefit the most individuals. Alexa allows visitors to browse websites by 'category' and their category 'Social Networking' was used for the purpose of this investigation. Due to the time available in which to manually analyze the data, a limited number of SNS could be chosen. In deciding how many to investigate, we found that the top five SNS were also ranked in the top thirty of all websites globally. Whereas, the sixth ranked SNS (Flickr), was ranked 164th. Because there was such a steep drop in the popularity of the ranked SNS, from the 5th to the 6th (and onwards), it seemed rational to investigate the top five ranked SNS in addition to the 6th. This would account for any confounding variables that might be linked to popularity. The 7th ranked SNS onwards were therefore excluded from the investigation. As a result, the six SNS Privacy Policies selected were those of: Facebook (FB) (Facebook., 2013), Twitter (T) (Twitter, 2014), LinkedIn (L) (LinkedIn., 2014), Pinterest (P) (Pinterest., 2014), Google+ (G+) (Google., 2014), and Flickr (F) (Flickr., 2015) as shown in Table 1. Their policies as available in August 2014 were analyzed.

Table 1 Further Details of SNS Selected

Social Networking Site	**Alexa Rank**	**Headquarters Location**	**Year of Release**
FB	1st	Menlo Park, California, US	2004
T	2nd	San Francisco, California, US	2006
L	3rd	Mountainview, California, US	2003
P	4th	San Francisco, California, US	2010
G+	5th	Mountainview, California, US	2011

Alexa (Alexa., 2014) does not provide exact metrics on the number of users worldwide, but instead ranks websites by traffic estimates based on data from their global traffic panel. Their global traffic rank is a measure of how a website is doing relative to all other sites on the web over the past three months. It is calculated using a proprietary methodology that combines a site's estimated average of daily unique visitors and its estimated number of pageviews over the past three months. The metrics of all six sites were estimates, because the SNS had not implemented the Alexa on-site analytics or published their results. Where this is the case, Alexa shows estimated metrics based on traffic patterns across the web as

a whole, identifying these patterns by looking at the activity of web users throughout the world and using data normalisation to correct for any biases. However, the more traffic a site gets, the more data Alexa has to calculate estimated metrics, meaning that the closer a site is to being ranked number one, the more reliable the estimates are. Given that the top six are being investigated, this increases the validity of their ranking.

Finding comparable and exact numbers of SNS users worldwide is not an easy task. This is because the definition of 'users' can be interpreted in different ways and broken down into various subcategories. For example, 'registered users' who simply 'have' an account but do not use it and 'active users' who have an account and use it frequently. Furthermore, 'active users' can be defined using different parameters. Statista (statista ,2016) detail the number of active users worldwide as of January 2016 for four of the SNS investigated: Facebook (1550 million); Twitter (320 million); LinkedIn (100 million); and Pinterest (100 million). However, they do not provide numbers for Google+ and Flickr. Finding user numbers for these SNS for the same period and using the same 'user' definition is difficult.

Flickr did state in June 2015 (Flickr, 2015) that they have a *'community of more than 112 million photographers'*, but this figure is not for the same period, and also does not say how it was measured. If it is the number of registered accounts, it is not comparable to statista's figures and even if it is for 'active users', it cannot be deemed comparable without details of the parameters used to calculate this which were not listed. The effect of how 'user' is defined is exemplified in the context of Google+. This is because for every Google account created, a Google+ profile was created automatically. Thus, counting 'registered users' could even less reflective about the user base of Google+ than other SNS as there may be many users with Google+ accounts that had never used them. Indeed, a digital marketing firm (Stone Temple Consulting., 2015) analyzed 516,246 randomly selected Google+ profiles and found that 90.1% of these had never posted anything on the service. It is worth noting that this study was conducted in 2014 and Google+ was discontinued in April 2019.

Another limitation made to shape the sample was due to the fact that online privacy policies often take a 'layered approach' and use hyperlinks to link to further explanatory information as depicted in Figure 1. For example, within their policy, LinkedIn stated that 'You may choose the parts of your profile that search engines index or completely opt out of this feature in your LinkedIn account settings'. When clicked, 'settings' redirects the user to their account settings. For this investigation, we had to decide whether the content we

assessed for similarity would extend to the information contained on the pages following these hyperlinks (second layer). At the familiarization stage of the investigation (Stage 1 discussed further in Section 3.3.5) we examined the hyperlinks. However, we found that they were generally links to more advice on the topics discussed (from the SNS itself and outside sources); links to account settings and other pages on the website; other SNS policies; other SNS services; and online contact forms. Furthermore, the UK Information Commissioner's Office have stated that when using a layered approach, the first layer of a privacy policy should contain the 'key privacy information' with 'more detailed information available elsewhere' (Information Commissioner's Office., 2016), opposed to new information. Therefore, we took a pragmatic approach and examined only the 'first layer' of the policies, defined as the content on the page when their 'privacy policy' or 'privacy' links were clicked on.

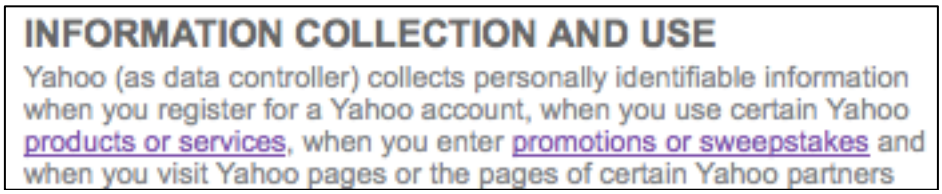


Figure 1 Excerpt from Yahoo (Fickr's) Privacy Policy

It is also worth noting that Facebook did not provide a 'first layer' at the time, as when you clicked on the 'Privacy' tab on their homepage, it showed a screen that broke their Privacy Policy into various sections as depicted in Figure 2. Therefore, the 'view complete data policy' was selected and treated as the first layer. Google also offered a 'download pdf' option as depicted in Figure 3 which was used as it represented the most comprehensive first layer that was available. All other 'first layers' were treated as the first screen shown when the 'Privacy Policy' link was clicked.

Facebook, Pinterest, Twitter and LinkedIn all had a privacy policy specific to their SNS, whereas, Google and Yahoo (who operates Flickr) had generic privacy policies covering all of their services. Yahoo did also offer a specific page on 'Flickr' as a service, but given that this was three hyperlinks away from the Flickr homepage and the second link was buried in the privacy policy (Figure 1) and so a decision was made not to include this in the data sample, because it was not part of the first layer.

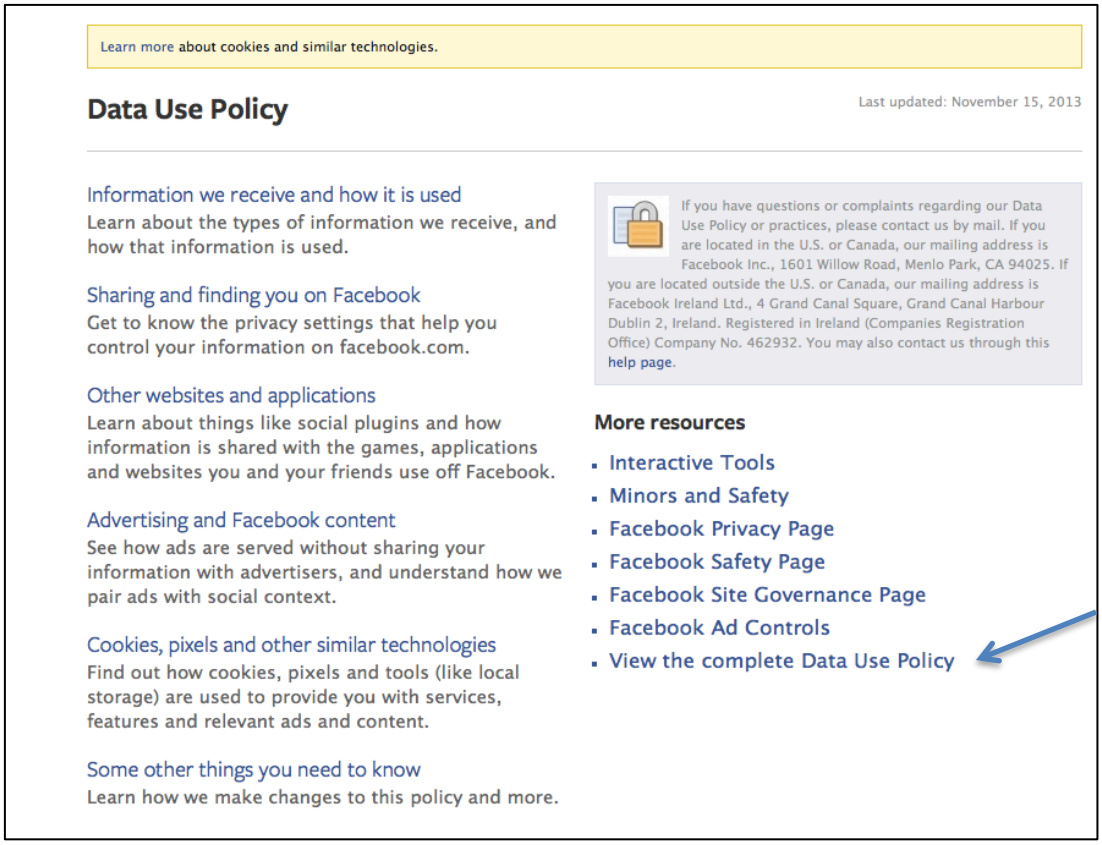


Figure 2 Facebook's Data Use Policy Page

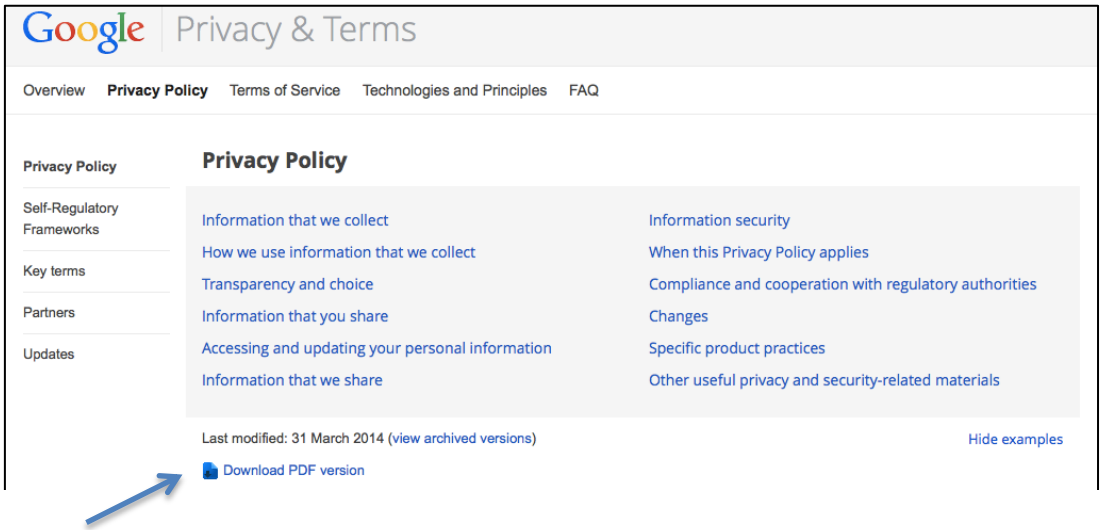


Figure 3 Google's Privacy Policy Page

3.3.3 Attributes for Comparison

With the aim of the investigation being to measure similarity as a precursor to standardization, it was important to select the appropriate attributes to measure and to compare on. Although similarity could be measured using various different attributes e.g. length, number of sections etc., here it would not provide a meaningful measure of similarity upon which the potential for standardization could be assessed. In terms of granularity, the clauses used by the SNS proved an appropriate attribute to compare, as

they convey enough information to make comparison meaningful. Indeed, investigating whether certain clauses were common or whether privacy policies consisted of largely bespoke clauses is of interest in efforts to standardise them. A 'clause' is defined as 'a part of a treaty, law or contract' (Oxford University Press., 2011).

For comparison, a second attribute was also measured for similarity and the coverage of forty recommendations from the ICO 'Privacy notices code of practice' (Information Commissioners Office., 2010) was chosen. Because the world is divided into legal jurisdictions, it was important to select one jurisdiction from which we could extract an appropriate subset of legal recommendations. The EU proved interesting, because of its single omnibus law at the time, the Data Protection Directive (DPD), which was aimed at harmonizing data protection laws throughout EU member states. This law meant that findings in relation to one EU country should be more generalizable to other countries within the EU, than to those outside of it. However, as discussed in Chapter 2 implementations of the DPD at this time differed between Member States, who decide the means to achieve the DPD's aims. In fact, the implementations of the DPD's obligation to inform has varied significantly between Member States in terms of what information should be provided, in what form, and at what time (Information Commissioner's Office., 2010). Because of this, one Member State's legislation and implementation was chosen to examine.

Whilst any could have been chosen, the UK seemed appropriate because the researchers were familiar with its legislation and were aware that the UK Regulator had (ICO) produced specific guidance aimed at the creation of legally compliant privacy policies, as discussed further below in Section 3.3.5. Furthermore, comparing the Privacy Policies produced by SNS all of which were headquartered in the US with legal requirements from a jurisdiction where they are not headquartered provides for an interesting juxtaposition. Indeed, conclusions about the possibility of a global standardized policy will be strengthened where similarity between policies originating from the US and recommendations for compliance based on UK and EU law can be found.

The UK Information Commissioner's Office (ICO) is an independent authority, set up to uphold information rights in the public interest in the UK. In an effort to help make policies more informative and in the role of the Information Commissioner to promote good

practice and compliance under Section 51 of the UK Data Protection Act, 1998 (DPA, 1998) ICO issued a 'Privacy notices code of practice' (Information Commissioner's Office, 2010) aimed at helping organizations 'collect and use personal data appropriately by drafting clear and genuinely informative privacy notices'. The ICO Code provides recommendations, aimed at aiding those processing information in complying with the DPA (Britain, 1998).

The ICO Code itself states that it can be used as a 'checklist to evaluate an existing privacy notice', which is why its recommendations were chosen as one of the attributes for this investigation. Whilst the code is not legally binding, the Information Commissioner at the time stated that he would take the standards of the Code into account if he received a complaint that information has been collected in an unfair or unreasonable way (ICO, 2009:4). The basic legal requirement was to comply with the DPA 1998, which organisations may have been able to do through alternate methods but following the Code would help organisations meet their legal obligations, by making sure organisations collect and use information fairly and transparently (ICO, 2009:7). Thus, comparing the privacy policies in this study for the presence of these recommendations is more appropriate for supporting conclusions about the potential for a legally compliant standardized policy than comparing them for the elements used in the Wu et al. (2010) study alone. Whilst the four elements Wu et al. (2010) used are reflected in the ICO Code, the Code provides further, more specific recommendations, to aid compliance with UK (based on EU) law.

It is worth noting that the UK and US are both members of the Organisation for Economic Co-operation and Development (OECD), which is an intergovernmental economic organization with 37 member countries founded in 1961 to stimulate economic progress and world trade. The OECD identified eight Privacy Principles of good practice in 1980 that were then updated in 2013 (OECD, 2013) and because of these shared principles which influence data protection and privacy legislation in OECD Countries, the UK and US could be predisposed to similarities in their approaches to privacy policies. However, both countries do have very different policy contexts when it comes to data protection and privacy as the US currently has no single comprehensive federal (national) law regulating data protection and privacy. Whilst, the US Federal Trade Commission's (FTC) Fair Information Practice Principles (Commission., 2000) have significantly shaped how privacy policies are written by US web companies, specific guidance on policies equivalent to ICO's (in stating exactly what they should contain) is often sector-based. For example, the model privacy form (Securities and Exchange Commission., 2009) which has been designed for

compliance with the US Gramm-Leach-Bliley Act is aimed at financial institutions, as this is whom the Act regulates and may not be relevant for (or include all the information required for) transparency in other sectors. These differences are another reason why investigating the similarities between the privacy policies is interesting, but also why ICO's sector neutral guidance is a more appropriate as a set of requirements for comparison. The assumption going into the study was that guidance from the independent body charged with upholding information rights in the UK should lead to compliance with UK law and therefore, provided the most appropriate set of recommendations for this investigation.

3.3.4 Methodology Overview

To study the similarity between the privacy policies, a combination of Thematic Analysis (Braun and Clarke, 2006) and Cross-Document Structure Theory (CST) was used to identify the clauses and place these into themes. Jaccard's Similarity Co-efficient was then used to provide a measure of similarity.

Cross-Document Structure Theory (CST) (Aleixo and Pardo, 2008) is a formal discourse theory for multi-document analysis, which establishes relationships among segments of different documents about the same topic. Human annotation is then used to assign similarities between texts (Zhang, Otterbacher and Radev, 2003). CST has two possible classification scenarios, binary and full. Binary classification is simply interested in the existence of cross-document relations, regardless of type. Whereas, full classification cares about the type of cross-document relationship and classifies the relationship as one eighteen defined relationships. Examples of these relationships include subsumption, identity and citation. Whilst this theory and methodology offered many benefits, it is recognised that it cannot be used to solve the AI-complete problem of what two pieces of text 'mean', offering only a heuristic approximation (Zhang, Otterbacher and Radev, 2003).

Therefore, as a second method Thematic Analysis (Braun and Clarke, 2006) was also used. This method is used to pinpoint, examine and record themes within data, and occurs in the six standard stages of Thematic Analysis outlined in Section 3.3.5. More detail of what happens at each stage is included in Section 3.3.5 which discusses how these stages were adapted for this investigation. Unlike CST, Thematic Analysis goes beyond the text and looks for implicit and explicit ideas within it (Guest, MacQueen and Namey, 2012). The process of data analysis can either occur inductively (where the themes emerge from the

text) or deductively (where the data analysed based on preconceived themes). For this investigation, only binary CST classification was completed and the coding and analysis in Stages 1-5 (as discussed in Section 3.5.5) was completed by the primary researcher only, taking a deductive approach using recommendations from the ICO Code, based on a framework agreed by all researchers which can be found in full in Appendix 1 and examples of which are depicted in Table 2. The coding and analysis was then discussed with the other researchers to produce the Stage 6 Final Report which can be found in Sections 3.4, 3.5 and 3.6.

Similarity for both attributes was measured using Jaccard's Similarity Coefficient, a statistic used for comparing the similarity and diversity of sample sets and results in a percentage similarity. It is defined as the size of the intersection divided by the size of the union of the sample sets (Jaccard, 1912) and the formula for it can be found in Figure 4.

$$s = \frac{|Q \cap D|}{|Q \cup D|}$$

$$= \frac{p}{p + q + d}$$

where
p = number of variables that are positive for both
q = number of variables that are positive in Q but not D
d = number of variables that are positive in D but not Q

Figure 4 Jaccard's Similarity Co-efficient formula

3.3.5 Methodology in Practice

The six stages of Thematic Analysis were undertaken as follows:

Stage 1: Familiarization with data. Here researchers immerse themselves in the data, gaining familiarity with its depth and breadth (Braun and Clarke, 2006). Therefore, in this stage the privacy policies were read multiple times, first passively then actively, to recognise meanings and patterns to support the subsequent phases of analysis.

Stage 2: Generating Initial Codes. This phase involved the production of initial 'codes' from the data. 'Codes' are defined as 'the most basic segment of the raw data that can be assessed in a meaningful way regarding the phenomenon' (Boyatzis, 1998). Unlike some legal documents, such as contracts or Acts of Parliament, which are broken down into numbered clauses, privacy policies are only broken into sections, which meant that the clauses had to be identified for the purpose of this investigation.

As the Thematic Analysis definition of 'code' and the definition of 'clause' (used above in Section 3.3.3) were compatible, this stage was used to identify the atomic clauses in the privacy policies. The policies were initially divided into sentences and beginning with Facebook's privacy policy (as the longest policy), a table was created, initially treating each sentence as a clause. A little of the surrounding data was kept in the table if relevant so that context was not lost (Bryman, 2001). Thus, anything in the table included in the working document in italics and brackets showed that it was supplementary information from another clause to provide context.

Following advice regarding this stage (Braun and Clarke, 2006), as many clauses as possible were 'coded' i.e. created from the privacy policies. Although this meant destroying the structure of the policies (by rearranging and splitting sentences) this and the resulting number of clauses did not matter because the purpose of the investigation was focused on similarity in (and a full picture of) the information conveyed in privacy policies by clauses opposed to focusing on measuring other factors about privacy policies for comparison, such as their length or the number of clauses the contain.

The clauses resulting from this stage formed the initial list of clauses for each Privacy Policy, the results of which can be seen in Table 3 and produced a total number of clauses of 986 clauses across all six policies (although the clauses had not been checked for duplication at this point). Here a technique from CST was introduced and sentence pairs were examined (Ryan and Bernard, 2003), similar to the Thematic Analysis 'compare and contrast' approach (Glaser, 1978). All policies were compared against each other and sentence 'pairs' was compared (one sentence from a policy compared with every sentence from another policy), individually asking each time:

• What is the sentence about?

• What question is it trying to answer?

• Is it equivalent to the current examined clause in these respects?

• Would adding or subtracting other information from the same Privacy Policy make the clause equivalent?

As part of this comparison, sentences were then examined to see whether multiple sentences needed to be combined to form a clause, or whether multiple clauses were contained within one sentence. For, example, Pinterest, had the sentence 'We may change this policy from time to time and if we do we'll post any changes on this page' and Google+ addressed this in two separate sentences 'Our Privacy Policy may change from time to time.' and 'We will post any privacy policy changes on this page.'. Therefore, the Pinterest sentence was split into two clauses and coded to the corresponding Google+ sentences in the table.

Because of the amount of data in the sample, when it had all been coded into the table with the relevant codes, for accuracy, each privacy policy was checked line by line to make sure that it was either in this table or in the tables of removed content (explained further below). The table was then re-checked clause-by-clause going through each row to check that where multiple clauses had been coded as equivalents, that this was still the case. Where four or more SNS had a clause in common, the privacy policies of the SNS identified as lacking the clause were double-checked to ensure that this was not due to error or omission. The data in the table was repeatedly checked until no more codes/clauses needed to created or moved, also known as achieving *theoretical saturation* (Strauss and Corbin 1990: 188).

As a result of breaking the policies down into atomic clauses, each clause could only be coded once (i.e. only be classed equal to one other clause), unlike other applications of Thematic Analysis, which code individual extracts of data into numerous codes. The reason for this was that firstly, if a clause was conveying the same information as two codes then they should all be treated as the same. Secondly, part of the motivation in the investigation was to get a full list of individual clauses (although the number of clauses was not important, the number of types of different information was) and therefore using a single clause twice might affect this. This came into issue if multiple meanings could be inferred and the clause could not be split further to allow it to address two clauses. In this situation,

no inference was made and the clause was placed with the clause it was most like lexically or treated as a new clause if it was too different. For example, different SNS used the words close, deactivate and delete regarding the closure of an account. Although overlap could be inferred, these were treated separately. However, for the most part, when differentiating between clauses it was generally less about the words used and more about what the policy was trying to convey in the sentence, which is why thematic analysis was used in addition to some elements of CST, so the meaning of the words was not completely overlooked.

Also, during this stage, some information from the policies was removed, including duplicate clauses in the same policy, sub-headings mentioned in the body of the subsequent text and sentences which preceded lists. For example, the subheading 'your information' was removed from Facebook's policy when the first line in the section began 'your information is'. However, 'Information for users outside the United States and Canada' was left in because following this, only contact information was provided, thus the subheading was required for context. The logic of removing these was to normalize the data. Including them would have inflated the number of clauses some SNS had and skewed the results. The removed items were kept in a separate table for data checks and because they could be of interest in future studies concerning privacy policies e.g. looking at the effect of repetition of clauses in privacy policies and the impact this has.

The results of this stage produced 669 individual clauses which contained the 986 individual clauses as shown in Table 4.

Stage 3: Searching for Themes Among Codes. This phase refocuses the analysis at broader themes, and in this study involved sorting the clauses into potential themes (Braun and Clarke, 2006). Codes differ from themes, which are often much broader and identify what data means (Braun and Clarke, 2006). Rather than coding inductively and creating themes from the clauses alone, it is at this stage that forty ICO Code (Office., 2010) recommendations were used as themes, into which the data was placed for the purpose of RSQ's 2-4. The ICO Code states that it can be used as a list for organisations to check their privacy policies against for compliance. Because of this the ICO Code was parsed manually and forty-six recommendations were identified using the process of Stages 1 and 2. Seven were identified as too broad or vague to assess in the context of this investigation e.g. include '*Any further information necessary, in the specific circumstances, to enable the*

processing in respect of the individual to be fair'. These seven were removed, leaving thirty-nine themes. One of the 39 themes was then split into two themes, which left forty themes for analysis.

Each one of the clauses from Stage 2 was then placed into at least one of the forty ICO Code recommendations or placed in a category of 'miscellaneous' if the clause did not fit into one of the themes. To create an objective description of what is meant to 'address' an ICO Code recommendation and because the analysis was only being done by a single researcher, a Table containing all of the recommendations and a definition for thematic coding was created, with an example of the clause we would code to this theme. Table 2 shows an example of three of these and then Appendix 1 shows the full table.

Table 2 Example from Coding Table used to Code Clauses into ICO Recommendations

ICO Code Recommendation	**Definition for Thematic Coding**	**Example of Clause**
Tell people how long you or other organisations intend to keep the data.	The privacy policy refers to how long it (or organisations it shares the data with) intend to keep data for.	Typically, information associated with your account will be kept until your account is deleted". **Facebook**
Tell people who their information will be shared with/disclosed to	The privacy policy advisers who user's information will be shared with/disclosed to.	"Secret boards are visible to you and other participants in the board, and any participant may choose to make the contents of the board available to anyone else." **Pinterest**
Tell people the purpose for using the information.	The privacy policy tells the user the purpose for using the information.	If you email us, we may keep your message, email address, and contact information) to respond to your request **Twitter**

Unlike in Stage 2, where each clause could only be coded once, in this stage it did not matter if the clauses were coded into more than one 'theme'. The reason for this was that this stage of Thematic Analysis was used to normalise the data for when similarity would be compared to answer RSQ's 1 and 2. Because of the differing lengths in the privacy policies and the different functionalities of SNS, it was clear that a number of clauses were bespoke to individual privacy policies. Whereas, if similarity was also measured by seeing whether a privacy policy includes at least one clause for each recommendation made in the ICO Code for comparison, differences in length or functionality would not skew the results.

As the focus of this investigation was to see whether the privacy policies contained clauses addressing the recommendation, we did not investigate whether the SNS were legally complying with them. For example, one of the forty ICO code recommendations was: *'Obtain assurances (in form of written agreements) from any organizations you share personal information with about what they will do with the information and what the effect on people is likely to be'.*

Two clauses coded into this recommendation from LinkedIn's policy were:

- 'These third-party developers have either negotiated an agreement to use LinkedIn platform technology or have agreed to our self-service API and Plugin terms in order to build applications ("Platform Applications")'.

- 'Both the negotiated agreements and our API and Plugin terms contain restrictions on how third parties may access, store, and use the personal information you provide to LinkedIn'.

Although this meant that LinkedIn had included information in its policy 'addressing' the recommendation, it would take further investigation (outside the scope of this study) to assess whether the assurances obtained are in fact legally compliant. There are two reasons why this type of in-depth legal analysis is outside the scope of this study. Firstly, because of the time it would take to complete such an analysis on compliance with each of the forty recommendations. Secondly, because that was not the aim of this study, which is to assess similarity as a potential for standardization. Future work explored areas that were identified in this study for which further in-depth legal analysis would be beneficial.

The reason that coding was not done inductively at this stage, to produce themes from the clauses is because the purpose here was to assess similarity with the long-term aim of creating a standardized policy. Because one of the main benefits of a standardized policy for organizations is to help them comply with the law, this is likely to be a huge incentive in their adoption of such a policy. Adoption of a standardized policy will be more attractive and therefore more likely if it aids organizations in compliance with the law. Whilst generating themes from the policies may be useful for future work, here we wanted to see how well the recommendations could be used as a framework for the policies in their current form (although some inductive coding was completed as discussed below). It could be argued that coding the clauses into the recommendations risks creating a circular

argument in the research design, because the aim of the recommendations is to address the law, you would expect to find them in the policies. However, as studies (Van Alsenoy et al., 2015) experience and indeed our results in Section 3.4 show, this is not always the case. Furthermore, the juxtaposition of US policies being compared with UK (based on EU) law may also mean this is not the case. Thus, by coding the clauses into ICO Code recommendations, we can identify any recommendations warranting further investigation, due to their lack of presence in the policies.

Once this coding of the clauses into the recommendations was completed, the 'miscellaneous' category contained a number of clauses, which could not be allocated to an ICO Code recommendation. We then coded inductively on these clauses, which identified themes that the policies included that went above and beyond the ICO Code recommendations. This provided the answer to RSQ4 (Are there any themes of information addressed in all of the privacy policies that were not included in the forty recommendations from the ICO Code?). These results identified an interesting area for future work which formed the next part of this thesis (and is discussed further in Section 3.6 and Chapter 4) to understand why these were not present in the ICO Code, and what the implications of this are.

When analysis was complete all clauses were coded to a theme (either one of the 40 ICO Code recommendations or one of the themes produced from the miscellaneous category).

Stage 4: Reviewing Themes. This stage involves two levels. Level one involves reading the collated clauses for each theme and considering whether they form a coherent pattern (Braun and Clarke, 2006). If not, the researcher considers whether the theme is problematic or whether the data extract simply does not fit there, in which case, the theme can be re-worked. Level two involves a similar process, but in relation to the whole data set, where the validity of how individual themes connect to the data (Braun and Clarke, 2006). This stage, including the production of a thematic map (Braun and Clarke, 2006) was not required, although discussion regarding possible overlap in the recommendations of the code can be found in Sections 3.4 and 3.5. Because the themes used here were pre-determined from the ICO Code, at this stage each clause that had been allocated to a recommendation was checked for coherence. As were the themes generated from the 'miscellaneous' clauses.

Stage 5: Defining and Naming Themes. This stage names themes and paraphrases their content, clearly defining what themes are, and what they are not (Braun and Clarke, 2006). This involves organising the collated data extracts for each theme and organising them into a consistent account with accompanying narrative (Braun and Clarke, 2006). As the themes were pre-determined recommendations, they were simply named as their recommendation in full. The thematic codes we had already produced (examples provided in Table 2 and full list in Appendix 1) were used as their definitions. For the themes generated from the 'miscellaneous' clauses', these were named and defined, and are discussed below in Section 3.4. Defining the scope of each recommendation allowed the categorical statement of whether a privacy policy had addressed the recommendation or not. This allowed the use of Jaccard's Similarity Coefficient which relies on a yes-no binary clarification to measure the similarity between the privacy policies both in terms of the clauses the used but also in addressing the recommendations of the ICO Code in order to answer RSQ's 1 and 2, the answers to which is discussed in Sections 3.4 and 3.5 .

Stage 6: Producing the Final Report. The task of this stage is to tell the complicated story of your data in a way which convinces the reader of the merit and validity of the analysis (Braun and Clarke, 2006) using extracts embedded within the analytical narrative. This illustrates the story you are telling about your data, where necessary going beyond description to make arguments in relation to your research question (Braun and Clarke, 2006). This report can be found in Sections 3.4, 3.5 and 3.6.

3.4 Results and Analysis

This section is divided into four subsections focused on answering RSQ's 1-4 of the study, with the results used to answer these RSQ's included, analysed and discussed. The answer to RSQ5 is then presented in Section 4. Before that, a discussion of the results as a whole is presented, with a particular focus on what these results implicate regarding the motivations of the investigation.

3.4.1 General Overview of Results

Table 3 and Figure 5 display the results from Stage 2 of Thematic Analysis and shows that Facebook and LinkedIn 's 'first layer' included significantly more clauses than the other SNS. Google (ranked third in descending order of number of clauses) had less than half the

number of LinkedIn (ranked second). Interestingly, there is no direct relationship between the number of clauses identified and the number removed, indicating that increased policy length did not necessitate repetition. Table 3 and Figure 5 also shows that the descending order of SNS in terms of number of clauses identified and number of clauses remaining, stays the same (Facebook, LinkedIn, Google+, Twitter, Pinterest, Flickr). However, the order varies in terms of the number and percentage of clauses removed. Flickr in particular had just over a quarter of clauses removed which is a significant amount given only 89 were identified initially which was the lowest total number of clauses.

3.4.2 Similarity of Clause Coverage Between Privacy Policies of SNS (RSQ1)

Table 3 and Figure 6 show how many clauses were present in all six SNS through to how many were only present in one. Interestingly, it shows that only **2.09%** of all possible individual clauses identified were common to all six policies, with **75.93%** of clauses identified as bespoke to just one.

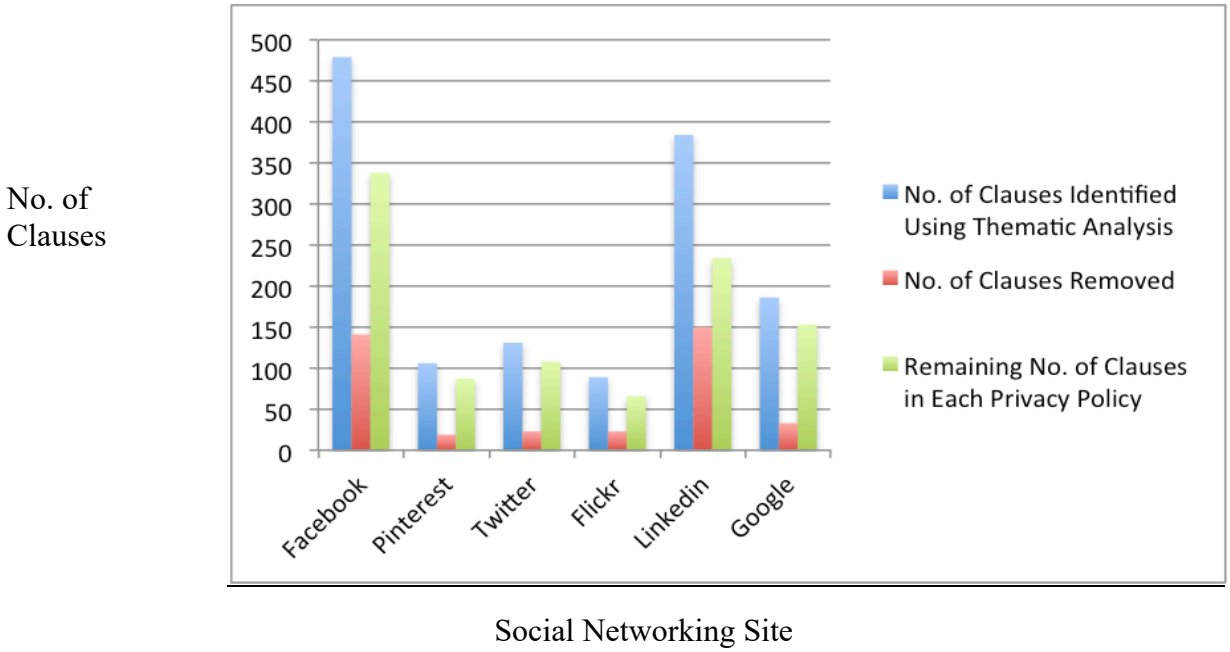


Figure 5 Number of clauses identified, removed and remaining

Table 3 Number of clauses identified, removed and remaining

	Facebook	Pinterest	Twitter	Flickr	LinkedIn	Google+	Total
No. of Clauses	479	106	131	89	384	186	1375
No. of Clauses Removed	141	19	23	23	150	33	389
% Clauses Removed	29.44	17.92	17.56	25.84	39.06	17.74	28.29
Remaining no. of Clauses	338	87	108	66	234	153	986

Figure 6 shows evidence of a power-law relationship between the number of clauses, and how many policies they appear in. A power law relationship is one where the frequency of an event varies with a power of some attribute of that event e.g. size (Clauset, Shalizi and Newman, 2009). Generally, as the number of clauses examined increases, the number of SNS they can be found in decreases. Although, there is an increase (rather than decrease) in the number of common clauses as the number of SNS increase from five to six. However, as Clauset, Shalizi and Newman, (2009:2) state, few empirical phenomena obey power laws for all values as often the power law only applies for values greater than some minimum, in which case it is stated that the tail of the distribution follows a power law.

Table 4 Clause Coverage by SNS

No. of SNS Clauses Were Present In	No. of Clauses	% of Individual Clauses
6	14	2.09
5	11	1.64
4	16	2.39
3	35	5.23
2	85	12.71
1	508	75.93
Total	**669**	**100**

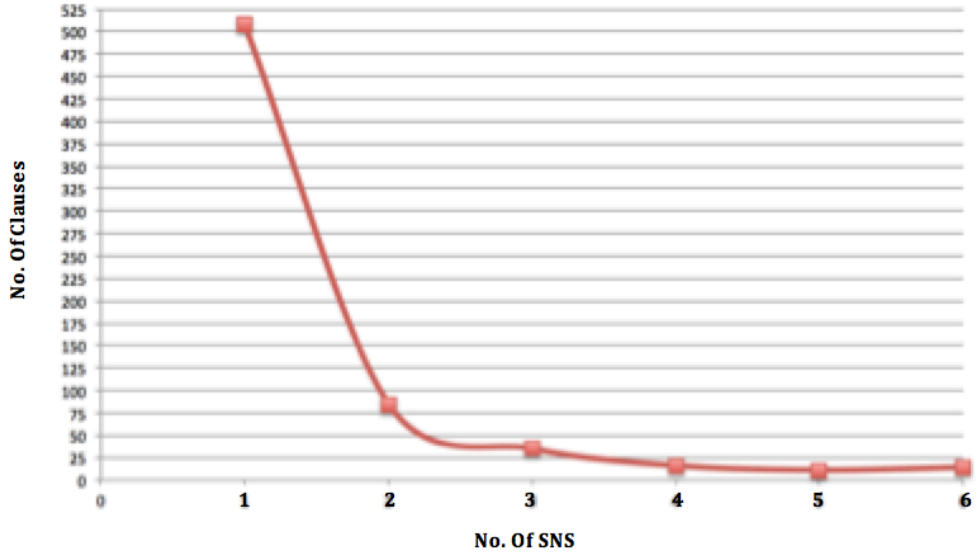


Figure 6 Clause Similarity Between the Privacy Policies of SNS

Table 5 Number of clauses identified, removed and remaining

	Facebook	Pinterest	Twitter	Flickr	LinkedIn	Google+	Total
No. of Clauses	479	106	131	89	384	186	1375
No. of Clauses Removed	141	19	23	23	150	33	389
% Clauses Removed	29.44	17.92	17.56	25.84	39.06	17.74	28.29
Remaining no. of Clauses	338	87	108	66	234	153	986

Table 6 Jaccard Similarity of Clause Coverage

	Facebook	Pinterest	Twitter	Flickr	LinkedIn	Google+
Facebook		0.09	0.14	0.08	0.15	0.10
Pinterest			0.27	0.18	0.13	0.19
Twitter				0.17	0.21	0.19
Flickr					0.11	0.15
LinkedIn						0.13
Google+						

In answering RSQ1, 'What is the similarity between the privacy policies of the top six SNS globally, in the clauses they use?' Tables 6 and 7 show the Jaccard Similarity (or Jaccard Index) and Dissimilarity (or distance) Coefficients between the privacy policies in terms of the clause coverage i.e. how alike they were in the clauses they used. As, described in the Section 3.3.4, this statistic is used for comparing the similarity and diversity of sample sets.

Table 7 Jaccard Dissimilarity of Clause Coverage

	Facebook	Pinterest	Twitter	Flickr	LinkedIn	Google+
Facebook		0.91	0.86	0.92	0.85	0.90
Pinterest			0.73	0.82	0.87	0.81
Twitter				0.83	0.79	0.81
Flickr					0.89	0.85
LinkedIn						0.87
Google+						

Table 6 shows that the similarity between the SNS in the clauses they use was low, with the range between **8-27%.** The least similar were Flickr and Facebook with 8% similarity and the most similar were Pinterest and Twitter with 27% similarity. Average similarity was 15%. Interestingly, Table 5 shows that Flickr and Facebook were at separate ends of the continuum in terms of number of clauses identified, with Facebook having the most and Flickr the least. This may explain their dissimilarity. Whereas, Table 5 shows that Pinterest and Twitter would sit next to each other on this continuum, with a similar number of clauses. This may indicate why they have a higher similarity.

The investigation highlighted three prominent reasons for differences between SNS in the clauses that they used and the reason for such a large number of bespoke clauses:

Firstly, differences in the functionality offered between SNS resulted in a large number of clauses devoted to communicating information about these. For example, LinkedIn were the only SNS to discuss the use of Polls, and Facebook alone offered a service called 'Instant Personalisation'. Other SNS would not include these clauses within their policies, because they do not offer the functionality. Therefore, they would not need to communicate information about these.

A second reason for a large number of bespoke clauses was due to semantics. Different words were often used between policies to discuss the same topics, but without being defined. For example, when discussing the termination of an account, the words 'close', 'delete' and 'deactivate' were all used across different policies. Whilst Facebook confirmed 'delete' meant permanent deletion, Pinterest only stated users had the ability to 'close your account at any time'. Without defining what 'close' meant, it was difficult to ascertain whether the clauses were comparable, meaning that they had to be treated as different which may have inflated the number of individual clauses.

Thirdly, some SNS elaborated on certain topics with more information than others, which resulted in more clauses. For example, although all SNS included a link to follow if users had any questions, comments or complaints, some also included their physical address and information regarding the complaints/question procedure. Additionally, some SNS provided definitions and examples of varying length and content for technical terms. For example, only Pinterest and Twitter elaborated on the definition of cookies to mean 'persistent' and 'session' cookies, which resulted in additional clauses, which were not present in other policies.

3.4.3 Similarity of ICO Code Recommendation Coverage Between Privacy Policies of SNS (RSQ2)

In answering RSQ2 (What is the similarity between the privacy policies of the top six SNS globally, in their coverage of forty recommendations made by the UK Information Commissioners Office (ICO) in their Code of Practice?) Table 8 summarises from Appendix 2 how many of the forty ICO Code recommendations each SNS addressed at least once individually. Table 9 and Figure 7 show how many recommendations were addressed by all six SNS, through to how many were addressed by none. They show that when looking at recommendation coverage, the largest percentages of recommendations covered, were for those covered by **none** (22.5%), or **all six** of the SNS (30%). These percentages account for over half of the total recommendations and show similarity between SNS in terms of the ICO Code recommendations that they do (and do not) address. Unlike Figure 6, there is no evidence of a power-law relationship in Figure 7 between the number of SNS and how many recommendations they address. Instead, the majority of recommendations were either addressed by either all SNS, or none.

Table 8 No. of ICO Code Recommendations Addressed At least Once by SNS

SNS	**No. of ICO Code Recommendations SNS Addressed at Least Once**	**% of ICO Code Recommendations at Least Once**
Facebook	**23**	**57.5%**
Pinterest	**18**	**45%**
Twitter	**19**	**47.5%**
Flickr	**19**	**47.5%**
LinkedIn	**26**	**65%**
Google	**26**	**65%**

Table 9 ICO Recommendations and No. of SNS They Were Addressed By

No. of SNS That Addressed Code Recommendation At Least Once	How Many Code Recommendations Addressed by All Six SNS	% of Code Recommendations Addressed by All Six SNS
0	9	22.5%
1	2	5%
2	5	12.5%
3	5	12.5%
4	3	7.5%
5	4	10%

Tables 10 and 11 show Jaccard Similarity and Dissimilarity in covering ICO Code Recommendations. Table 10 shows the similarity of SNS with the Code, ranges from **45-65%**, However, interestingly Table 9 shows that two pairs of SNS addressed exactly the same number of ICO Code recommendations, which could initially indicate similarity. However, they could have addressed different ICO codes with the minimum overlapping which is why Jaccard's coefficient is required to calculate if they are similar in addressing the **same** recommendations/themes.

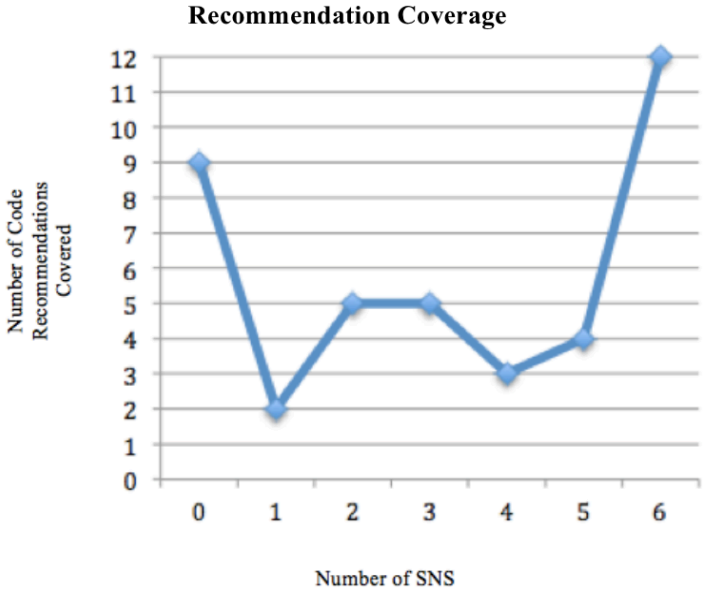


Figure 7 No. of ICO Code Recommendations Covered by Multiple SNS

Table 10 Jaccard Similarity in Covering ICO Code Recommendations

	ICO Code	Facebook	Pinterest	Twitter	Flickr	LinkedIn	Google+
ICO Code		0.58	0.45	0.48	0.48	0.65	0.65
Facebook			0.52	0.68	0.68	0.81	0.63
Pinterest				0.76	0.61	0.52	0.69
Twitter					0.65	0.61	0.66
Flickr						0.66	0.61
LinkedIn							0.68
Google+							

Table 11 Jaccard Dissimilarity in Covering ICO Code Recommendations

	ICO Code	Facebook	Pinterest	Twitter	Flickr	LinkedIn	Google+
ICO Code		0.42	0.55	0.52	0.52	0.35	0.35
Facebook			0.48	0.32	0.32	0.19	0.37
Pinterest				0.24	0.39	0.48	0.31
Twitter					0.35	0.39	0.34
Flickr						0.34	0.39
LinkedIn							0.32
Google+							

The fact that the similarity percentages between the ICO Code Recommendations and the privacy policies of SNS are not closer to 100% similarity may be because the ICO Code recommendations are based on UK and EU law, whereas the SNS are based in the US.

However, in answering RSQ2 'What is the similarity between the privacy policies of the top six SNS globally, in the coverage of forty recommendations, made by the UK Information Commissioners Office (ICO)?', Table 10 shows that similarity, between the SNS themselves in addressing the code recommendations, ranges from **52%-81%**. This evidences a higher percentage of similarity amongst the SNS in the specific recommendations they addressed, than their overall similarity with the ICO Code.

These percentages corroborate Figure 7, that there were certain recommendations that SNS collectively did, or did not, address. However, it is important to note that failing to address an ICO Code Recommendation and a lower similarity with the ICO Code does not automatically necessitate non-compliance with it. Failing to cover ICO Code Recommendation could be because it was not applicable and thus the privacy policy would not need to address it. For example, none of the SNS addressed the recommendation that

'*Where individuals are required by law to provide personal details, be open and explain why information is being collected and what it will be used for'* i.e. none of the policies stated that individuals were required by law to provide certain personal details. This may be because individuals are not required by law to provide SNS with personal details, or equally, because individuals are, but SNS failed to address this in their policy. Which scenario is correct cannot be ascertained without further investigation and legal analysis, outside the scope of this thesis for the reasons that have been explained above. Such analysis would also require access to information, which SNS often do not divulge in full, such as what personal data the SNS collects linked to the exact purpose for processing this.

Table 10 also shows that the least similar (with each other) in addressing ICO Code recommendations were jointly Facebook and Pinterest (52% similarity) and LinkedIn and Pinterest (52% similarity). The most similar were Facebook and LinkedIn (81% similarity). Interestingly, Facebook and LinkedIn had the highest numbers of clauses (Table 3) both prior and after repetitive clauses were removed and although Pinterest did not have the lowest number of clauses, it did have the second lowest with only 21 more than Flickr, who had 66 (Table 3). This indicates that the more clauses SNS have, the more ICO Code Recommendations they are likely to share, another example of a power law relationship. However, as stated above, failing to address recommendations is not indicative of non-compliance, and therefore a lesser length or lower similarity with the ICO Code should not be assumed to mean a less legally compliant policy.

3.4.4 ICO Code Recommendations Not Present in Privacy Policies of SNS (RSQ3)

In answering RSQ3 (Are there any recommendations of the ICO Code, which all privacy policies do not address?) analysis identified nine of the forty ICO Code Recommendations that were not addressed by any of the six SNS. These were as follows:

Try and predict whether you will be likely to do things with it (the personal data) in future without drawing up a long list of future possible uses if you are unlikely to use it for those purposes. LinkedIn and Facebook were the only SNS to touch on the future of their services in their privacy policies, although not enough to count as addressing this ICO Code Recommendations. Google+ did state that they would ask for *'consent before using information for a purpose other than in this privacy policy'*, however, this is not specific about future uses and does not indicate what these purposes may be to address the

Recommendations. Facebook stated that: '*Granting us permission to use your information not only allows us to provide Facebook as it exists today, but it also allows us to provide you with innovative features and services we develop in the future that use the information we receive about you in new ways'*. Therefore, although this indicates that Facebook envisages there will be future uses of data, it does not specify what these may be, just that an individual has given permission for them. This is a particularly wide licence, lacking specificity on what these features may be. Indeed, Facebook has a track record of imposing such wide licences, with the Electronic Privacy Information Center (EPIC) filing a complaint with the US FTC about this (EPIC, 2009).

In providing information about the future, LinkedIn was more focused on the possibility that they would collect new types of information, rather than put data they have collected to new uses, stating that '*LinkedIn is a dynamic, innovative environment, which means we are always seeking to improve the services we offer you. We often introduce new features, some of which may result in the collection of new information (for example, when the Endorsements feature launched, we began collecting information about skills for which Members were endorsed and the individuals who endorsed them). Furthermore, new partnerships or corporate acquisitions may result in new features, and we may potentially collect new types of information'.* Again, this was still not specific enough to count as addressing the ICO Code Recommendation as this focuses on collecting new personal data for new uses, not putting the personal data already collected and to use for new purposes.

About the right to complain to the Information Commissioner if there is a problem. None of the SNS mentioned the right to complain to the UK Information Commissioner. This is likely to be because this is the data protection authority for the UK and it could perceived for it to be impractical for the SNS to list the equivalent for every county using its service, especially as a criticism of Privacy Policies in their current form is that they are too long. Furthermore, because of the aforementioned issues relating to legal jurisdiction on the web, users of SNS may not have the right to complain to the UK Information Commissioner or even have an equivalent Regulator in their country to which they could complain.

Interestingly, both Facebook and LinkedIn referred to California's 'Shine the Light law', which is only applicable to California residents. This may be because both are based in California, even though it would certainly not be the only law that is applicable to them.

Have separate notices aimed at different groups of individuals you deal with. Each SNS only provided one privacy policy (although versions in different languages were often available). This may be because it only deals with one 'group' of individuals, as exactly what 'group' means is unclear. The example given in the ICO Code is that a local authority may use information about old age pensioners to administer free access to local leisure facilities or the use of information about shopkeepers to collect business taxes (ICO, 2010). In this sense, this recommendation may not apply to SNS, because they may not have different uses for different user data but use all user data in the same ways. Again, as discussed previously, it is beyond the scope of this thesis to make assessments about the compliance of the SNS with this ICO Code Recommendation.

Where individuals are required by law to provide their personal details be open with people and explain clearly why their information is being collected and what it will be used for. None of the SNS addressed this. As discussed earlier, this may be because there is no information the SNS are required by law to obtain, or that there is, but that the SNS has not included the details of this within their policy.

In marketing contexts, when organisations ask for permission to share customer information with third parties e.g. companies in the same group, this should be backed up with more detail information such as the names of the companies involved for those who want it.

This is certainly applicable to the SNS and all six discussed sharing information with third parties although they did not specify whether this sharing was for marketing. The specific third parties were not detailed in the privacy policies although general representations of types of third parties were sometimes provided. It may that this information was missing from the privacy policies because it was provided (and this recommendation addressed) on another layer of the policy, accessible by following a hyperlink. It may also be that this information is communicated at the time when a marketing permission is obtained from an individual. Therefore, although it can be concluded that it was not addressed in the first layer of the policy, this information may have been available elsewhere. The practice of listing this information elsewhere is questionable, given ICO's guidance on key information being in the first layer and regardless of whether or not the sharing of data with third parties is for marketing purposes, the GDPR (and previously DPD) does require that data controllers provide information about the recipients of personal data under the obligation

to inform. However, it does not specify that it needs to be provided all at once with the other information prescribed under the Obligation to Inform and therefore if SNS are providing it at the point of collecting marketing permissions this could still be viewed as compliant. Again, a full analysis of the compliance of this activity is beyond the scope of this thesis.

If an organisation intends to collect personal information with the intention of selling or renting it, you should make it clear to individuals that the information they provide could be supplied to anyone and used for any purpose and tell them this when they provide their details; and

That if their information is rented, individuals are told that if the business is insolvent, bankrupt, being closed down or sold that their information will be returned to its owner. None of the SNS addressed either of these ICO Code Recommendations, although they all discussed details about the transfer of personal data to other parties if the business becomes insolvent. This may mean these recommendations do not apply because none of the SNS sell or rent information. However, with companies like GNIP (2016) selling access to 'social media data' from various SNS, specifically listing that access to Twitter, Facebook, Flickr and Google+ data is available, it is unlikely that these ICO Code Recommendations would not be relevant. Thus, for these four SNS at least, this ICO Code recommendations would appear to apply.

Avoid using confusing terminology e.g. technical language. Pinterest, LinkedIn and Google all stated at the beginning of their policy that they had tried to keep their policies as simple as possible. Pinterest acknowledged that some of their terms were a little technical, and Google advised that if readers were not familiar with terms like cookies etc. they should read about them first. However, all policies used technical language, such as 'cookies', 'API' and 'plugins' etc., and therefore arguably failed in addressing this recommendation. The recommendation is to 'avoid' rather than 'try to avoid', which is more akin to what the policies did. Some technical language was used and then followed by an explanation of what it meant. For example, Twitter, Facebook and Pinterest all provided a definition of 'cookies. However, the term API was often used without definition. It must be noted that complying with this recommendation would prove almost impossible for SNS, and websites in general. As Robinson, et al., (2009) acknowledges, national laws require full descriptions of data processing activities, which prove difficult to describe in a form the consumer can

understand without the use of technical language. A better recommendation may be to avoid using technical terms but where this is not possible, to always provide a description/explanation to accompany technical language.

If you collect information from vulnerable individuals (such as children) have an appropriate privacy notice to their level of understanding – would they understand the consequences? None of the SNS offered policies aimed at vulnerable individuals, which could be because they do not collect information from them. Pinterest, Twitter and LinkedIn all stated that they had a minimum age to use their site, whereas Google and Flickr did not mention a minimum age in their policy, which may be because they have a generic policy, applicable to many services where some services are be available to children and some may not, however if any of their services are there should be a mention of children in here. Upon brief further investigation (which will not change the results or scope of this study but was merely undertaken to ascertain whether SNS should have been providing notices specific to children), elsewhere on its site, Google+ mentions a minimum age of 13 to use its services and in Yahoo's Terms of Service which are a separate legal document to its privacy policy, it states a minimum age of 13 to use the SNS provided by Yahoo.

Despite having a specific policy for its service, Facebook also stated that it had a minimum age of 13 in its Terms of Service rather than its privacy policy. The common factor of the minimum age of 13 may be because of the US FTC's Children's Online Privacy Protection Act (1998), which applies to the online collection of personal information from children under 13, and places additional requirements upon websites which do. Stating a minimum age of 13 is a bold statement that the website is not aimed at children, which the FTC will consider during investigations. This indicates, that regarding children, SNS may not be required to provide notices to their level of understanding. However, in some countries children may be considered to be older than 13 and there may also be other vulnerable individuals that require an appropriate privacy policy, especially given the amount of information, which an SNS can obtain. Other vulnerable individuals could include those with mental health or learning difficulties over the age of 13. As all of the SNS only had one privacy policy, despite being used by vulnerable individuals, they are certainly not in compliance with this recommendation.

3.4.5 Themes Addressed by All Privacy Policies of SNS But Not Recommendations in the ICO Code

As mentioned in Section 3.3.5, if a clause was not allocated to an ICO Code Recommendation, then it was placed into a category of 'miscellaneous'. To answer RSQ4 (Are there any themes of information addressed in all of the privacy policies that were not included in the forty recommendations from the ICO Code?) inductive Thematic Analysis was performed on these 'miscellaneous' clauses. This identified four themes, which appeared in all six privacy policies, but not in the ICO Code.

The Process of Updating the Privacy Policy: Whilst the ICO Code mentions that privacy policies should be reviewed regularly, it does not advise that websites should include details about the process of doing so or involve users in this process. However, all SNS did include information about their process of revising the policy, with some mentioning more than others. In particular, Facebook was very detailed about the process of revising the privacy policy and stated that:

> *'If we make changes to this Data Use Policy we will notify you (for example, by publication here and on the Facebook Site Governance Page). If the changes are material, we will provide you additional, prominent notice as appropriate under the circumstances. You can make sure that you receive notice directly by liking the Facebook Site Governance Page…Unless we make a change for legal or administrative reasons, or to correct an inaccurate statement, we will give you seven (7) days to provide us with comments on the change. After the comment period, if we adopt any changes, we will provide notice (for example, on the Facebook Site Governance Page or in this policy) of the effective date'.*
>
> **Facebook (2014)**

The detail Facebook provided might be linked with the fact that they had the largest number of clauses, as trying limit length did not appear to be a concern. It could also be due to the criticism that they have received regarding previous revisions of their privacy policy (Anderson, 2009). Either way, this theme of detailing the process of updating the policy was common to all six SNS.

Functionality: As mentioned in relation to RSQ1, a large number of clauses contained in the SNS were used to explain the functionality the SNS utilised, which was often required to explain the purpose for using or sharing personal data. For example, both Facebook and

LinkedIn had to explain Platform technology (an underlying system on which application programs can run) in order to explain who they share personal data with through it and the purpose of this. Technically, the SNS could fulfil the ICO Code's recommendations without explaining their functionality in detail. However, explaining these functionalities and their role in the collection, use and sharing of personal data can provide the contextual information that makes privacy policies more informative and enables users make informed decisions regarding the processing of their personal data, thus increasing the transparency of personal data processing.

List of Personal Data Collected: Interestingly, the ICO Code did not recommend that organisations should provide information on exactly **what** personal data they collect, despite requiring that they state the purpose of obtaining, using and disclosing it. However, providing information about the specific personal data the SNS processed was clearly present in some form in all of the privacy policies of the SNS. For example, SNS stated that they processed various personal data including photos, associated metadata, messages, responses to ads etc. A recommendation to include this may not have been included in the ICO Code because it is not solely aimed at SNS, or even websites, but governs data collection both online and offline. In the offline context, using the example of filling in a questionnaire, an individual may be fully aware of what personal data they are providing to the organisation as they write it down. However, online users may not be so aware of *what* personal data is being collected about them, especially as discussed previously, '*increasing amounts of data are not collected from the individuals concerned, but are instead observed, derived and inferred*' about individuals (OECD, 2014). Because of this, it is impossible for SNS or any type of website to be truly transparent about personal data processing without providing information about the specific personal data it is processing in some form.

How They Receive Information: The final theme that was present in all of the privacy policies but not a recommendation in the ICO Code is to provide information about *how* the personal data of individuals is received by the SNS e.g. through friends, through a user's computer etc. Again, the reason that this is not a recommendation of the ICO Code may be for the same reason above, that the ICO Code is not aimed specifically at online contexts, where the sources of information are not as transparent and users may not be as aware of *how* information is being collected about them. However, providing this information clearly

makes data processing more transparent as users may not always be aware of how information is collected (Rogers, 2011:223). Although one of the forty ICO Code recommendations stated that 'organisations that intend to combine information from different sources should explain this', this does not recommend that SNS should always provide individuals with this information, regardless of whether they intend to combine information. Even where SNS do look to combine information from other sources, this recommendation doesn't explicitly require SNS to detail what those sources are and SNS could interpret this as just a requirement to explain how data will be combined opposed to where it came from.

3.5 Discussion, Recommendations and Limitations

In the words of Aristotle, 'the whole is greater than the sum of its parts' and it is only by combining the answers to RSQ's 1-4, that an understanding of their implications regarding the 'bigger picture' of privacy policies can be understood and that the answer to RSQ5 *'To what extent is standardization possible between the privacy policies of SNS?'* possible.

3.5.1 Discussion

The answer to RSQ1 indicated that there is not a lot of similarity between the privacy policies of SNS in the clauses that they use, with similarity ranging between **8-27%.** Although there were common clauses that could be drawn out, a power-law relationship existed between the number of common clauses and the number of SNS which shared them. This meant that there were only a small percentage of clauses shared by all six SNS, with the majority of clauses bespoke to only one SNS. From looking at this alone, one would conclude that the similarity between SNS is so low that standardization seems, if not impossible, definitely a long way off. Indeed, a lot of work would be required to create the level of similarity required for standardization by clause.

However, the answer to RSQ2 showed that if you look at similarity in terms of themes of information addressed rather than the specific clauses used to do so, the similarity between SNS is far higher, ranging from **52-81%**. This indicates that SNS express similar themes of information in their privacy policies, but in different ways. This indicates that standardization is not as impossible as the answer to RSQ1 suggests. Indeed, it was noted that in relation to re RSQ1 in Section 3.4.2 that the differences in clause coverage were

largely due to differences in functionality, semantics and amounts of elaboration between SNS. By looking at similarity in terms of theme, these differences are not as influential, particularly in relation to the semantics of the language used and the amount of elaboration the SNS provide. For example, if one SNS used ten clauses to address an ICO Code Recommendation and another SNS used two, and with differing language, they have both still addressed the same theme and therefore would be considered similar.

The answers to RSQ'S 2 and 3 also showed that there were ICO Code Recommendations which every SNS addressed and recommendations that were addressed by none of the SNS. However, it highlighted that it cannot be assumed that a failure to address a recommendation is due to a SNS failing or choosing not to do so, as the particular recommendation may not be relevant to them. Furthermore, some recommendations were almost impossible to comply with in the context of SNS, such as the recommendation to 'avoid using confusing terminology'. This is because of the technical functionality of SNS, which they rely on to process personal data. To properly convey how information will be used or disclosed, a level of technical language will always be required.

Answering RSQ4 showed a number of themes present in all of the policies, which were not recommendations in the ICO Code. However, these clearly provided additional information to users regarding use of their personal data, which would give them more control over their personal data. As discussed previously, one reason for this may be that the ICO Code was not aimed exclusively at online environments and therefore, makes assumptions about people's awareness of the information they provide. Furthermore, the age of the ICO Code may also attribute to this. Dated December 2010, the ICO Code was, at the point of the study, under review and in the process of being updated. ICO may also have assumed that information within these themes would have been caught by their 'catch-all' theme *'Any further information necessary, in the specific circumstances, to enable the processing in respect of the individual to be fair'* which was removed from forty themes we used for being too broad and non-specifics.

So, whilst RSQ3 shows that the ICO Code provides a number of themes, which are present in the policies, in practice it proved that it cannot be treated as an exhaustive list of what should be included in a global standardized policy when all stakeholders are considered. Indeed, as one of the criticisms of privacy policies in practice is that they are targeted at

meeting applicable legal requirements rather than serving a real transparency benefit towards the consumer, it appears that there are elements of transparency which go beyond the legal requirements in the UK. Thus, when looking to standardize policies of SNS or other types of website, a thorough examination should also include an assessment of a representative sample the relevant privacy policies that will be subject to standardization, to ascertain a full list of themes, in addition to examining other sources.

3.5.2 Recommendations

Given this discussion, the outcome of this investigation indicated that in answer to RSQ5 *'To what extent is standardization possible between the privacy policies of SNS?* Is that standardization is possible, albeit at a thematic level currently opposed to at a clause level. It is also only possible if the issues raised throughout this study are addressed. Thus, five initial recommendations were made to facilitate the development of a standardized privacy policy for SNS:

1.Begin with an as-exhaustive-as-possible list of themes, which a SNS should address, rather than focusing on clauses initially. Because the investigation showed high similarity between the policies in the ICO Code recommendations they covered, SNS policies are already in a better position to begin to be standardized by theme. This could form a visually familiar table for users as a first step, consisting of two columns, with the list of standardized themes in a standardized order on the left. The SNS clauses can then be allocated to those themes on the right. In addition to looking at the legal requirements and the advice of data protection authorities to create this list of themes, a representative sample of the privacy policies which will be utilising this standardised policy should also be examined as a source from which themes can be gleaned.

2. Include a theme of general functionality and a separate theme of functionality specific to that SNS. A theme of general functionality would include functionality common to all SNS and the data collection and use associated with this (such as log data etc.). specific functionality would include functionality that the specific SNS offers and uses above that minimum processing required, which will result in further personal data processing. If this is provided in this format users could easily identify differences between SNS by looking at the specific functionality theme, in addition to familiarizing themselves with standard processing

for this type of technology in the general functionality theme.

3. Definitions, explanations and examples of technical terms should be standardized so that each privacy policy uses the same ones. For example, when referring to 'cookies' there should be a set, approved definition, explanation and example of a cookie. Given that it is almost impossible to avoid using technical terms in relation to describing the activities of SNS, at least by doing this, the amount or type of information a user gets in this context will not vary with the SNS they use. This will lessen confusion and potentially support familiarity with definitions and examples.

4. Certain words should also be standardized. The same benefits of standardization can be realised for certain words. For example, close, delete and deactivate should not be used interchangeably, but either one word is used, or their individual (but separate) definitions in relation to terminating an account should be standardized i.e. 'close account' always means one thing, as does 'delete account'. This would also lessen confusion and increase transparency.

5. Make sure that when standardizing, there is a way for users to easily ascertain when a theme is not addressed and why. As mentioned, if an ICO Code Recommendation was not addressed it was unclear whether this was because it was not applicable, or because the SNS simply failed to do so. Fulfilling this recommendation would solve this issue by quickly making it clear when a theme of information would not be covered and why. This would make SNS justifications for not addressing a theme clear to users, regulators and researchers and would evidence that they have considered all the applicable requirements when constructing a privacy policy.

3.5.3 Limitations

Prior to concluding on this particular study, it is important to discuss its limitations. One limitation is that this investigation only looks at the policies at a single point in time (August 2014), and whilst other comparative analyses have been discussed (Section 3.2.3), these did not compare the policies for the same elements as this investigation. This makes it difficult to provide information on the evolution of privacy policies at different points in time, especially given that policies are often frequently revised. However, this investigation has

provided the starting point for future work, which could repeat this investigation on the policies as they change, to allow for a discussion on the evolution of privacy policies.

Another limitation of the study is that the coding and analysis in Stages 1-5 was completed by one researcher and then discussed with the other researchers. Whilst this provided methodological consistency, coding in parallel could have enriched the work by providing multiple perspectives. Therefore, future work could conduct inter-rater reliability or intra-rater reliability on the findings.

Furthermore, as discussed in Section 3.3.2, only the first layer of the privacy policies was examined for various pragmatic reasons. This could have impacted the findings, because the nine ICO Code Recommendations that we did not find addressed in the 'first layer' of the policies could have been found in these further layers. Future work could extend this analysis to all layers of the policies on the same domain, to conclude whether certain elements have been addressed at all.

This analysis could also be complimented by a full legal analysis of what the implications of these exceptions are, whether the SNS are in contravention with the ICO Code, and whether it is appropriate that this information is not in the first layer of the policies.

3.6 Conclusion

The preliminary study discussed in this chapter had two aims. First, to understand whether privacy policies are similar enough in practice for standardisation of them to be possible, and second, to understand more about the phenomena that are 'privacy policies', with a focus on the information that they contain.

In relation to the first aim, the study found that the privacy policies of SNS demonstrated homogeneity and promising potential for standardization, albeit at a thematic, rather than clause level. Five recommendations were then made to support achieving this in practice. In relation to the second aim, the study found that there was a number of ICO Code Recommendations for transparency which were not addressed by the privacy policies. It also found that there were a number of themes of information that all of the privacy policies included, but that went beyond the recommendations of the ICO Code.

Given the overall research goal of this thesis, there were many ways the findings from this study could have been taken forward. In relation to the first aim, the recommendations made in Section 3.5.2 could have been followed up with further work on how to put these into action, in order to standardize privacy policies on a thematic level. For example, computer supported algorithms could be used to test whether this standardization can take place in practice, and whether it can be done by computers instead of manually by humans. Following this level of standardization, further levels could be explored, such as standardizing some of the clauses, or the specific information to be provided within themes. For example, the ICO Code Recommendation of detailing *'Who to contact if they want to complain or know about how their information will be used',* could be standardized so that organizations have to provide the same specific information, such as: telephone number; physical address; email address etc. In addition, although this study indicated that standardization by clause is not currently feasible, overcoming the differences in functionality, semantics and elaboration with the recommendations made in Section 3.5.2 would allow for another assessment of the potential for standardization by clause.

In relation to the second aim, the results from the study highlighted something of interest and in particular questioned two implicit assumptions that had been made going into it. First, was that ICO's recommendations for a compliant privacy policy would provide a full list of the information required to make data processing transparent; and second, that that the problem with increasing the transparency of processing lies simply in improving the communication of the information that privacy policies contain.

In relation to the first, from this study alone it is not clear whether this assumption is correct, or whether the guidance from ICO reflects the position of the law. One of the most surprising findings was that all six of the policies touched on the specific personal data that they collect and process, even though the ICO Code did not recommend that they do so. This finding seemed unusual, given that the ICO Code aims to set the highest bar for compliance with the law and for transparent processing, yet the policies were providing information beyond its recommendations. It also seemed unusual given that privacy notices often form part of the basis of consent, and well-known models of informed consent online (e.g. Friedman, Lin and Miller, 2005) require an explicit answer to the question of 'what information will be collected'.

In relation to the second assumption, it is one upon which much of the transparency work in HCI is based. Taking the Usable Privacy Research Project (2016) as an example, the work is based on the policies as they already exist, even work on the creation and analysis of a website privacy corpus (Wilson, et. al., 2016) did not include a reference to the specific personal data being processed.

Whereas, the results of this study highlight that there is also a need to understand whether the information organisations are providing in their policies is sufficient enough to enable transparency in the first place. Improved communication of information will not make data processing as transparent as it can be if the information provided is not enough. This becomes even more important where computing is being used to process this information behalf of the individual. This is not necessarily a fault of HCI research, as it may work on the assumption that it is the role of other disciplines, such as law, to investigate this. However, it does show that without this additional research, work within the HCI community on improving the transparency of processing will reach (if it has not already) a glass ceiling in the ability to make personal data processing transparent.

Therefore, whilst the findings in relation to either aim of this preliminary study could be pursued further, if the problems raised by second aim were left unresolved, they could stand in the way of the creation of a transparent, standardized, and legally compliant privacy policy. To research all of the problems raised would be a gargantuan task, far beyond the scope of a single PhD and therefore the most surprising one was chosen. The work undertaken in the next chapter was devised to continue to meet the research goal of the thesis. In particular, it focuses on whether there is a legal requirement in the EU and UK, for organisations to provide information about the specific personal data that they are processing.

Chapter 4 Categorising Personal Data

4.1 Introduction

The previous chapter discussed the preliminary study that was undertaken as part of this thesis. The findings highlighted that when looking at the similarity between the privacy policies of SNS, all six policies included information about the specific personal data that they processed, yet this was not a recommendation within the ICO Code of Practice for privacy policies.

This finding was unusual, because it seemed logical that for the processing of personal data to be transparent, organisations (especially those providing online services) would need to provide information about the specific personal data that they process. The study also highlighted and challenged two assumptions that were made going into the preliminary study. First, that ICO's recommendations for a compliant privacy policy would provide a full list of all the informational requirements required to make data processing transparent; and second, that the problem with making data processing more transparent lies simply in improving the communication of the information contained within privacy policies.

This chapter discusses the second investigation undertaken as part of this thesis, the aim of which was to investigate the assumptions the preliminary study highlighted in more detail. The investigation was published in a Journal paper in the Computer law and security review (Cradock, Millard and Stalla-Bourdillon, 2017). The overall research question for this study was:

RQ2: When is there a legal requirement in the EU and UK under the obligation to Inform to provide information about the specific personal data being processed and what is the requirement for this?

This research question was devised to investigate whether there is a legal requirement in the EU and the UK to provide information about the specific personal data being processed and if so, what this requirement is. It also sought to then understand whether not recommending this was a deficiency in the ICO Code.

4.1.1 Methodology

Legal doctrinal research method (Van Hoecke, 2011), critical analysis (Vibhute & Filipos Aynalem, 2009) and argumentation theory (Van Eemeren, et al. 2013) were used to answer these questions, describe the state of affairs in EU and UK law. Legal doctrinal research method is the law's equivalent of scientific enquiry. It collects empirical data (statuses, cases etc.), words hypotheses on their meaning and scope, and then tests them using the classic canons of interpretation. Theories are then built on this that are then tested and from which new hypotheses are derived (Van Hoecke, 2011).

4.2 Categories of personal data

At the time of the investigation, the Data Protection Directive was the primary law governing data protection in the EU. As discussed previously, this has now been replaced by the General Data Protection Regulation. This section discusses the previous position under the Data Protection Directive and the UK Data Protection Act 1998 for context and then discusses the current position under the GDPR and the UK Data Protection Act 2018. It also looks at guidance provided in relation to these laws by the UK regulator and other data protection bodies.

4.2.1 Data Protection Directive

As discussed in Chapter 2, the Data Protection Directive (DPD) created an 'obligation to inform' in Articles 10 and 11, which required data controllers (as the organisations which determine the purposes and means of the processing) to provide certain information to data subjects (the individuals to whom the personal data being processed relates) about the processing of their personal data.

Under the DPD, the obligation was split into two processing scenarios, each of which had slightly different informational requirements. Article 10 governed cases where personal data was collected from the individual, and Article 11 governed cases where the personal data had been obtained from elsewhere.

In providing the informational requirements, both Articles stated that data subjects should be provided with at least:

- The identity of the controller and his representative, if any;
- The purposes of the processing, for which the data are intended; and
- Any further information necessary, having regard to the specific circumstances, to guarantee fair processing in respect of the data subject

Whist the first two points made it relatively clear what information must be provided by organisations, the last point was a wide and case-specific requirement. To provide further clarity, both Article 10 and 11 each gave three examples of information that might fall within this third point, and could be necessary to inform data subjects of, depending on the circumstances. Both Articles provided the examples of informing data subjects of:

- The recipients or categories of recipients of the personal data; and
- The existence of the right of access to and the right to rectify the personal data concerning them;

However, the Articles differed on the third example they each provided. Article 10(c) provided the example of informing the individual of *'whether replies to the questions are obligatory or voluntary, as well as the possible consequences of failure to reply'*; whereas, Article 11(c) included the example that when information is obtained 'not from the data subject', they may need to be informed of the *'the categories of data concerned'*. Whilst it was easy to see why Article 10's example was not included in Article 11 (because it would only apply to a scenario where the individual is providing the information and could not apply if the data being obtained from elsewhere) it was not as clear why Article 11's example was not listed in Article 10.

Because informing data subjects of the 'categories of data concerned' was not stated anywhere within the text of Article 10, one could conclude that this means that whenever personal data is obtained from the individual, there is no requirement to inform them of the personal data being processed. However, despite not being listed as an example, in theory, providing this information could have still been required under Article 10(c), as the examples provided were not exhaustive. Specifically, Article 10(c) requires the data controller to inform the individual of '**any** information required' for the processing to be

'fair' and so in certain circumstances a data controller may have been obligated to inform individuals of the 'categories of data' being processed. Yet, even if this were found to be so, this requirement would have been assessed on a case-by-case basis, as confirmed by the European Commission in their first report on the implementation of the DPD (European Commission, 2003). This means that even if a subsequent legal case had clarified a scenario which required the provision of this information under Article 10, a data controller would only be under an obligation to provide this in circumstances similar to those in the ruling.

Furthermore, even though informing individuals of the 'categories of data concerned' was mentioned within Article 11, this was only as an example of 'any further information necessary' for the processing to be fair. This means that it was not mandatory for this information to be provided to individuals in every case where personal data about them was obtained from elsewhere, only when it was necessary to guarantee 'fair' processing. Compared to Article 10, this at least indicated that the DPD envisioned there would be situations in which data controllers will be obligated to provide individuals with information about the 'categories of data concerned', however it did little to clarify what these circumstances will be.

Further uncertainty arose under the DPD when considering whether it is Article 10 or Article 11 that applied. The obligation to inform distinguished between situations where the data is 'collected from the data subject' (Article 10) and where the data is 'obtained not from the data subject' (Article 11). However, looking at the differences between 'provided', 'observed', 'inferred' and 'derived' personal data in relation to the obligation to inform as discussed in Chapter 2, it was unclear which situation applied where. It seems clear that 'provided' data would fall under Article 10, because it is provided with the awareness of the data subject (and therefore certainly obtained from them (OECD, 2004)). Yet, for personal data that is 'observed' by others and recorded in a digital format (OECD, 2004) e.g. data originating from online cookies or sensors, how would this be classed? It could be argued that the individual is the source of the data, as their actions generate the data in some way and therefore Article 10 would have been applicable. However, it could also be argued that the cookie or the sensor is the source of the data and therefore Article 11 should apply.

This uncertainty is also true for personal data that is 'inferred' or 'derived', where in both cases data is generated from other data. It is unclear whether the obligation to inform only covered the instance of the original personal data collection or whether as the data controller began to derive and infer personal data from this original data (e.g. the profitability of the individual) that they fell under an Article 11 obligation, as this new personal data that is being created had not strictly been obtained from the individual. This could be an instance where a data subject should have been informed of the 'categories' of personal data that would derived or inferred under Article 11.

Despite this uncertainty, it is worth noting that the distinction between the two Articles is somewhat arbitrary due to the catch all requirement that they both include which requires data controllers to inform individuals of 'Any further information necessary, having regard to the specific circumstances, to guarantee fair processing in respect of the data subject'. This wide requirement could arguably see the provision of the same information under both Articles. However, the inclusion of the example of 'categories of data' in Article 11 does imply that the legislators believed it was more likely to be required in scenarios where personal data is not obtained from the data subject.

Thus, under the Data Protection Directive it was not always clear which Article the data controller's processing activities were governed by. Furthermore, whilst certain circumstances may have mandated the provision of information about the specific personal data being processed, which would be done by listing the categories of personal data, data controllers were not required to provide this information consistently to individuals.

4.2.2 Article 29 Working Party Guidance

Although from the legislation it was unclear when, and whether, data controllers needed to inform individuals of the 'categories of data concerned', the Article 29 Working Party (WP) repeatedly referred to a need for data controllers to inform individuals of the personal data they process under the obligation to inform. However, they did not always refer to this obligation using the term 'categories'.

The Article 29 Working Party was an independent body that gave expert advice on data protection within the EU under the DPD,

Even as early as 1999, the WP were concerned about processing operations that were being performed without an individual's knowledge, stating that:

> *"Internet software and hardware products should provide the Internet users information about the data that they intend to collect, store or transmit"*
> **Article 29 Working Party (1999)**

The Working Party echoed this guidance again in relation to online data protection in 2000 (Article 29 Working Party, 2000). The WP have also stated that Individuals should be given:

> *"accurate and full information of all relevant issues, in particular those specified in Article 10 and 11 of the Directive, such as the nature of the data processed"'*
> **Article 29 Working Party (1999)**

And that:

> *"According to Article 10 ... each data subject has a right to know ... in the context of apps ... what type of personal data is being processed'* and that *'the relevant data controller must inform potential users at the minimum about: ... the precise categories of personal data the app developer will collect and process"*
> **Article 29 Working Party (1999)**

Although this quote was in the context of apps and smart devices, and the one before in the context of electronic health records, for both, the WP based their opinion on Article 10 DPD, despite the lack of this specific requirement within this Article. Even if the WP were basing this guidance on Article 10(c) DPD, then this would not be an information requirement for data controllers in every case of processing, as it could only be interpreted as applying in the specific circumstances referred to by the guidance i.e. apps, smart devices and electronic health records (as discussed in Section 4.2.1).

Interestingly, in 2014 the WP did extend their guidance beyond these specific scenarios, when they advised Google that to overcome issues with its one-for-all privacy policy, it should provide *'an exhaustive list of the* ***types of personal data*** *processed'* **(**Article 29 Working Party, 2014**).** This confirmed that, in the opinion of the WP, there were situations

beyond apps and smart devices, or electronic health records, in which individuals should be informed exhaustively of the types of personal data being processed about them.

Whilst this guidance from the WP seemed promising in providing some clarity to the RSQ2 of whether data controllers are required to provide information about the specific personal that they are processing, for a number of reasons, the guidance of the WP did not provide clarity for data controllers on this.

First, because the WP has discussed this requirement of data controllers in relation to specific scenarios (such as apps, search engines and smart devices) it is unclear whether it is only in relation to these facts that it applies or whether the WP believe that regardless of context, data controllers should be providing individuals with this information. Interestingly, in relation to apps and smart devices, the WP reasoned that the obligation is required because:

> *'Being told what data are being processed is particularly important given the broad access apps generally have to sensors and data structures on the device, where such access in many cases is not intuitively obvious"*
> **Article 29 Working Party (1999)**

This justification, and the problem of unobvious and broad access, is also true of other online contexts, especially due to the increase in observed, derived and inferred data (OECD, 2014). Therefore, it would seem logical that to create transparency, this obligation should be extended to all online personal data processing, and at least, to any other scenarios where this reasoning applies. However, without clarification on this matter, the extent of this obligation remained unclear.

Second, instead of consistently referring to this as a requirement to inform an individual of the 'categories' of personal data processed, in accordance with the wording in Article 11 DPD, the WP used various inconsistent terms in its guidance. It referred to this requirement simultaneously, as a requirement of data controllers to inform data subjects of the 'nature of the data', the 'types of data', and of the 'categories of personal data'. Using such differing terminology to refer to this requirement without clarifying what these terms mean (and

whether they are equivalent), makes it unclear exactly what obligation the WP thinks data controllers are under.

Thirdly, although their guidance was authoritative, and highly influential, the WP only held an advisory status, and therefore its opinions and recommendations (including these) were not legally binding. This meant that if a data controller did not inform individuals of the categories of personal data it processed, it would still have been for a court or the regulator to confirm that they were not fulfilling their data protection obligations. Without that confirmation, and clarification of the circumstances in which it applies, it is unclear what the outcome would be and what obligation data controllers were under.

4.2.3 The UK Data Protection Act 1998

As discussed in Chapter 2, the nature of the DPD as a Directive meant that it had to be implemented into each EU Member State's (MS) national law, and so examining these implementations could provide some further clarity on RSQ1. However, taking the example of the United Kingdom (UK), it was just as unclear as to when a data controller was under an obligation to inform the data subject of the categories of personal data that they were processing about them.

The UK implemented the DPD through the Data Protection Act 1998 (DPA 1998) and the Article 10 and 11 DPD information requirements were transposed (almost verbatim) into Schedule 1, Part II 2(3) of the DPA 1998. Interestingly, informing data subjects of 'the categories of data concerned' was not stated anywhere within the DPA's informational requirements.

Unlike the DPD, the informational requirements were only referred to once, in Schedule 1, Part II 2(3) DPA 1998. This removed the differing examples of 'any further information, which is necessary' which were contained in Article 10(c) and 11(c) DPD. This is an important difference, as it was these examples that suggested that 'any further information necessary' might differ depending on whether data is obtained from the data subject or from elsewhere under the DPD.

As discussed in Section 4.2.1, it could be argued that the distinction between the two Articles in the DPD was arbitrary, however in removing the differing examples the DPA 1998 also removed the reference to providing the 'categories of data concerned'. It was the provision of this example which indicated that informing data subjects of the 'categories of data concerned' might be a specific requirement under the obligation to inform at all.

Under the DPA 1998, the only difference between obtaining data from the individual and 'any other case' appears to be between the time of disclosure of the information, under Schedule 1 Part II, 2(1) and 2(2). This provided that if data is obtained from the data subject, the information must be provided at the time the data controller first processes the data and that in any other case of acquiring the personal data, the information must ideally be provided at the time of first processing, if not as soon as practicable after.

Thus, the DPA 1998 provided even less clarity on when a data controller might be under an obligation to inform the data subject of the 'categories of data concerned' than the DPD.

4.2.4 The General Data Protection Regulation

Whilst confusing, in some ways it can be seen as quite logical that the WP might be inferring a requirement that was not stated explicitly within the DPD. Indeed, it has been asserted that the 1980 OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (OECD, 2013) upon which the DPD is based, were '*developed primarily with 'provided data' in mind*' (OECD, 2014). It is therefore not surprising that the DPD also reflects the presumption that data is collected from individuals with some degree of involvement or awareness.

Following this presumption, the logical consequence is that there is no need to inform individuals of exactly what personal data is being collected, as individuals would be involved or aware of the data being collected which is why there is a focus on informing individuals of the purpose it will be put too. The drafters of the DPD could not have foreseen the explosion in personal digital technology that would follow the creation of the DPD, and dramatically change the personal data collection and generation practices of data controllers', which in practice rebut this presumption. Thus, the WP may have had no choice but to try to bridge the gap between the focus of the DPD on 'provided data' and

the reality of data collection as it became, where 'provided data' was just the tip of the iceberg in terms of what data controllers were collecting and processing.

Given this, and the fact that the GDPR has been heralded as the modernisation of the legal framework for data protection law within the EU, if the WP's guidance on this matter were authoritative, the logical conclusion would be that 'categories of data concerned' would be listed as a mandatory information requirement under the equivalents of both Article 10 and Article 11 DPD in the GDPR (Articles 13 and 14 GDPR respectively). At the very least, one would expect to see 'categories' of personal data as an example of something that an individual may need to be informed of for the processing to be fair under both Articles.

However, Article 13 GDPR (which replaced Article 10 DPD) still does not mention informing data subjects of the categories of personal data at any point. The GDPR did introduce new mandatory information requirements under Article 13(1), and also new examples of what the data controller *might* need to inform data subjects of for processing to be fair and transparent under Article 13(2). Yet, despite this, the GDPR still does list the categories of personal data as something data subjects may need to be informed of when personal data is obtained from them. This seems completely at odds with the WP's guidance, which has repeatedly referred to Article 10 DPD when inferring this requirement.

Interestingly, under Article 14 GDPR (which replaced Article 11 DPD), informing individuals of the categories of personal data is no longer merely an example of further information that *'might'* be necessary to ensure fair processing (Article 11(c) DPD). Under the GDPR, it is now a mandatory informational requirement to be given to the data subject in every case of data collection that is not from the data subject (Article 14(1)(d) GDPR).

Thus, on the one hand, the GDPR has increased the importance of data controllers informing data subjects of the categories of personal data. On the other hand, it is still not clear if it is ever an obligation for data controllers to inform individuals of this if they obtain the personal data 'from the data subject', let alone something that is mandatory in every case. Furthermore, the GDPR itself does not contribute any guidance on where the distinction between obtaining personal data from the data subject and from elsewhere lies and so the issues with understanding which Article applies when observed, derived or inferred data is being processed is still unclear.

It could be argued that the reason that the categories of personal data has not been listed in the context of personal data obtained from the data subject, is due to the fact that for some of these scenarios of data collection, individuals may be fully aware of the data they provide. However, Articles 13(4) and 14(5)(a) GDPR allow for this, stating that information is not required to be given if the data subject already has it. Therefore, it would seem more logical to make it a mandatory requirement, and then allow data controllers to rely on Article 13(4) and 14(5)(a) GDPR when necessary, rather than not include it at all. Doing so would reverse the current presumption, that individuals are aware of the data being processed about them, which is more fitting with the guidance of the WP. Furthermore, even where a data subject is aware of the categories of personal data being processed, listing them allows a data controller to attach other information to them to increase the transparency of processing, as discussed in Section 4.3.2.

Interestingly, taking this approach was discussed during the legislative process of the GDPR. The European Parliament Committee on Civil Liberties, Justice and Home Affairs (LIBE) rapporteur's draft report on amendments to the Commission's proposed GDPR (2012), suggested in Amendment 126 to insert *'(aa) category of data processed'* into the (then) Article 14(1) GDPR (now Article 13(1) GDPR). The reasoning was that the GDPR:

> *"can be simplified by merging information and documentation, essentially being two sides of the same coin. This will reduce administrative burdens for data controllers and make it easier for individuals to understand and exercise their rights"*.
>
> **LIBE (2012)**

However, the suggestion did not make its way into the LIBE Committee's Final Report, nor further than this in the legislative process of the GDPR.

4.2.5 The UK Data Protection Act 2018

As discussed in Chapter 2, as a Regulation instead of a Directive, the GDPR replaced the DPD and was directly applicable in Member States when it came into force in 2018, without the need for implementing national legislation. However, as the GDPR allowed for some degree of flexibility via various derogations and exemptions, and so in the UK the Data

Protection Act 2018 (DPA 2018) was introduced to replace the Data Protection Act 1998. The DPA 2018 sits alongside GDPR and tailors how the GDPR applies to the UK, for example by providing the UK-specific exemptions.

With the GDPR being EU Legislation It is worth noting that although the UK voted to leave the EU in June 2016 and formally ceased to be a member on the 31st January 2020, the GDPR still applies until the end of the transition period on 31 December 2020. Following this the GDPR will be brought into UK law as the 'UK GDPR' and will remain in domestic law, but the UK will have the independence to keep the framework under review.

In looking at the obligation to inform in the context of the DPA 2018, again instead of separating the obligation to inform into two sections and making a distinction between the informational requirements when personal data is being obtained from the data subject and when it is being obtained from elsewhere, the DPA 2018 only refers to the requirements once, in Chapter 3, Section 93(1). It also refers to it as the 'Right to information' instead of the obligation to inform.

However, unlike the DPA 1998, where the process of combining the two scenarios into one lead to an omission of the reference to providing information on the categories of data concerned, the combination in the DPA 2018 has led to the *'the categories of personal data relating to the data subject that are being processed'* being an informational requirement that must be provided in all scenarios under Section 93(1)(c).

Presumably this is because the GDPR made it a requirement in all cases of processing of personal data obtained from elsewhere (Article 14) to provide information about the categories of personal data being processed. Thus, in combining the two scenarios, the UK had to make this a requirement for both scenarios to provide this information otherwise it would be weakening the protection of the GDPR. This means that the DPA 2018 actually goes beyond the requirements of the GDPR and takes the approach that is suggested in Section 4.2.4, by making it a mandatory requirement and then providing an exemption in Section 93(3) that the controller is not required to give a data subject the information if they already have it.

Section 207 DPA 2018 details the territorial application of the DPA 2018 and states that it applies to data controllers and processors that are in the United Kingdom (regardless of where the processing takes place) but also to controllers and processors outside of the UK that are offering goods or services to, or monitoring the behaviour of, data subjects in the UK. This means that in answer to RSQ1, according to the law, data controllers based in the UK or data controllers that are based outside but are offering goods or services to, or monitoring the behaviour of, data subjects in the UK must provide those individuals with information about the specific categories of personal data being processed.

4.2.6 The UK Information Commissioner's Office Guidance

As discussed in Chapter 3, whilst the DPD and DPA 1998 were in force, the ICO's 'Privacy notices code of practice' ("the Code") was the leading authority on complying with the obligation to inform in the UK. A new version of the Code was published in 2016, to reflect the state of the art at the time and the impact of the GDPR. The Code aimed to provide recommendations to support data controllers in drafting legally compliant, clear, and informative privacy notices.

The previous version of the Code (dated December 2010) which was used for the preliminary study in 2014 did not mention informing data subjects of the categories of personal data processed (or the 'types', or 'nature' of the data), even as an example of something that might be required in particular circumstances. Although it stated that when deciding whether to give 'any further information necessary', in the interests of fairness, one must take into account the nature of the data, it did not state that this must be disclosed to the data subject (merely that it must be taken into account).

Yet, the lack of mention of any such requirement in the Code did not prevent ICO from advising Google in 2015 that they have an obligation to inform individuals of the personal data they are processing. Following investigations into its 'one for all' privacy policy, in the undertaking with Google, ICO instructed them to provide:

> *"…clear, unambiguous and comprehensive information regarding data processing, including an exhaustive list of the types of data processed by Google and the purposes for which data is processed"*.

This inconsistency made it even less clear when data controllers are under an obligation to provide data subjects with this information, as their guidance in practice conflicted with their guidance in the Code. Again, this could have been ICO bridging the gap between the hard law and the reality of processing as it had become. Yet, unfortunately, the 2016 version of the Code still did not clarify the situation.

As with the previous version, there was still a strong focus in the 2016 version on the data controller considering what information is collected internally rather than providing information about this externally. The 2016 Code stated that to cover all the elements of fairness, an organisation will need to consider 'what information is being collected?' (ICO, 2016). It also recommended that to help decide what to include in their privacy notices, data controllers should map out how information that flows through their organisation and is processed, including *'what information you hold that constitutes personal data'.*

Unlike the previous version, the 2016 Code did make it clear that a data subject will need to be informed of the categories of personal data processed. However, this only appeared once, at the end of the Code, and only in relation to *'data not obtained directly from the data subject'*, reflecting the new mandatory information requirement under the GDPR (discussed in Section 4.2.4). There was still no discussion of when (if ever) the data subject should be informed of the categories of personal data processed if the data is obtained directly from them.

Although not using the term 'category', the 2016 Code did use the different terminology of 'types' of data and 'the information you collect'. For example, the Code stated that *'depending on the circumstances, you may decide it is beneficial to go beyond the basic requirements of the law'* and tell people the *'the links between different types of data you collect and the purposes that you use each type of data for'.* Furthermore, the 2016 Code provided an example of a privacy notice on a mobile screen, and one of the sections to click on was labelled *'what information do we collect from* you'. Thus, it is clear that the 2016 Code envisaged situations where an individual must be informed of these; however, it is unclear exactly what these situations were. Although linking purposes to types is described

as going beyond the requirements of the law, it is unclear whether 'types' should always be listed and going beyond the law would be the act of linking them to a purpose.

Despite not listing the 'information you collect' or 'categories' or 'types' of personal data as a basic piece of information that a data controller should always include in a privacy notice (ICO, 2016), when discussing taking a layered approach to a notice (which allows a data controller to provide the key privacy information immediately and have more detailed information elsewhere for those that want it), the Code states that:

> *"…there will always be pieces of information that are likely to need to go in the top layer of a notice, such as who you are, what information you are collecting and why you need it"*.
>
> **Information Commissioner's Office (2016)**

This meant that under the 2016 ICO Code it was unclear exactly when a data controller was under an obligation to provide such information. It was also still unclear whether these terms all equate to the same information requirement, or whether the differing terminology reflects different information requirements.

Interestingly, despite the DPA 1998 not making a distinction between the information requirements required when data is obtained directly from the individual or from elsewhere, on the matter of which GDPR Article applies to the processing, the 2016 ICO Code stated that *'there are also some differences in what you are required to provide, depending on whether you are collecting the information directly from data subjects or from a third party'* (ICO, 2016). This suggested that 'data not obtained from the data subject' means data 'collected from a third party'. This could be interpreted as meaning that data collected from a first party cookie provided by the controller would be classed as 'data obtained from the data subject' under the GDPR. If so, this would increase the importance of informing the data subject of the categories of personal data processed under Article 13 GDPR, as it is certainly not intuitively obvious to an individual what personal data is obtained from cookie. However, as it is ICO that have elaborated in this way it remains to be seen whether this reflects general consensus under the framework in other EU Member States which is outside the scope of this thesis. Furthermore, it is still not completely clear

what 'obtained from a third party' entails and clarification with example situations would still prove useful here.

Interestingly, the introduction of a new right under the GDPR provides potential for discussion and clarification on this matter. A new right to data portability for individuals is introduced under Article 20 GDPR. This is a right of data subjects to receive the personal data concerning him or her in a structured, commonly used and machine-readable format, and have the right to transmit those data to another controller. Article 20(1) GDPR provides various qualifications for the right, one of which is that the personal data must be 'provided'. Therefore, the detailed discussion expected on this right may include discussion that clarifies the difference between data obtained from the data subject and from elsewhere or confirms this guidance from ICO.

In particular, the Article 29 Working Party in their 'Guidelines on the Right to Data Portability' (2017) have been endorsed by the European Data Protection Board, the body that replaced the WP. In this guidance the WP state that *"In this regard, WP29 considers that the right to data portability covers data provided knowingly and actively by the data subject as well as the personal data generated by his or her activity".* Again, whilst persuasive rather than binding, this would suggest that personal data collected through cookies and via connected products would be subject to the requirements under Article 13 GDPR and therefore information on the categories of personal data being processed would not always be required, despite the fact that individuals may not be aware of this.

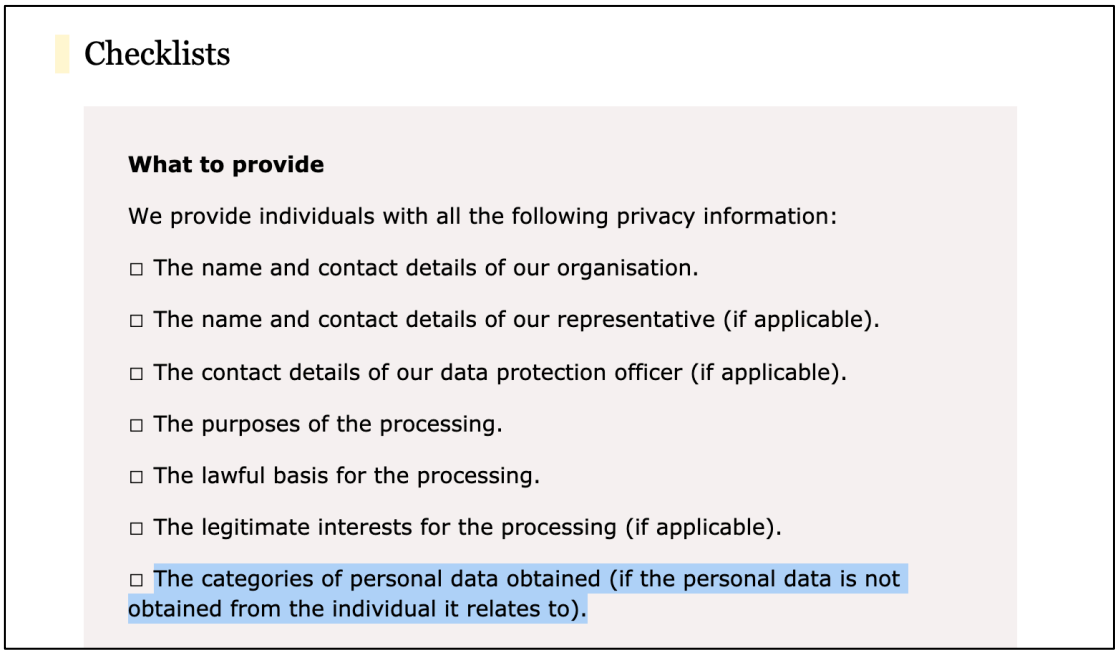
Indeed, the WP stated that provided data included *'as raw data processed by a smart meter or other types of connected objects, activity logs, history of website usage or search activities'*. Whilst the WP may have made a wide interpretation of the definition of 'provided' data in relation to the right to data portability so that there would be more situations in which it could apply, the knock on effect is that it could potentially limit the requirement for controllers in other countries to specify the categories of personal data they process in these scenarios. This is particularly important as this guidance has been endorsed by the EDPB, which is explained further in the next Section.

Thus, although the 2016 version of the Code made some improvement on the previous version by acknowledging this as an information requirement, it was still unclear when

exactly a data controller should inform the individual of the 'categories' of personal data processed and what this consists of.

The introduction of the DPA 2018 has changed the way ICO provide guidance on the obligation to inform. They no longer have an 'ICO Code on Privacy Notices' as this has been superseded by a general 'Guide to the GDPR' which includes a page on the obligation to inform. There are two specific places on ICO's website where they provide guidance on the obligation to inform under the DPA 2018.

In their webpage on the 'Right to be informed'[7], ICO advise in their checklist for what to provide that controllers must provide the '*categories of personal data obtained (if the personal data is not obtained from the individual it relates to)'* as depicted in Figure 8.


Checklists

What to provide

We provide individuals with all the following privacy information:

□ The name and contact details of our organisation.

□ The name and contact details of our representative (if applicable).

□ The contact details of our data protection officer (if applicable).

□ The purposes of the processing.

□ The lawful basis for the processing.

□ The legitimate interests for the processing (if applicable).

□ The categories of personal data obtained (if the personal data is not obtained from the individual it relates to).

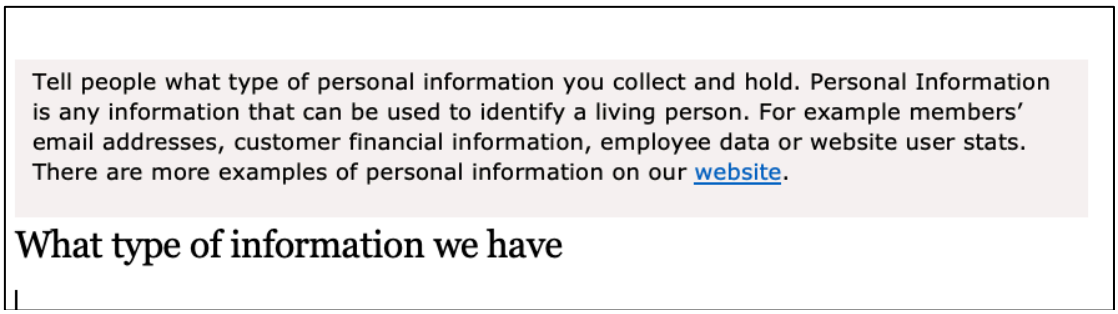
Figure 8 ICO Checklist for the Right to be Informed

This goes against the requirements of the DPA 2018, which always requires the provision of information about the categories of personal data being processed, unless the individual already knows them. As discussed previously, whilst it may be true that if the personal data is obtained from the individual, they are aware of the categories being processed, there

[7] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-be-informed/

will be instances where this is not the case and therefore ICO should not be making this distinction.

Another area where ICO provide guidance on this is in their 'Privacy Notice Template' which is designed to help organisations create legally compliant privacy policies. Despite advising that the categories of personal data only need to be provided if the personal data is not obtained from individuals above, in the template ICO does not make such a distinction and advises that information on the personal data being processed should always be provided.


Tell people what type of personal information you collect and hold. Personal Information is any information that can be used to identify a living person. For example members' email addresses, customer financial information, employee data or website user stats. There are more examples of personal information on our website.

What type of information we have

Figure 9 ICO Privacy Notice Template

It also uses the term 'type of information' opposed to 'categories of personal data' and seems to indicate this should be a list of the specific personal data processed rather than categories of this.

ICO also refer to this requirement in their webpage aimed at data subjects on their right to be informed[8], which states that *"an organisation must inform you if it is using your personal data. It should provide you with information on the following: … What type/types of data it is using".*

Therefore, despite the DPA 2018 making it clear that it expects information about the categories of personal data being processed to always be provided, ICO has provided guidance which contradicts this and has then provides guidance that contradicts itself. As the DPA 2018 is binding it has the authority here, however given the influence of ICO and the fact that most organisations subject to the DPA 2018 will use these resources rather than looking at the text of the legislation, it could lead to confusion and non-compliance with the obligation to inform in practice.

[8] https://ico.org.uk/your-data-matters/your-right-to-be-informed-if-your-personal-data-is-being-used/

4.2.7 European Data Protection Board Guidance

The European Data Protection Board (EDPB) is an independent European body whose purpose is to ensure consistent application of the GDPR, and to promote cooperation among the EU's data protection authorities. It replaced the Article 29 Working Party when the GDPR came into force. Whilst the EDPB has released some of its own guidance, opinion and recommendations for best practice, it has also endorsed various GDPR-related Article 29 Working Party Guidelines which had already been published by them prior to the EDPB. None of the WP Guidance documents and opinions discussed in Section 4.2.2 were endorsed by the EDPB, although this may partly be due to the age of them rather than their credibility. The question of when and if information about the categories of personal data should be provided by the controller is discussed at various points throughout the documents the EDPB provides.

In the Article 29 WP Guidelines on transparency under Regulation 2016/679 (2018) which were endorsed by the EDPB, the WP provided a table in the Annex in which they provided comments on the individual information requirements of Article 13 and 14. In this table, the WP stated that 'categories of personal data' are not required under to be provided under Article 13 as depicted in Figure 10. The table stated the reason they are required to be provided under Article 14 is that *"…in an Article 14 scenario because the personal data has not been obtained from the data subject, who therefore lacks an awareness of which categories of their personal data the data controller has obtained".* However, this contradicts the WP guidance on data portability discussed in Section 4.2.6 which stated that provided data included *"raw data processed by a smart meter or other types of connected objects, activity logs, history of website usage or search activities"'*. It seems ambitious to believe that individuals would be aware of the categories of personal data that are processed as part of this type of activity given the technicality of it. It also contradicts the previous guidance of the WP under the DPD, discussed in Section 4.2.2 that in the context of apps, Article 10 requires that each data subject has a right to know what type of personal data is being processed.

In the same guidance, the working party does provide some clarification on which Article inferred data would fall under, in stating that:

"…Pursuant to the principles of fairness and purpose limitation, the organisation which collects the personal data from the data subject should always specify the purposes of the processing at the time of collection. If the purpose includes the creation of inferred personal data, the intended purpose of creating and further processing such inferred personal data, as well as the categories of the inferred data processed, must always be communicated to the data subject at the time of collection, or prior to the further processing for a new purpose".

Article 29 Working Party (2018)

Required Information Type	Relevant article (if personal data collected directly from data subject)	Relevant article (if personal data not obtained from the data subject)	WP29 comments on information requirement
Categories of personal data concerned	Not required	Article 14.1(d)	This information is required in an Article 14 scenario because the personal data has not been obtained from the data subject, who therefore lacks an awareness of which categories of their personal data the data controller has obtained.

Figure 10 Article 29 Working Party Table on Information Requirements

In the same guidance, the working party does provide some clarification on which Article inferred data would fall under, in stating that:

"…Pursuant to the principles of fairness and purpose limitation, the organisation which collects the personal data from the data subject should always specify the purposes of the processing at the time of collection. If the purpose includes the creation of inferred personal data, the intended purpose of creating and further processing such inferred personal data, as well as the categories of the inferred data processed, must always be communicated to the data subject at the time of collection, or prior to the further processing for a new purpose".

Article 29 Working Party (2018)

This indicates that 'inferred' personal data would be thought of as data not obtained from the individual and fall under Article 14 because the WP is stating that the categories of the inferred data must be communicated to the data subject at the time of creation and have previously stated that this is not a requirement under Article 13.

In their Guidelines 1/2020 on processing personal data in the context of connected vehicles and mobility related applications, the EDPB confirmed the legal requirement for controllers to provide information on the categories of personal data under Article 14 when personal data is not obtained from them directly. The EDPB stated that:

> *"…When data have not been collected directly, the vehicle and equipment manufacturer, service provider or other data controller shall, in addition to the information mentioned above, also indicate the categories of personal data concerned"*.
>
> **European Data Protection Board (2020)**

Additionally, in their Guidelines 04/2020 on the use of location data and contact tracing tools in the context of the COVID-19 outbreak (2020), the EDPB appeared to extend this requirement to inform individuals of the categories of personal data being processed under Article 13 also. The EDPB stated that any contract tracing application should identify the *'The categories of data as well as the entities to (and purposes for which, the personal data may be disclosed)'*. As this is in the context of an app, which arguably falls within the WP's definition above of 'provided data' then the EDPB would be inferring this as a requirement under Article 13 GDPR. This would also be consistent with the previous guidance of the WP in relation to apps under the DPD discussed in Section 4.2.2.

Therefore, in answering the question of when are data controllers required to provide individuals with information about the categories of personal data that they are processing, guidance from the EDPB does not clarify the situation, as whilst the EDPB seems to suggest that it is always required under Article 14 and is sometimes required under Article 13, the guidance of the WP that is endorsed by the EDPB clearly states that it is not required under Article 13.

4.2.8 Summary and Conclusion to RQ2

In these sections, this thesis has discussed the obligation to inform in both the EU and UK in order to answer RQ2 'When is there a legal requirement in the EU and UK to provide information about the specific personal data being processed under the obligation to inform and what is the requirement for this?'

This investigation found that both the previous and current execution of the obligation to inform makes it difficult for data controllers to understand what obligation they are under in relation to informing individuals of the personal data that they process.

Under the Data Protection Directive (DPD), the concept of 'categories of data' was introduced as the method for describing the personal data that a controller processes. However, this was only referred to as an example of something that individuals may need to be told when personal data about them is obtained from elsewhere. In relation to the collection of personal data directly from the data subject, this requirement was not included, even as an example of something that individuals may need to be informed about. Whilst, arguably data controllers could be under an obligation to provide this information with the 'catch-all' requirement for them to provide any information that is required for the processing to be fair, it certainly wasn't a mandatory requirement for controllers to provide this in all processing scenarios. The DPD also did not provide any further guidance on the specific situations that would mandate provision of this information.

Despite this lack of clarity, the Article 29 Working Party (WP) repeatedly referred to the requirement for controllers to inform individuals about the personal data they process, although they did not always use the term 'category' to describe this requirement. The WP also referred to this requirement in relation to both scenarios, where data is obtained from the individual and when it is not. Whilst this indicated that controllers were required to provide this information, the advice of the WP was persuasive rather than binding on them. Furthermore, by referring to the requirement using different terms such as 'type' and 'nature of the data' it was unclear exactly what information controllers were expected to provide.

Further uncertainty was created in the UK under the Data Protection Act 1998 (DPA 1998), the UK's implementation of the DPD. This was because whilst the DPD referred to two different processing scenarios, with slightly different information requirements, the DPA 1998 combined these into one set of requirements, which applied to all processing scenarios. However, in doing so, they removed the example of 'categories of data' as something that individuals may need to be informed about. Whilst arguably providing this information could still be required under the 'catch-all' provision (of providing any information required for processing to be fair) the omission of 'categories of data' as an example arguably lessened the importance of providing this. Furthermore, the Information Commissioner's Office (ICO) as the UK Regulator did not refer to this requirement (even as an example of information that may need to be provided to individuals) in their 'ICO Privacy Notices Code of Practice' which was their guidance on compliance with the obligation to inform at the time. Thus, whilst an obligation to provide individuals with this information under the DPA 1998 could have been inferred, it was certainly not mandatory and also not something that controllers were actively made aware of.

The reason there was this uncertainly under the DPD may have been because it was created at a time when personal data was predominantly 'provided' by individuals and therefore the assumption was that individuals were aware of the personal data that was processed by controllers. Yet, in the time of the DPD, increasingly, personal data was obtained from elsewhere, and also provided by individuals without them knowing, which may explain why the WP repeatedly inferred this requirement. Whilst the introduction of the GDPR in 2018 provided some clarity, it did not make it completely clear when organisations are under an obligation to inform individuals about the personal data that they are processing. The GDPR did make it a mandatory requirement to inform individuals of the 'categories of data' when data is not obtained from them yet remained silent on whether this would be required in situations where personal data is obtained directly from the individual. Furthermore, the distinction between provided data and 'data obtained from elsewhere' was still not entirely clear in the context of online processing.

Advice from the Article 29 Working Party (WP) and the European Data Protection Board (EDPB) did not provide clarity on this. In fact, the WP's guidance, which was endorsed by the EDPB, contradicted its previous position, by stating that controllers were not required to provide information on 'categories of data' in scenarios where personal data was

obtained from the individual. Given that the WP had previously advised that 'provided data' included data generated by cookies and online services etc., much of which is unobvious to the individual, this seemed an unusual position to take. Although guidance from the EDPB did confirm that providing this information would always be required in situations where data is not obtained from individuals (in line with the GDPR) and has at least once extended this requirement to situations where data is provided by them, it is still unclear from their guidance whether there is a legal requirement for controllers to inform individuals of the categories of personal data that they are processing about them when they obtain data directly from them.

Interestingly, whilst the implementation of the DPD in the UK arguably lessened the requirement for (or at least clarity around) controllers to provide information about the categories of personal data that they process, the introduction of the DPA 2018 actually created clarity on this question. In combining the requirements of the two different processing scenarios of the GDPR into one, it made it a mandatory requirement for controllers to provide information on the categories of personal data processed in all situations. However, despite this clarity under the DPA 2018, ICO have created confusion in the guidance they have produced. They have simultaneously advised both, that the categories of personal data do not need to be provided in situations where personal data is obtained from the individual, but also that controllers should always provide information on the types of personal information that they collect and hold. This guidance not only contradicts the requirements of the DPA 2018, but also itself. Whilst in some situations where data is obtained from individuals there will not be a requirement to inform them of the personal data being collected, because the individual will be aware of this, if the WP's definition of what 'provided' data means is correct then there will certainly be situations where the individual is not aware and will need to be informed for processing to be fair and transparent. Thus, whilst the DPA 2018 is clear, it will be the guidance of ICO which controllers are more likely to look to, which makes this lack of clarity and incorrect guidance a problem.

In conclusion and in answer to RQ2, the position under the GDPR remains unclear on when organisations are under an obligation to provide individuals with information on the categories of personal data that they process about them and whilst the law is clear in the UK, the guidance on it is not. Therefore, to support controllers in understanding their

obligations clarification and consistency from regulators is required here. In deciding what the recommendation should be from them on this matter it is important to understand whether there is a benefit in providing this information, whether a consistent approach to doing so has emerged organically in practice or at least an approach that can be adopted to increase transparency. These points are analysed in the next chapter.

Chapter 5 Categorising Personal Data in Practice

5.1 Introduction

As discussed in the previous chapter, in answering RQ2 this thesis found that the law and guidance on it is not clear on whether data controllers are required to provide information about the specific personal data that they process about individuals under the obligation to inform and where they do require this, it is the 'categories of data' that must be detailed. However, other than detailing what 'special categories' of personal data are in (Article 9 GDPR), there is no further guidance on how to categorise personal data in the law or in the guidance on it by data protection bodies. This led to the next phase of research in this thesis which sought to answer RQ3:

RQ3: Does various current approaches to categorising personal data and informing individuals of these achieve the aims of transparency under the European Union Data Protection Framework?

As this question is rather broad, it was broken down into the following research sub questions:

RSQ1: What are the benefits of categorising personal data and does informing individuals of this increase the transparency of personal data processing?

RSQ2: Is there a consistent approach to categorising 'non- special category personal data' in practice?

RSQ3: Are any of the current approaches to categorising personal data in practice sufficient in making the processing of personal data transparent?

RSQ4: How are SNS categorising personal data and informing individuals of this in practice and does this achieve the benefits for transparency?

RSQ1 then sought to understand what the benefits of categorising personal data are and whether providing information on which of these are being processed has the potential to increase the transparency of personal data processing for individuals. RSQ2 was designed to understand what the current approaches to categorising personal data are in practice, and RSQ3 investigates whether they achieve the potential benefits for transparency. The purpose of these research sub questions was to explore whether the approaches in practice are sufficient, or whether the categorisation of personal data has potential, as an area where improvements can be made to the way data controllers are transparent about their personal data processing in accordance with the goal of this research. RSQ4 looks at how SNS are categorising personal data and informing individuals of these (as the behaviour which prompted this research) in practice and whether this achieves the benefits for transparency that a good categorisation of personal data offers.

5.2 Methodology and the Data

This part of the thesis uses the same methodology as the previous chapter of critical analysis (Vibhute & Filipos Aynalem, 2009) and argumentation theory (Van Eemeren, et al. 2013) which uses logical reasoning to reach conclusions. This investigation looked at the policies of the same six SNS as the investigation discussed in Chapter Three. However, since the first investigation, all six had been updated and therefore this investigation looked at the policies as available in January 2016. Again, Facebook, Twitter, LinkedIn and Pinterest all have a single policy; specific to them and Google+ is governed by Google's 'all-in-one' policy, covering all their services.

Interestingly, (as mentioned in Chapter Three) at the time of the first investigation, the Yahoo! 'all-in-one' policy was displayed when the 'privacy' link was clicked on the Flickr home page. Given that the page about 'Flickr' as a service was three hyperlinks away from the Flickr homepage and the second link was buried deep in the Yahoo! policy, the content of this page was not included in the data sample. However, when collecting the policies for analysis in January 2016, the situation was reversed. The 'privacy' link on the Flickr home page instead leads directly to the Flickr-specific page initially, with a link to the Yahoo! all-in-one policy contained within this. Because of this, both policies were analysed as part of the data sample.

5.3 Results

5.3.1 Benefits of Categorising Personal Data

As discussed in the previous chapter, because of the lack of clarity in the law and the guidance in the EU and the UK on whether organisations need to provide individuals with information about the categories of personal data that they are processing, there is a need for clarification on this. The ambivalence of regulators on this matter could be interpreted as them not viewing this an important piece of information to provide or that distinguishing between personal data has any benefit. However, various parties have highlighted the importance of this information for the processing of personal data to be transparent. This conflict lead to RQ2:

RQ2: What are the benefits of categorising personal data and does informing individuals of these in general and the current approach in practice to doing so increase the transparency of personal data processing?

RQ2 was designed to understand whether there is a benefit in distinguishing between personal data, and if so, what the benefits of providing information on this to individuals is for increasing the transparency of personal data processing are and whether the current approaches in practice are achieving transparency. This question needed to be answered in order to understand whether pursuing this area of research and potential improvement to the way data controllers are transparent about their personal data processing (the research goal of this thesis) would be worthwhile. It would also indicate the clarification that European regulators should provide.

The following sections discusses the findings in more detail, which are not intended to be exhaustive, but argues that there are benefits for transparency that providing information on the categories of personal data being processed creates. The reason that this is not necessarily an exhaustive list of the benefits is because those identified are enough to support the conclusion that distinguishing between personal data beyond the binary distinction of personal data and special category personal data has many benefits and that informing individuals of these can increase the transparency of personal data processing.

The benefits presented here were split into two overarching themes, benefits from knowing the category of personal data being processed and benefits from using the categories as an anchor for further information about the personal data processing.

5.3.2 Benefits from Categorising Personal Data

The findings from RSQ1 found that distinguishing between personal data and knowing which simply knowing which ones are being processed has many potential benefits for transparency.

Firstly, distinguishing between personal data and being informed of the categories of personal data being processed can enable an assessment of any risk involved in the processing of them. This assessment could be undertaken by individuals under the obligation to inform but also various other parties, including data controllers or supervisory authorities. Understanding the differences between categories of personal data allows for an assessment of the different risks that may be involved in the processing of them. Indeed, the same category of personal data may have different levels of risk for different people, some may view one category being processed as having a higher risk to their privacy than others. Thus, categorising personal data and being informed of which one is being processed allows individuals to make decisions about the risk involved and therefore whether to participate in processing or not.

Second, distinguishing between personal data and being informed of the categories of personal data reduces the amount of information that needs to be provided to individuals to increase transparency under the obligation to inform. If personal data were appropriately divided into categories that allowed individuals to understand more about the personal data processing simply from knowing which category (or categories) were processed, this could reduce the amount of information currently required to create this understanding. The benefit here is similar to the benefit envisaged by providing standardised icons under Article 12(7) GDPR i.e. giving a meaningful overview of the processing. One of the key criticisms of the manifestation of the obligation to inform is that it is that it generally results in long and complicated privacy notices, which are never read (McDonald and Cranor, 2008). This length and complexity reduces the transparency of processing, because it dissuades individuals from reading them. By not reading privacy policies, in practice, data subjects understand very little information about the processing

of their personal data. Thus, appropriately categorising personal data has the potential to reduce the amount of information that needs to be provided, removing the disincentive for individuals to engage with this information.

5.3.3 Categories as Anchors

In the findings to RSQ1, a second overarching benefit of distinguishing between personal data and informing individuals of the categories of personal data being processed was highlighted. This was that categories can be used as an anchor to which further information about processing can be attached. This has the potential to increase the transparency of personal data processing by contextualising the information provided under the obligation to inform. Various information can be attached to categories of personal data, but two examples are provided here.

Firstly, informing individuals of the categories of personal data being processed allows controllers to contextualise other information they provide under the obligation to inform, by linking it to a category of personal data. An example of this is the time for which the personal data will be kept by the data controller. Article 13(2)(a) and Article 14(2)(a) GDPR advise that a data controller may need to inform individuals of the 'period for which personal data is stored' for the processing to be fair and transparent. Thus, first specifying the categories of personal data processed then allows other information such as 'time limits for storage' to be attached to it. This allows an individual to understand what personal data will be deleted and when. This is also true for other information, such as source from which the category of personal data was obtained or for which specific purposes it will be used.

Secondly, informing individuals of the categories of personal data being processed allows for the attachment of different levels of protection or obligations and rights, in relation to different personal data. For example, under the GDPR (Article 9(1)), 'special categories of personal data' are defined. Article 9(1) GDPR defines these categories as the *'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*. These categories are then given a higher level of protection than 'normal' personal data, in that processing of them is

prohibited, unless under an applicable exception as provided for in Article 9(2) GDPR. Thus, by identifying categories, a different level of protection can be attached, including different rights and obligations. This has the benefit of providing a more nuanced approach to data protection. This nuance could be used to support a risk-based approach to regulation, which is arguably required giving the ever-expanding definition of personal data (as discussed in Chapter 2.

5.3.4 Summary

In answering RSQ1, as can be seen from this non-exhaustive list, there are various benefits for transparency of personal data processing to be achieved by appropriately categorising personal data. It can enable an assessment of the risk involved in the processing and reduce the information that needs to be provided to individuals for the processing of their personal data to be transparent. It can also be used to contextualise other information provided under the obligation to inform. Indeed, these benefits form the initial four requirements for a good categorisation under the obligation to inform which are discussed further in Chapter 5. Given these benefits, it would seem logical for regulators to advise that controllers are always required to inform individuals of the categories of personal data being processed about them, unless they can prove that the individual is already aware of these. However, in providing this recommendation it is important that a clear and robust categorisation of personal data is also provided, which begs the question of what this categorisation should be in practice. The only detail on this provided by the GDPR and DPA 2018 is a categorisation of personal data as either special category or non-special category personal data, with only special category personal data broken down into further categories. This led to RSQ3 and RSQ4 which sought to understand, given the lack of guidance from the legislation, how organisations are categorising personal data in practice in order to understand whether any of these categorisations could be used in their current form to meet the benefits for transparency of categorisation highlighted in Sections 4.3.1 and 4.3.2. The findings in relation to these RSQ's are discussed in the following sections.

5.4 Categorisation of Personal Data in Practice

5.4.1 Introduction to RSQ2

As has been demonstrated in Chapter 4, it is clear that clarification of the law is required. However, understanding when a data controller is under an obligation to inform the data

subject of the 'categories' of personal data is not the only issue that needs attention. Even if the law was clarified, so that it was clear when controllers are under an obligation to inform individuals of the categories of personal data that they are processing about them, there is still the issue of exactly what a 'category of personal data' is in relation to the obligation to inform. This question will need to be answered in order to realise the benefits which categorising personal data and informing individuals of these has for transparency. Whilst it may initially seem obvious what a 'category' of personal data is for this obligation, an informal roundtable discussion hosted by the OECD (2014), involving a cross-section of more than sixty-five privacy experts, recognised that categorising personal data could in fact be approached in numerous ways, both explicitly and implicitly.

The only guidance the legislation provides on how to categorise personal data is to distinguish between personal data and 'special category' personal data. The GDPR then provides of a categorisation of 'special categories' of personal data. These categories are distinguished from other personal data because the processing of them warrants further protection and obligations for controllers under the framework (Article 9(1) GDPR). Thus, when the requirement to inform individuals of the 'categories of data' is required under the obligation to inform, it could be interpreted that the requirement refers to informing the individual of whether 'special categories' of personal data are processed and what these are. Yet, interpreting the requirement in this way would miss out on informing individuals about a large amount of the personal data that is processed, which would have a big impact on the transparency of personal data processing. Additionally, in practice the preliminary investigation for this thesis discussed in Chapter 3 found that SNS were categorising non-special category personal data and providing information on this in their privacy policies. This also indicated that SNS saw a benefit to the transparency of personal data processing of categorising personal data and informing individuals of these.

Whilst ICO, the WP and the EDPB have provided examples of what items, types and categories of personal data are (see Figure 9 for example), none of these have provided a comprehensive approach to categorising non-special category personal data in practice that could be identified and adopted readily in relation to the obligation to inform.

These findings lead to RSQ2:

RSQ2: Is there a consistent approach to categorising 'non- special category personal data' in practice?

RSQ2 was designed to understand what the current approaches to categorising non-special category personal data are in practice and whether there is consistency in them so that a categorisation of personal data could readily be adopted and recommended by the regulator to be used in relation to the obligation to inform. In answering RSQ2, this thesis looked beyond SNS, to categorisations of personal data that other sources were also using.

In terms of the method used to answer this question, a systematic literature review (Okoli, C. and Schabram, K,. 2010) was considered, which uses systematic methods to collect secondary data. However, initial analysis identified that there were many different methods being used in practice to categorise personal data and therefore there was not a requirement in answering RSQ2 to systematically look for all of the different ways that personal data could be categorised. Thus, argumentation theory was the method used to answer this question by collecting random secondary sources and constructing a logical argument to answer the question. Additionally, whilst a formal measure of similarity could have been used to indicate consistency, such as Jaccard's Similarity Coefficient which was used in the preliminary study discussed in Chapter 3, initial analysis also identified early on the dissimilarities between the categorisations in practice. Therefore, to answer RSQ2 an exact measure of similarity was not required, just an indication of whether there was consistency which could easily be assessed by eyeballing the data without the need for such measure.

5.4.2 Categorisation of Personal Data in Practice

Table 12 contains the results of the analysis that was undertaken in order to answer RSQ2 (Is there a consistent approach to categorising 'non- special category personal data' in practice? and evidences the various different ways that personal data has been differentiated between in practice by different parties. As discussed, the notion that 'categorising' personal data is the correct approach to differentiating between the personal data is because the text of the DPD, GDPR and DPA 2018 all referred to this being the granularity of information required when discussing the information requirements of the obligation to inform.

However, as discussed in Section 4.2, confusingly the WP and ICO have referred to this information obligation using various different additional terminologies, from the 'nature of the data' to the 'type of personal data', without clear and consistent examples of whether these terms are similar or different, and what they encompass. This can also be seen in practice, as Table 12 shows as many different sources have also referred to 'items' and 'types' of personal data. Some sources linked these items and types of personal data back to categories of personal data, but others did not.

Thus, in answering RSQ2, Table 12 evidences the divergence in approaches to categorising and differentiating between personal in practice because there is not further guidance on how to do this. These examples were not gathered using a systematic literature review but were gathered as part of the literature review which was undertaken to learn more about transparency. The selection criteria for examples were that there was some form of approach to try and divide the personal data in a way, be that into categories, types or items. With so many different approaches, even if the law were to be clarified so that a data controller could understand when they are under an obligation to inform the data subject of the categories of personal data they process, without further clarification, it is still unclear exactly what information they would need to provide. This means that clarification would also be required on which, if any, of these approaches to categorising personal data is the one referred to in relation to the obligation to inform. This requirement led to RSQ3, which sought to understand whether any of these categorisations could achieve the benefits for transparency that were highlighted in Section 4.3.

Table 12 Examples of Different Approaches to Distinguishing Between Personal Data in Practice

Source	Category of Data	Type of Data	Item of Data
Leon et al (2013) study	Computer-related information, Demographic and preference info, Interactions with the website, Location information, Personally identifiable information	N/A	Length spent on each website page, operating system, age, gender, hobbies, country where visiting website from, name, credit card number.
Platform for Privacy Preferences (P3P) 1.0 Specification Section 3.4 (Reagle and Cranor, 1999)	Physical Contact Information	N/A	Telephone number, address
	Online Contact Information	N/A	Email
	Unique Identifiers	N/A	N/A
	Purchase Information	N/A	Method of payment
	Financial Information	N/A	Credit or debit card info.
	Computer Info	N/A	IP no., domain name, browser type, operating system.
	Navigation and Click-Stream Data	N/A	Pages visited, how long users stay on each page
	Interactive Data	N/A	Queries to a search engine, or logs of account activity.
	Demographic and Socioeconomic Data	N/A	Gender, age, income
	Content	N/A	Text of email, bulletin board postings, or chat room communications
	State Management Mechanisms	N/A	N/A
	Political Information	N/A	Membership/affiliation with groups such as religious organizations, trade unions, professional associations, political parties, etc.
	Health Information	N/A	Sexual orientation, use or inquiry into health care services or products, and purchase of health care services or products.
	Preference Data	N/A	Favourite colour, musical tastes.
	Location Data	N/A	GPS position data
	Government-issued Identifiers	N/A	N/A
	Other	N/A	N/A

Source	Category of Data	Type of Data	Item of Data
Allen & Overy's Guidance on Binding Corporate Rules (2016)	Employment Data, Client Data	N/A	N/A
E-Privacy Directive (2002)	Traffic Data	N/A	Routing, duration, time or volume of a communication, protocol used, location of the terminal equipment of the sender or recipient, network on which the communication originates or terminates, beginning, end or duration of a connection
	Location Data	N/A	Latitude, longitude and altitude of the user's terminal equipment, direction of travel, level of accuracy of the location information, the identification of the network cell in which the terminal equipment is located at a certain point in time, time the location information was recorded
OECD Privacy Expert Roundtable (2014)	*(Categorisations in relation to the sensitivity of the data)* Health Data Ethnic Origin	N/A	N/A
	(Categorisations in relation to the subject of the data) Employee Data Minor's Data Non-citizens Data	N/A	N/A
	(Categorisations in relation to the context in which the data is being processed) Electronic Communications Data, Credit Reporting Data, Archival Data, Social Security Administration Data	N/A	N/A
	(Categorisations in relation to the degree of identifiability) Identifying Data De-identified Data	N/A	N/A

Source	Category of Data	Type of Data	Item of Data
	Anonymous Data Pseudonymous Data		
	(Categorisations in relation to how the data has been collected) Directly collected data Indirectly collected data	N/A	N/A
	(Categorisations in relation to the manner in which the data originated) Provided Data Observed Data Derived Data Inferred Data	N/A	N/A
UK Data Protection Register Requirements under DPA 1998	N/A	Personal Details Family, Lifestyle and Social Circumstances Financial Details Employment and Education Details Goods or Services Provided	N/A
MyDex White Paper '*The Case for Personal Information Empowerment: The rise of the personal data store*' (2010)	N/A	Data that identifies me	Name, address
	N/A	Data conferred by other parties	Passport number, my credit reference rating
	N/A	Information gathered by me	Search and research results
	N/A	Data generated by my dealings with other parties	Transaction and interaction records)
	N/A	Information created by me	My plans, my preferences
General Data Protection Regulation (2016)	Data revealing: Racial or ethnic origin Political opinions, Religious or philosophical beliefs, Trade-union membership,	N/A	N/A

Source	Category of Data	Type of Data	Item of Data
	Genetic data, Biometric data Data concerning health Data concerning a natural person's sex life or sexual orientation	N/A	N/A

5.5 Categorisation Ability to Increase Transparency

5.5.1 Introduction to RSQ3

In Chapter 4, the findings showed that there is a requirement for data controllers to provide individuals with information on the categories of personal data processed about them, but that there was a lack of clarity between the law and the guidance on it in the EU and the UK as to when exactly this applies. RSQ1 of this chapter found that there were many benefits for transparency of differentiating between personal data and informing individuals of the categories of personal data being processed. RSQ2 found that the lack of a robust and sophisticated categorisation of non-special category personal data by the law and data protection regulators has led to many different approaches to categorising personal data emerging in practice. Thus, to realise the benefits of categorisation of personal data for transparency, these categorisations need to be examined to understand whether they increase the transparency of personal data processing and could be adopted as a categorisation of personal data by European data protection regulators. This led to RSQ3:

RSQ3: Are any of the current approaches to categorising personal data in practice able to achieve the potential benefits categorising personal data and informing individuals of these under the obligation to inform?

The purpose of RSQ3 was to consider whether any of these categorisations could achieve the benefits for transparency that were discussed in Section 5.4 which would mean that they could potentially be adopted by a European regulator.

In order to answer this question, the categorisations themselves could be assessed against these benefits individually. However, it was clear from initial analysis that there were various overlaps between the categorisations in practice in terms of the information they contained and the approaches they took to categorising personal data. It was for this reason that the six steps of Thematic Analysis as described in Chapter 3 were applied to the categorisations to generate themes which described the approach they took to categorisation of personal data. Generating themes allowed for an understanding of the motive on theory underlying the categorisations. These themes, as approaches to

categorising personal data, were then assessed against the benefits for transparency that were identified in Section 4.3. This provided an assessment of whether they are sufficient to make the processing of personal data transparent and thus could be adopted by a regulator.

In applying Thematic Analysis to the categorisations, four themes of approaches to categorisations emerged:

1. Categorisations in relation to what the personal data is e.g. Name, address etc.
2. Categorisations in relation to identifiability
3. Categorisations in relation to sensitivity
4. Categorisations that focuses on the source of the data

These themes also formed four of the requirements for an appropriate categorisation of personal data under the obligation to inform which are discussed further in Chapter 5.

Table 13 Themes of Categorisation

Key

Theme Met	
Theme Not Met	

	What the Personal Data Is	Identifiability	Sensitivity	Source
Leon et al study;				
Platform for Privacy Preferences (P3P) 1.0 Specification				
Allen & Overy's Guidance on Binding Corporate Rules				
E-Privacy Directive				
OECD Privacy Expert Roundtable				
UK Data Protection Register				
MyDex White Paper '*The Case for Personal Information Empowerment: The rise of the personal data store*'				
General Data Protection Regulation (2016)				

Table 13 shows how the individual categorisations in Table 12 map to the themes of information identified by the Thematic Analysis. As can be seen from Table 13, some categorisations fall into more than one theme e.g. the special categories of personal data as specified by the DPD and GDPR detail both what the data is, and what is considered to be personal data but also a categorisation by sensitivity by categorising this personal data as 'special' compared to non-special category personal data. The most common approach in all the categorisations was to describe what the data is, which all approaches did to some extent.

In order to assess whether any of these approaches to categorisation were likely to achieve transparency in practice, the next section looks at the themes of categorisation through the lens of the benefits of categorisation identified in Section 4.3. These benefits are outlined in Table 14 with an overall assessment of how the categories meet them in practice, which is explained in the following four subsections.

Table 14 Themes of Categorisation and Benefits for Transparency of Categorisation

Key

Benefit Met	
Benefit Partially Met	
Benefit Not Met	

Benefit	Identifiability	Sensitivity	What the data is	Source
Enables an assessment of risk				
Reduces the overall information that needs to be provided to increase transparency				
Allows controllers to contextualise other information provided under the obligation to inform				
Allows for the attachment of different levels of protection or obligations and rights in relation to the categories				

5.5.2 Categorisation in relation to identifiability

Categorising in relation to the degree of identifiability e.g. identifying data, de-identified data, anonymous data and pseudonymous data has a number of benefits and applications in general under the data protection framework. It can be useful for helping to ascertain when data protection laws apply, as the GDPR and DPA 2018 only apply to information that relates to an identified or identifiable individual. It can also inform an assessment of the appropriate technical and organisational measures that are required to ensure the security of the personal data, with the level ranging from low to high depending on how identifiable the individual is from the data alone. However, in relation to the benefits that will help achieve transparency in practice under the obligation to inform, categorisation in relation to identifiability is less promising as it offers some benefits but does not achieve all of them in practice.

Categorising personal data in relation to identifiability can be useful for understanding the ability of the data controller to link the data to the individual, which arguably allows for an assessment of risk. For example, whilst still falling within the definition of personal data, the processing of pseudonymous data is seen to be less risky than personal data that has not been pseudonymised. Therefore, if a controller advises which category it is processing, this can in theory help the individual to make a decision of the risk to them. It can also allow for the attachment of different levels of protection, for example 'anonymous data' does not attract any obligations or rights as it is outside the scope of the framework.

Indeed, Schwartz and Solove (2012) proposed a model of Personally Identifiable Information (PII 2.0) based on distinguishing between identified' and 'identifiable' data called PII 2.0. PII is the concept of protected data in the US Data Protection Framework but it differs slightly from the definition of personal data in the EU. Schwartz and Solove (2012) viewed identification as a continuum of risk rather than a dichotomy and that their approach avoids the US reductionist view and the EU's expansionist view of 'protected data' (2012:1817). Their model places information on a continuum that begins with no risk of identification at one end and ends with identified individuals at the other. They divide this spectrum into three categories, each with its own regulatory regime. Information can be about an (1) identified, (2) identifiable, or (3) non-identifiable person. Because each section does not have hard boundaries, they define them as standards, allowing broad

discretion to take account of relevant factors (2012:1877).

However, whilst providing information on this type of categorisation would not increase the amount of information that needs to be provided, it would not reduce it, which is why this type of categorisation is identified as partially met in relation to this benefit in Table 14. The categories would be quite easy to provide but the categories do not provide summaries which make it easier to reduce information by incorporating other information into them. Furthermore, although this categorisation would allow controllers to contextualise other information, it is arguable what benefit this would give individuals in terms of transparency beyond the benefits in relation to risk analysis mentioned above. For example, an individual may be informed that *'we keep pseudonymous data for three months after you close your account'*. Whilst this does link further information to data, there is still the question of what data it is that has been pseudonymised. Indeed, to check whether the data controller is compliant, knowing what information is being processed is required, to allow them to make subjective and granular decisions about these aspects. Without this granularity, although data subjects may know when data protection laws apply, they will not be able to assess compliance.

5.5.3 Categorisation in relation to sensitivity

Categorisation in relation to sensitivity by distinguishing between 'personal data' and 'sensitive personal data' and informing individuals of which category a data controller processes could prove useful for data subjects. It could help them understand the general sensitivity of the data in question and help them to keep a check on data controller compliance with other obligations under the framework in relation to processing special categories. This approach to categorisation also allows different rights and obligations to be attached to different categories. Indeed, under the GDPR the processing of special category personal data attracts additional obligations for the controller because of the increased risk to the rights of individuals that processing this personal data attracts.

It can also help a data subject to assess the risk of the personal data processing by knowing whether it was sensitive personal data or non-special category personal data that is being processed. However, this is quite a binary distinction and so whilst it allows for an objective assessment of risk, it does not allow for a subjective assessment as the processing of data that is classed as sensitive will increase the risk for everyone, compared to the processing

of non-special category personal data. To make data processing transparent, categories need to have a lower level of abstraction than just 'personal data' and 'sensitive personal data'. This lower level could be informing the individual of the specific personal data which allows the individual to make subjective risk assessments. For example, in using an online service, being informed that *'data collected by cookies will be processed to infer sensitive categories of personal data'* is objectively risky, whereas being informed that *'data collected by cookies will be processed to infer your sexuality'* may present more risk to some than others.

However, despite these benefits, as with categorising in relation identifiability, whilst this approach does not increase the information that needs to be provided, it also does not reduce it. The approach also allows other information to be attached to the category but does not really contextualise it, because it is such a binary categorisation.

5.5.4 Categorisation of what the data is

Because of the deficiencies with the approaches to categorisation discussed in Sections 5.6.3 and 5.6.4, both of which called for a lower level of abstraction, categorising personal data in relation to what the specific fields of personal data are in a systems sense has potential promise. Indeed, in providing a lower level of abstraction for 'non-sensitive' personal data, some of the other categorisations in Table 12, such as those of Leon et al (2013), P3P (1999), and the e-Privacy Directive (2002), could prove useful. Focusing on a data controller informing the individual of these could see a move towards the creation of a taxonomy of personal data, as called for by the World Economic Forum (2014).

This approach can allow for an objective and subjective assessment of risk as discussed in 4.5.3, it can also contextualise other information provided, for example by explaining exactly what personal data is used to infer other personal data so that individuals can understand what they need to withhold to prevent such an inference being made. It also allows for the attachment of different levels of protection to data, as under the GDPR with special category data where specific categories are identified.

However, a big issue with this approach to categorisation in increasing transparency is that it would increase the amount of information that would need to be provided to individuals

infinitely. As discussed in Chapter 2, the definition of personal data is very wide and given the vast amount of personal data that controllers, and in particular online services process, this would be an almost impossible approach to categorising personal data to achieve in practice.

It would also face other issues. First, there is the issue of creating a taxonomy that is simultaneously able to accommodate new forms of personal data as new technologies emerge; remain simple enough for data controllers and data subjects to comprehend; and be able to deal with personal data that belongs to more than one category. Indeed, one of the criticisms of the Platform for Privacy Preferences Project (P3P) (2013) (and its lack of adoption) was because their approach to categorising data was too complex, even for webmasters (Schwartz, 2009), and their taxonomy only included seventeen 'data types'. Given the broad definition of personal data and the rate at which new forms of data are being created and used, such a taxonomy is likely to become confusing quickly. This would suggest that any approach to categorisation would need to include far fewer categories than seventeen.

Second, is the issue of deciding on the granularity (or level of abstraction) of the categories, which requires making trade-offs between specificity and practicality, in particular if you are trying to categorise all possible personal data into less than seventeen categories. An example of a P3P category is 'computer information', yet even this is a wide category, which does not make it intuitively obvious what it includes. This means that exactly what is being processed is still not transparent. Whilst increasing the granularity, to state that the data subject's 'IP Address' is processed may increase transparency, it will also result in even longer privacy notices and cognitive overload, given the amount of different data types there already are in existence, let alone those to be created and that will be inferred from collected data.

Thus, whilst in accordance with Table 14, categorizing personal data in relation to what it is has the most promise for achieving the benefits for transparency of the current approaches, it would not be possible alone to use it to achieve in maximum transparency in practice.

5.5.5 Categorisation in relation to source

Due to the issues with categorising in relation to what the personal data is, categorising in relation to source could provide potential here. For example, the categories produced by the OECD's Privacy Expert Roundtable, of 'provided', 'observed', 'derived' and 'inferred' personal data (2014) could prove useful.

To some extent the approach allows for an assessment of risk, because if an individual is informed that by using a service personal data will be observed or inferred about them they can understand that it will not just be the data that they provide that the controller will have access to about them. However, in terms of true risk assessment there is still a need to have more granular information about what this personal data is, and what exactly is being 'observed', 'derived' and 'inferred', which would support them in making subjective choices about this and/or acting as a check on data controllers.

However, as with the categorisations in relation to identifiability and sensitivity, whilst further information can be attached to these categories, it would not contextualise and increase transparency in the same way that specifying the specific personal data being processed does, to allow individuals to foresee the implications of processing. It also does not reduce the information that needs to be provided to individuals by summarising other information, although adopting this approach would not increase the amount of information that would need to be provided.

5.5.6 Conclusion to RSQ4

In answer to RSQ4 (Are any of the current approaches to categorising personal data in practice able to achieve the potential benefits categorising personal data and informing individuals of these under the obligation to inform?' it is clear that none of the approaches alone are able to achieve the benefits of categorising personal data for transparency. Whilst, informing individuals of the specific data that is processed is the most promising, in practice it would not be feasible to categorise personal data in this way and would increase the length of privacy policies as these continued to be used to provide this. Therefore, a new approach to categorising personal data will be required alongside any recommendation from the regulator to always inform individuals of the categories of personal data that are being processed under the obligation to inform.

This RSQ took a top down approach to look at whether there was a complete categorisation of personal data highlighted by this research that could be adopted as the categorisation of personal data under the obligation to inform. Whilst no complete categorisation of personal data that was sufficient to increase the transparency of personal data processing was 'out there', in looking to propose a new approach it was also important to understand how controllers are categorising personal data in practice and what we can learn about their behaviour to inform the proposal of a new category. This led to RSQ4 of this investigation.

5.6 Categorisation by SNS in Practice

As discussed, RSQ3 focused on the top down approach of categorization and found that there is not an approach to categorization of personal data in practice that can be adopted to increase transparency under the obligation to inform. This meant that in making a recommendation for regulators to mandate that controllers inform individuals of the categories of personal data that they are processing about them, a new approach to categorizing personal data would need to be proposed.

As it was the finding of the preliminary study in Chapter 3 of this thesis that SNS were providing information on the specific personal data that they are processing which prompted the investigation discussed in this Chapter, to propose a new categorization it would be beneficial to understand how controllers are addressing this in practice which led to RSQ4 'How are SNS categorising personal data and informing individuals of this in practice and does this achieve the benefits for transparency?'.

To be consistent with the preliminary study, the privacy policies of the six SNS which formed part of the first study in this thesis (as discussed in Chapter 3) were examined to see how they discussed the personal data that they processed, the justifications for which are outlined in Section 3.2.2.

The six stages of Thematic Analysis which are described in Chapter 3 were then used to identify themes from the clauses in the privacy policies that included information about the personal data that they processed. The results identified four trends in the way all six of

the privacy policies provided information about the actual personal data they collect and process: they were not exhaustive in the information they provided; they focused on the source of the personal data; they separated the information about the personal data processed from the purpose that it would be used for; and they used differing to describe the personal data. These themes are discussed further in the next four subsections.

5.6.1 Themes in Providing Information About Personal Data

First, all six policies were not exhaustive in detailing the personal data that they processed. They all used words or phrases indicating this when they provided information about the personal data that they processed from, 'this includes', to 'for example', 'among other information' and 'like'. LinkedIn even confirmed that it does not aim to be exhaustive in its policy about all the personal data that it processed. This practice means that the information SNS are providing about the personal data that they process is not transparent. Because information today is increasingly observed, derived and inferred about individuals (OECD, 2014), without being informed exhaustively of what is collected and processed, individuals have little way of ascertaining this. Furthermore, with incomplete information any assessment of risk will be invalid because it will be based on incomplete information. As Google has received instructions from both the WP and ICO on this matter (as discussed in Section 4.2.6) which has advised them to provide an exhaustive "*list of the types of data processed by Google"* this was a surprising finding in relation to them. However, given the issues with actually doing this in practice without further guidance from a regulator, to actually do this in practice would be difficult.

A second theme was that the SNS often focused on the process of collecting the personal data, and in particular, the source of the personal data. For example, in Facebook's section entitled '*what kinds of information do we collect',* they focused on the collection process e.g. *'things that you provide'* or *'the information we receive from our partners'*. As discussed in Section 5.6.5, this approach of focusing on sources of personal data in informing individuals does have some benefits for transparency, but overall if this were to be endorsed as the categories of personal data that need to be provided under the obligation to inform it would not achieve all the benefits of increasing transparency in practice.

A third theme was that when the privacy policies included examples of the specific personal

data that they collect, they were generally discussed separately to the purpose for which they would be used, with examples of specific information linked to a specific purpose used sparingly. This practice also impedes the transparency of processing, as it means that which data is being used for which purpose is unclear. This prevents individuals from understanding exactly what organisations are doing with their data, as without knowing what information services are collecting and which purpose will be applied to it and thus whether they intend to infer more information from this, it is impossible for individuals to begin to understand everything that an organisation collects, processes and 'knows' about them. Again, this process affects the ability of individuals to truly assess the risk involved in processing, it also means that the categories of personal data are not being used to contextualise other information provided in the privacy policy.

Fourth, the SNS referred to the information collected and processed using different levels of granularity. For example, Facebook's descriptions of what they collect ranged from 'content' to 'frequency and duration of your activities', through to their most granular description of the 'date a file was created'. This practice both helps and hinders transparency. Referring to 'what is collected' using wide descriptions like *'information about your use of financial products and services that we offer'* makes it almost impossible for individuals to understand exactly what is being processed. Whilst more granular descriptions, like 'photos' or 'IP address' can increase the transparency, they often do not (alone) convey all the information that can be derived or inferred from them (such as location) depending on the purpose applied and the metadata included. Thus, whilst referring to what is collected using different granularities may be beneficial and could allow for the attachment of different rights and obligations, if services are not granular enough, processing is not transparent and even where services appear to be very granular, without other information, transparency can sometimes still be impeded.

5.6.2 Conclusion to RSQ4

Thus, in answering RSQ4, 'How are SNS categorising personal data and informing individuals of this in practice and does this achieve the benefits for transparency?', four different approaches were present in all of the privacy policies of SNS, none of which alone achieves the benefits for transparency that an appropriate categorisation of personal data would be able to do in practice. The findings do indicate that it may be a combination of

approaches to categorising personal data that is required in practice or that further information will need to sit behind the 'categories' of personal data to achieve transparency in practice.

5.7 Conclusion

As discussed, the preliminary study of this thesis discussed in Chapter 3 found that all six of the SNS investigated provided information on the specific personal data that they processed, yet this was not a recommendation made by the UK Regulator in their guidance on what to include in a compliant privacy policy. This finding was unusual, because it seemed logical that for the processing of personal data to be transparent, organisations (especially those providing online services) would need to provide information about the specific personal data that they process. This Chapter discussed the third investigation undertaken as part of this thesis, the aim of which was to understand whether providing information about the personal data being processed has benefits for transparency and whether there are any approaches that have organically arisen in practice.

In answering RSQ1 of this chapter, the investigation found that this lack of clarity reduces the transparency of personal data processing, because there are many benefits for transparency by categorising personal data and informing individuals of which ones of these are processed. This suggests that European Regulators should confirm that there is a requirement for controllers to always specify the categories of personal data that they are processing and amend their guidance to reflect this. However, to do so they will also need to provide guidance on how to categorise 'non-special category personal data'.

In answering RSQ2, the investigation found that despite the lack of direction from the law on what a category of non-special category personal data is, different parties have interpreted their own approaches to categorising personal data, but in different ways. This means that a consistent approach to categorising non-special category personal data has not organically emerged under the obligation to inform. Yet, in answering RSQ3, the investigation found that none of these approaches are sufficient in practice at making the processing of personal data transparent to a level that equates the information available to subjects to that possessed by data controllers and achieves the benefits of transparency which an appropriate categorisation can create.

As it was the behaviour of the SNS that were investigated in the preliminary study of discussing the specific personal data they processed, RSQ4 looked to understand how SNS were doing this in practice, as this may inform the creation of a new categorisation of non-special category personal data. However, it found that the approaches of the SNS also did not achieve the benefits for transparency of an appropriate categorisation of personal data. This finding increases the importance for regulators to confirm how to categorise 'non-special category personal data', because even if they do not confirm that they must **always** be provided, it will be a requirement in some situations to do so, and controllers do not have the guidance they need to understand how to do this. This lack of clarity reduces the transparency of personal data processing and the answers to RSQ3 and RSQ4 mean that there is not an approach to categorising non-special category personal data that can be readily adopted in practice by the regulator.

These findings suggest that this is an area where improvements can be made to the way data controllers are transparent about their personal data processing in accordance with the goal of this thesis. In particular, they suggest that a new categorisation of non-special category personal data is required in order to increase the transparency of processing.

Thus, the final contribution of this thesis is to suggest a new way to approach the categorisation of personal data which increases the transparency of personal data processing, and which a regulator can adopt under the obligation to inform. This is discussed further in the next chapter.

Chapter 6 A New Approach

6.1 Introduction

6.1.1 General Overview

As discussed in Chapter 4 there is a lack of clarity in the law on when controllers need to inform individuals of the personal data that they process. Chapter 5 highlighted that this is despite the importance of individuals being informed about the personal data that is processed about them for transparency. It also highlighted that a consistent approach to categorising personal data has not occurred and none of the approaches discovered in practice would suffice. These findings suggest that there is a need for a new approach to categorising personal data. A new approach needs to be one that increases the transparency of personal data processing under the obligation to inform, by achieving the benefits of informing individuals of the categories of personal data that are processed about them, but also that avoids the issues with the current approaches to categorising personal data.

This chapter discusses the proposed approach to categorising personal data that this thesis proposes by answering RQ4:

RQ4: Could the DIKW model be adapted and used to increase the transparency of personal data processing in relation to the obligation to inform under the European Union Data Protection Framework?

RQ4 was developed understand whether a new categorisation of personal data, based on a model from computer science, can be used to increase the transparency of personal data processing under the obligation to inform.

This chapter is structured as follows. It begins by reviewing the literature on different approaches to constructing categories in general. It then introduces the Data, Information, Knowledge and Wisdom (DIKW) Hierarchy, a proven model in computer science which forms the basis for the new approach to categorisation personal data. The chapter discusses why using knowledge from computer science is important here, what the model is, different versions of it and the criticisms it has received. It then confirms the version of

the hierarchy that is being used in this thesis and why the model was chosen. This chapter then goes on to discuss previous work that has linked this model to data protection law and why despite this, and other suggested models of personal data, a new approach is still required. The chapter then discusses the model presented in this thesis, including how it will work in practice and the potential impact it can have. It describes how it can be used to create a legally certain requirement for data controllers to inform individuals about the categories of personal data that they process under the obligation to inform and how it would work within the current framework. Finally, the chapter validates the model using the requirements analysis and case study methodology to test that the model is fit for purpose, this method is discussed further in the next section.

6.1.2 Method

As discussed, in proposing the new approach to categorisation, this needs to be validated, to demonstrate that it has the potential to increase the transparency of personal data processing under the obligation to inform. Various methods could be used to validate the approach, for example an expert review where data privacy experts review the potential performance of the categorisation in practice could be used or a laboratory experiment where individuals are presented with a privacy policy which uses this categorisation to compare whether they understand more about the processing than the privacy policies that SNS are currently using could also be used. This thesis used a combination of requirements analysis and the use of two case studies to validate the new approach. These methods were chosen for a number of reasons. First, requirements analysis is generally used to validate software systems and therefore as this thesis has focused mainly on software systems in the form of SNS, it was believed that it would prove useful as a method in validating a system which is aiming to increase the understanding individuals have of software systems. Second, given the contributions made and work already undertaken as part of this thesis, there was not the time left to use other methods, however as discussed further in Chapter 7, these methods could be used for future validation of the proposed categorisation. Thirdly, as this is an interdisciplinary thesis it seemed appropriate to use a method from computer science validate the approach. Fourthly, by using case studies we could explain how the new approach could be used as well as testing it for validation. These methods are discussed further in Section 6.6.

6.2 Requirements for Categorisation

To create an appropriate categorisation of personal data, which increases the transparency of personal data processing in practice, it was important to understand more about how to construct a categorisation of a phenomena in general. Category is defined as a *'type, or group of things having some features that are the same'* (Oxford Mini English Dictionary, 2011). Categorisation is the process in which these things are classified into a set of categories, usually based on their similarities. Constructing meaningful classifications of objects is an omnipresent scientific problem (Michalski and Stepp, 1983) and the study of categories and categorisation is relevant to various disciplines.

The ability to categorise is one of the most basic human cognitive processes (Corrigan, Eckman and Noonan, 1989) and helps us to comprehend the infinite number of unique objects in the world (Markman, 1983). Categorising allows us to recognise familiar information, assimilate new information and to distinguish among objects and properties (Bornstein, 1984), it also requires some form of high-level cognition, to track causal patterns and seek explanatory coherence (Keil, 2005). As discussed in Section 6.2.2, advancements in Machine Learning mean that the ability to categorise is no longer only that of humans, but also of machines and computer systems and academic work on human categorisation in Psychology and Philosophy is being built upon by researchers in Artificial Intelligence to simulate these processes in computer systems. Another field which provides guidance on how to approach categorising knowledge realms is Taxonomy, the science of classification is also an area.

Categorisations are generally broken down into three levels, superordinate categories, basic level categories and subordinate categories (Mervis and Crisafi, 1982). Subordinate categories are the lowest level and have a low degree of generality and class inclusion which means that they are clearly identifiable and highly detailed. Basic level categories are the next level up and include the members of the subordinate categories, they provide general recognisable gestalts and more recognisable common features. Superordinate categories are found at the top of a categorisation and include the basic level categories, they display a high degree of generality (Mervis and Crisafi, 1982). If we take the example of 'car' as a basic level category, then 'vehicle' could be the superordinate category and 'Ford Mustang' could be the subordinate.

6.2.1 Theories of human categorisation

Despite consensus that categorisation is a fundamental human ability, its exact nature is controversial (Corrigan, Eckman and Noonan, 1989) but four key theories of how humans construct categories have emerged.

The first of these is the classical view, which was originated by Plato and systematised by Aristotle. It purports that categories are mutually exclusive and clearly defined, with necessary and sufficient conditions which all members share (Corrigan, Eckman and Noonan, 1989). Those of the classical persuasion believe that categories are enduringly real and truly 'out there' in the real world, and in our heads (Smith, 2005). However, studies have revealed that the categories humans produce are malleable and adapted to fit the idiosyncrasies of the moment (Barsalou, 1993a, 1993b; Bransford & Johnson, 1972; Malt, 1994). Studies have shown that humans do not collectively categorise things in the same way, because our histories, cultures and the context of the moment affect how we construct categories. Thus, although the classical view dominated early work on categorisation, its theoretical problems have been evident for some time (Smith, 2005). There is little evidence of a successful version of the theory in practice and studies show that humans struggle to define properties of even everyday categories (Rosch and Mervis, 1975; Smith and Medin, 1981; Barsalou, 2005; Wittgenstein, 1953). Data also contradicts the idea of necessary and sufficient features in human categories (Smith, 2005) because such features would require all members to be equally good. However, Rosch (1973) found that people judge some members as 'better' than others e.g. a robin is judged to be a better example of a bird than a penguin.

These issues led to the second and third theories of the way humans form categories and the emergence of probabilistic theories in the 1970s, which attempted to explain human category judgements by tying them to general cognitive processes of memory, attention, association and generalisation by similarity (Smith and Medin, 1981). Two key theories emerged, Prototype Theory and Exemplar Theory, which although alike in the processes they assume, differ slightly in their interpretations (Smith, 2005). However, it has been argued that the theories may not formally be distinguishable (Estes, 1986).

Prototype Theory posits that categories have lists of characteristics, instead of necessary and sufficient features (Smith and Medin, 1981). It asserts that humans create a prototype

by averaging characteristics, to produce a typical example of a category (Posner, 1969; Rosch, 1973). Thus, a Robin is a more typical member of the category 'birds' than a penguin, because it contains more characteristic attributes and is thus closer to the prototype (Rosch, 1973).

Whereas, Exemplar Theory purports that humans categorise based on examples we recall at the point of categorisation instead of a prototype developed over time. It asserts that humans remember instances of categories and associated properties, to which general processes of memory retrieval, association and generalization by similarity give rise to the in-task category judgements (e.g. Nosofsky, 1984; Smith & Medin, 1981). The theory posits that only exemplars are stored, opposed to summary representations of categories and that abstractions result from scanning and summarising these exemplars (Barsalou, 2005).

Despite the success of probabilistic theories in modelling judgements humans make of common categories (McRae, Cree, Westmacott & de Sa, 1999; Viggliocco, Vinson, Lewis & Garrett, 2004), both theories struggle to account for exactly how people reason about categories (Smith, 2005). Rips (1989) and Keil (1994) both found that people make category judgements shaped more by a defining feature than an overall probabilistic theory. For example, people will maintain that an organism that has no properties 'like a bird' other than bird DNA and bird parents is nonetheless, still a bird.

Deficiencies in probabilistic theories led to the fourth key theory of how humans form categories and prompted researchers to study people's beliefs and reasoning about exactly which characteristics are relevant to category membership (Smith, 2005). This led to assertions that people possess intuitive theories of different domains that are analogous to scientific theories (Gelman, 2003). Findings indicated that some characteristics have more causal force than others, leading to research into studying intuitive theories and their development. This created new fields of study about categories, including conceptual combination, induction and causal reasoning (Keil, 2003; Smith, 2005). It also led to interesting insights about how reasoning can differ between domains, including differences between biological versus non-biological domains (Gelman, 2003).

Despite these developments, intuitive theories have been criticised for a lack of sufficient definition and consensus over them. They also only assess a subset of the data that is

traditionally viewed as the realm of a categorisation theory, as certain phenomena like induction, conceptual change, conceptual combination and causal relatedness are singled out as theoretically more important than phenomena concerning the recognition of instances. Intuitive theories do not explain why robins are judged to be psychologically better birds than penguins (Smith, 2005) and the fact that people readily make these judgements is seen as irrelevant (Armstrong, Gleitman & Gleitman, 1983). This means that there is not one conclusive theory that explains how humans construct categories.

6.2.2 Computers and Categorisation

As discussed, categorisation is now not only a human process, but also one of machines and computer systems. A key area where categorisation has been researched in relation to machines and systems is 'conceptual clustering' as a machine learning task (Michalski, 1980).

Machine Learning is a subfield of Artificial Intelligence (AI) in Computer Science, focused on machines 'learning' from experience and improving performance by automating knowledge acquisition and refinement (Fisher, 1987). Although Machine Learning is not concerned with building automated imitation of intelligent behaviour (like traditional AI), the ability to turn experience into expertise, or to detect patterns in complex data is a cornerstone of human intelligence. Machine Learning uses the abilities of computers to compliment human intelligence, often performing tasks beyond human capabilities (Shalev-Shwartz and Ben-David, 2014).

Whilst humans have inductive bias and common sense built in, which bias the generalisations and approaches we take, and help us filter out random, meaningless learning, computers do not. Prior to the development of Machine Learning, any learning task exported to a machine would generally require well-defined principles and concepts (Alpaydin, 2016). Machine Learning is developing the principles and algorithms which imitate these abilities and allow systems to move beyond learning by memorisation (Shalev-Shwartz and Ben-David, 2014).

Shalev-Shwartz and Ben-David (2014) highlighted two aspects of a problem that will call for the use of Machine Learning: the problem's complexity, which places it beyond human

capability; and the need for adaptivity. Complexity may be caused by human introspection, which prevents us from being able to provide explicit explanations of how tasks should be executed e.g. describing driving or image recognition. It may also come from the fact that the size of the datasets involved are too large and complex for humans to comprehend. Adaptivity may also require Machine Learning techniques, as many tasks change over time and from one user to another. Because programmed tools written by humans are very rigid, it is more effective for the machine to develop its own algorithm (Alpaydin, 2016).

Machine Learning has branched into several subfields, each dealing with different types of learning tasks (Gollapudi, 2016). One type of learning model which has made developments in relation to categorisation is the model of unsupervised learning.

In supervised learning tasks, the learning algorithm receives data with some extra information attached to it e.g. receiving emails which are tagged as spam/not spam when trying to get a learning algorithm to learn how to detect spam emails (Shalev-Shwartz and Ben-David, 2014). This additional information allows the learning algorithm to create a rule for labelling newly arriving email messages, based on the examples and information it has. Whereas, in unsupervised learning tasks all the learner gets is a large body of data with no direction on what it is trying to predict. The learner's task is to detect patterns and anomalies and segment datasets by some shared attributes, simplifying them in the process (Shalev-Shwartz and Ben-David, 2014).

There is also an intermediate learning setting, where whilst the training examples contain more information than test examples, the learner is required to discover additional information about the test examples. Shalev-Shwartz and Ben-David (2014) give the example of trying to learn a value function that describes for each setting of a chess board, the degree by which white's position is better than black's when the only information available to the learner at training time is positions occurred throughout the game, labelled by who eventually won that game.

Thus, as an unsupervised learning task, clustering data into similar subsets of data can be thought of as categorisation by machines. Shalev-Shwartz and Ben-David (2014) describe clustering as the *'task of grouping a set of objects such that similar objects end up in the same group and dissimilar objects are separated into different groups'*. Clustering analysis

is one of the most widely used techniques for exploratory data analysis, across all disciplines. It allows researchers to understand a dataset, by identifying meaningful groups among the data points and does not start with specific labels in mind (Gollapudi, 2016). However, as with human categorisation, because there are no inherent labels to direct prediction, there is no clear success evaluation procedure. Even with full knowledge of the underlying data, the 'correct' clustering will not be clear, as it depends on the reason for clustering, which is not always apparent in unsupervised learning. Thus, whilst clustering allows grouping of the data in a meaningful way, it often still requires interpretation from a human.

Even within clustering there are various 'types' of clustering algorithms (Xu, D., & Tian, Y., 2015). Whilst traditional clustering methods can group the data in different ways, they are unable to describe the generated clusters, making the clusters difficult to interpret (Boubacar and Zhendong, 2014). They also arrange objects solely on the basis of similarity within the dataset without taking account external concepts, which can be useful in characterising object configurations (Michalski and Stepp, 1983).

Conceptual Clustering is a type of clustering algorithm which accepts object descriptions and produces a classification system (Fisher, 1987). Michalski and Stepp (1983) describe the process of conceptual clustering as:

Given:

- A set of abstracts (physical or abstract)
- A set of attributes to be used to categorise the objects
- A body of background knowledge which includes the problem constraints, properties of attributes, and a criterion for evaluating the quality of constructed classifications.

Find:

- A hierarchy of object classes, in which each class is described by a single conjunctive concept. Subclasses that are descendants of any parent class should have logically disjoint descriptions and optimize an assumed criterion (a clustering quality criterion).

These hierarchies can be built in a top-down or bottom-up manner. A top-down approach will divide the objects into a small number of classes, each of which may be divided into subclasses. This process is iterated until a termination condition is met. In bottom-up clustering, each object is initially considered to be in its own class, they are then grouped together, and the resulting groups are then brought together into superclasses, until the top level is reached (Kaufman, 2012).

Michalski and Stepp (1983) acknowledge that is no universal answer to the difficult problem of how to judge the quality of a clustering but suggest two major criteria. First, that descriptions of clusters are "simple", so that it is easy to assign objects to classes and to differentiate between classes. Second, is that class descriptions should fit the data, which may indicate the need for complex descriptions. Thus, the demands for simplicity and good fit conflict and the solution is to find a balance between the two. Fisher (1987) adds that 'good' concepts should maximise the number of predictions that can be made about objects of the environment.

6.2.3 Taxonomy

Another area of science which provides guidance on how to classify objects is Taxonomy, the science of classification. The words taxonomy, classification (Bowker and Star, 1999), typology (Bailey, 1994; Doty and Glick, 1994) and framework (Schwarz et. al, 2007) are often used interchangeably, although 'taxonomy' has been found to be the most commonly used term in a literature survey on the topic (Nickerson, Varshney and Muntermann, 2013). Ontologies are also often referred to, although this is a distinct concept.

A taxonomy is a set of dimensions consisting of mutually exclusive and collectively exhaustive characteristics, such that objects under consideration will fall within one, and only one, dimension (Nickerson, Varshney and Muntermann, 2013). In that sense it is similar to classical categorisation theory. Taxonomies help the understanding and analysis of complex domains, by providing structure to knowledge within a field. This enables the study of and hypothesis about relationships among concepts (Nickerson, Varshney and Muntermann, 2013).

Whilst useful, developing a taxonomy is a complex process. As a discipline, Biology provides some guidance and perhaps the most famous taxonomy, the traditional Linnaean taxonomy (Simpson, 1961). Once a taxonomy is constructed, phenetics and cladistics can be used to determine where an object falls within it (Fleishman, Quaintance, and Broedling, 1984). Phenetics (or numerical taxonomy) classifies solely on the basis of similarity. Characteristics are identified and then statistical techniques are used to cluster organisms into similar groups (Sokal and Sneath, 1963). Whereas, cladistics examines the evolutionary relationships among organisms and groups based on evolutionary heritage (Eldredge and Cracraft, 1980). Thus, two organisms may be closely related in a cladistic taxonomy yet a phenetic analysis would put them into different categories. Taxonomy development has also been well studied in the social sciences (Bailey, 1984).

In a review of methods and approaches to developing taxonomies, Nickerson, Varshney and Muntermann (2013) classified such methods into three main types:

- Inductive: This approach is similar to phenetics in Biology and determines dimensions and characteristics by observing empirical cases and using statistical techniques, such as cluster analysis (or less rigorous techniques) to find similar groupings.
- Deductive: This method first derives a taxonomy from theory or conceptualization opposed to from empirical cases (similar to cladistics in biology). Analysis of empirical cases may then follow, to evaluate and modify the taxonomy.
- Intuitive: This approach is essentially *ad hoc*. A taxonomy is proposed based on what makes sense from a researcher's perception, of the objects under examination. There is no explicit method or conceptual, theoretical, or empirical foundation in this approach. Nickerson, Varshney and Muntermann (2013) found that although authors using this approach reviewed the literature in their problem area, their taxonomy was often not based on this.

Other approaches were found, which fell outside of these categories including morphological analysis and the use of existing taxonomies. Whilst Nickerson, Varshney and Muntermann (2013) found that there were various different approaches to taxonomy development, they could not find a definitive rigorous method and so they present the method depicted in Figure 11.

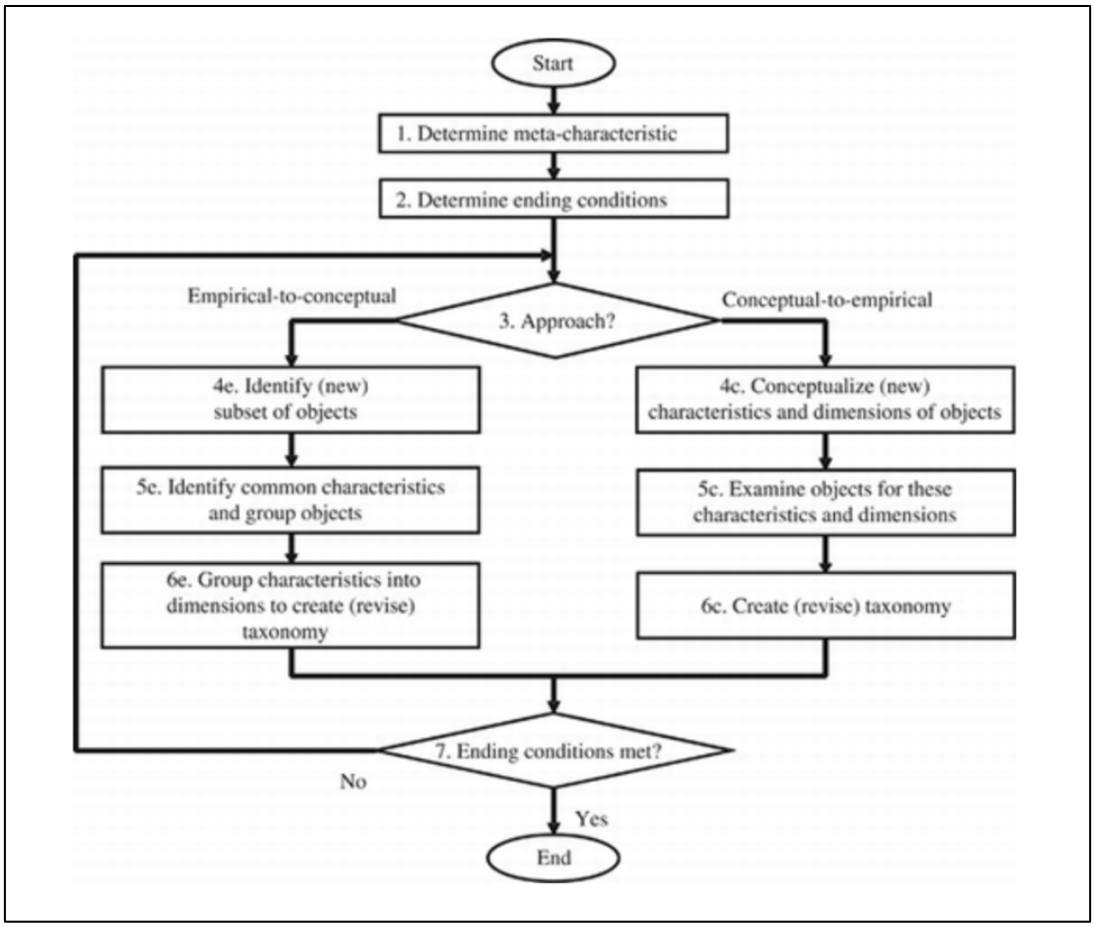


Figure 11 Nickerson, Varshney and Muntermann's (2013) Method of Categorisation

1. The first step is to define a meta-characteristic which dictates the choice of the characteristics below it. To do this, define the users of the taxonomy and both the purpose and expected use of it. An example of a meta-characteristic is the '*high-level interaction between the application user and the application*' in developing a taxonomy of mobile applications.

2. The second step is to specify the objective and subjective ending conditions, which determine when to terminate the iterative process of classification. Tables 15 and 16 show examples of objective and subjective ending conditions that can be selected.

Table 15 Nickerson, Varshney and Muntermann's (2013) Objective Ending Conditions

Objective Ending Condition	**Comments**
All objects or a representative sample of objects have been examined	If all objects have not been examined, then the additional objects need to be studied
No object was merged with a similar object or split into multiple objects in the last iteration	If objects were merged or split, then we need to examine the impact of these changes and determine if changes need to be made in the dimensions or characteristics
At least one object is classified under every characteristics of every dimension	If at least one object is not found under a characteristic, then the taxonomy has a 'null' characteristic. We must either identify an object with the characteristic or remove the characteristic from the taxonomy
No new dimensions or characteristics were added in the last iteration	If new dimensions were found, then more characteristics of the dimensions may be identified. If new characteristics were found, then more dimensions may be identified that include these characteristics
No dimensions or characteristics were merged or split in the last iteration	If dimensions or characteristics were merged or split, then we need to examine the impact of these changes and determine if other dimensions or characteristics need to be merged or split
Every dimension is unique and not repeated (i.e., there is no dimension duplication)	If dimensions are not unique, then there is redundancy/duplication among dimensions that needs to be eliminated
Every characteristic is unique within its dimension (i.e., there is no characteristic duplication within a dimension)	If characteristics within a dimension are not unique, then there is redundancy/duplication in characteristics that needs to be eliminated. (This condition follows from mutual exclusivity of characteristics.)
Each cell (combination of characteristics) is unique and is not repeated (i.e., there is no cell duplication)	If cells are not unique, then there is redundancy/duplication in cells that needs to be eliminated

Table 16 Nickerson, Varshney and Muntermann's (2013) Subjective Ending Conditions

Subjective ending condition	***Questions***
Concise	Does the number of dimensions allow the taxonomy to be meaningful without being unwieldy or overwhelming?
Robust	Do the dimensions and characteristics provide for differentiation among objects sufficient to be of interest? Given the characteristics of sample objects, what can we say about the objects?
Comprehensive	Can all objects or a (random) sample of objects within the domain of interest be classified? Are all dimensions of the objects of interest identified?
Extendible	Can a new dimension or a new characteristic of an existing dimension be easily added?
Explanatory	What do the dimensions and characteristic explain about an object? Do the explanations go beyond just descriptions of objects?

3. In the third step, researchers decide whether to take an empirical or conceptual approach, based on available data and their knowledge about the domain. If little data are available but the researcher has significant knowledge of the domain, the conceptual-to-empirical approach is advised. Whereas, if the researcher has little understanding but significant data about the objects, starting with the empirical-to-conceptual approach is advised.

4e. In the empirical-to-conceptual approach, the researcher identifies a subset of objects to classify. This could be a random sample, systematic sample, convenience sample, or some other type of sample.

5e. Common characteristics of these objects, as logical consequences of the meta-characteristic are identified using the knowledge and intuition of the researcher. These must discriminate among objects as a characteristic that has the same value for all or nearly all objects is not useful, even if it does follow from the meta-characteristic.

6e. Once a set of characteristics has been identified, objects are grouped formally using statistical techniques or informally, using a manual or graphical process. The resulting groups form the initial dimensions of the taxonomy. 'Conceptual labels' for sets of related characteristics are created and each dimension contains characteristics that are mutually exclusive and collectively exhaustive. Nickerson, Varshney and Muntermann (2013) found that most taxonomies reviewed presented four or fewer dimensions, but a few gave taxonomies with more than ten. There is no agreement on what represents an appropriate number of dimensions.

4c. In the conceptual to empirical approach, the researcher conceptualises the dimensions of the taxonomy without examining objects, purely based on their notions about how objects are similar and dissimilar. As a deductive process, little guidance can be given but the researcher uses their knowledge of existing foundations, experience, and judgment to deduce what they think are relevant dimensions. Whether its characteristics follow from the meta-characteristic is the test of the appropriateness of a dimension.

5c. The researcher examines objects for these dimensions and characteristics. Are there objects that have each of the characteristics in each dimension? If not, then the dimension may not be appropriate and must be reviewed.

6c. The result of this process is an initial taxonomy, where objects fit into mutually exclusive and clearly defined dimensions, based on a conceptual-to-empirical approach.

7. In the final step, at the end of either approach, the researcher uses their insight, experience and skill to ask if the objective and subjective ending conditions have been met. The process is repeated until they are and each time the researcher can choose whether to change the approach taken. In the empirical-to-conceptual case, the researcher begins with new objects to determine whether the existing characteristics and dimensions are sufficient. In the conceptual-to-empirical case, the researcher begins reviewing the taxonomy and then examines empirical cases to determine usefulness in classifying objects.

Upon completion of the method, the taxonomy must be evaluated for usefulness. Taxonomies do not need to be perfect, but they must be useful. However, determining sufficient conditions for usefulness is difficult and evaluation may come down to seeing if it is used. Nickerson, Varshney and Muntermann (2013) recommend speculating on potential use of the taxonomy by asking whether the user's purpose can be satisfied with the taxonomy? This could be done by asking users about their potential use of the taxonomy or evaluating what the taxonomy tells the users in relation to the purpose of it.

6.2.4 Implications for This Thesis

There are a number of things that can be learnt from these approaches when looking to propose a new categorisation of personal data. Indeed, the issues with the way humans conduct categories discussed in Section 6.2.1 increases the need for a robust and consistent categorisation of personal data to be created. The fact that the categories humans create are malleable and adapted to fit the idiosyncrasies of the moment means that it is not surprising that different groups have interpreted ‘categories of personal data’ in different ways and will likely continue to do so without guidance on the correct categorisation.

Whilst it was not necessarily a good representation of how humans create categories, the type of categorisation required for personal data will be the classic theory and taxonomy, where categories are mutually exclusive and clearly defined with necessary and sufficient features that all members share. The categories will need to be mutually exclusive because different rights and obligations will be attached to the categories and so it must be clear which one applies. They will need to be clearly defined so that controllers and data subjects can understand which ones are processed and thus which rights and obligations apply. In

terms of necessary and sufficient features, whilst these are the aim, as Prototype Theory shows characteristics may be enough to indicate membership of a category.

Work on conceptual clustering shows how when attributes of categories have been identified, machine learning could be used to categorise data. It also highlights the fact that where there are no inherent labels to direct prediction, there is no clear success evaluation procedure and therefore one will need to be defined to confirm that personal data is classified correctly within the categorisation. It also shows that given the broad and ever-expanding definition of personal data, a top-down approach which divides the objects into a small number of classes, each of which may be divided into subclasses will be required.

Finally, work on Taxonomy frames that currently groups are taking 'intuitive approaches' to categorising personal data, by categorising based on what makes sense from a researcher's perception, of the objects under examination. There is no explicit method or conceptual, theoretical, or empirical foundation in this approach. To counter this, this thesis uses a combination of taking a deductive approach (where a categorisation of personal data is derived from theory or conceptualization opposed to from empirical cases) and use of existing taxonomies to propose a categorisation of personal data with such a foundation. Nickerson, Varshney and Muntermann's (2013) process for developing a taxonomy is then used in Section 6.5 to create the new categorisation, taking a conceptual-to-empirical approach.

6.3 The DIKW Hierarchy

As discussed in Chapter 2, there are various criticisms of the current definition of 'personal data', which can lead to a wide interpretation of what should be protected under the framework despite previous attempts in some countries to limit the scope of its definition. Schwartz and Solove (2012:1817) have termed this an 'expansionist' approach and Zwenne (2013) has highlighted the '*serious risk that this will lead to a dilution of data protection or privacy law, in the sense that the law will apply to everything and nothing, making it a law without meaning*'. However, whilst a wide definition of personal data is problematic, a narrow definition could leave individuals open to privacy-related harms, by not protecting information that warrants it, as the Article 29 Working Party have asserted (2007:5).

With the data which is considered to be 'personal' only set to widen with increasing technological progress, the best starting point for the law is to consider that: no personal data can be permanently anonymised and that all information can become 'personal data' by being linked to someone or being processed to reveal further personal data. This thesis argues that the focus of the framework should not be on narrowing the definition of personal data, nor abandoning the concept of personal data as Ohm (2010) suggests, but on creating more structure within the definition of personal data so that appropriate rights and obligations can be tailored accordingly. As discussed in Chapters 2 and 4, within this wide definition of 'personal data', the only further structure that the law provides currently is 'categories' of personal data and the law only specifies categories of 'special category personal data'. This thesis seeks to provide the equivalent categorization of 'non-special category personal data'.

Given the inextricable link between advances in technology and the expanding definition of personal data, when looking to categorise personal data, it is important to understand if there are any models in computer science which could be utilized for this purpose. In doing so, the Data, Information, Knowledge, Wisdom (DIKW) was discovered.

6.3.1 Introduction to the hierarchy

The Data, Information, Knowledge and Wisdom Model (DIKW Model) is one of the fundamental concepts in information science, computer science, knowledge management science, psychology, cognitive science, philosophy, and many others (Rowley, 2007). The model is also known variously as the 'Knowledge Hierarchy', the 'Information Hierarchy' and the 'Knowledge Pyramid' (Rowley, 2007).

The model is used both to define data, information, knowledge and wisdom, but also to contextualise the concepts with respect to one another, in order to identify how one concept can be transformed into the entity above it e.g. data to information, information into knowledge etc. (Rowley, 2007). It is widely accepted in the knowledge management circles as a way to represent the different levels of what we see and what we know (Cleveland 1982; Zeleny 1987).

6.3.2 The origins of the hierarchy

Many agree that the first recorded use of the Hierarchy was in a poem by T.S. Eliot in 1934 in which Eliot wrote:

> *"…Where is the Life we have lost in living?*
> *Where is the wisdom we have lost in knowledge*
> *Where is the knowledge we have lost in the information"*.
> **T.S. Eliot (1934)**

However, in more recent texts, authors often point to Ackoff (1989) as the source, although Cleveland (1982) also makes an early mention of the Hierarchy in information science literature. Ackoff made reference to the model in his acceptance address for the presidency of the International Society for General Systems Research in 1989. This speech was then printed in the article 'From Data to Wisdom', which does not cite any earlier sources of the hierarchy. Ackoff (1989) estimated that "on average about forty percent of the human mind consists of data, thirty percent information, twenty percent knowledge, ten percent understanding, and virtually no wisdom" (Ackoff, 1989:3). This phraseology allows views his model as a pyramid, which it has been likened to ever since (Rowley, 2007).

6.3.3 Different versions of the hierarchy

The exact origin of the hierarchy is not of supreme importance, because, despite the model being considered a fundamental model in both knowledge management and information system literatures, there are many interpretations and reinterpretations of the model. The Data Information Knowledge Wisdom model consists of four levels, but other variations have included more or less. For example, the five-level data, information, knowledge, understanding, wisdom hierarchy (Ackoff, 1989) and the data-information-knowledge-wisdom-enlightenment five-level hierarchy (Zeleny, 1987).

Earlier instantiations also prohibited backwards movement (e.g., acquiring information from knowledge), however, modern research has questioned this assumption as possessing knowledge can allow a user to derive information or even data (Stenmark, 2002). Indeed, Tuomi (1999) strongly argues for a reversal of the DIKW model because, information is created only after there is knowledge and data emerges as a by-product of cognitive artefacts.

Ackoff's proposed model had the following levels: data, information, knowledge, understanding and wisdom. More recent commentators have disputed that understanding is a separate level, and instead is present to some extent at each level (Bellinger, et al. 2004). At around the same time as Ackoff, Zeleny (1987) proposed an additional level, of enlightenment at the top. Enlightenment is not only answering or understanding why (which Zeleny defines as wisdom), but going further and attaining the sense of truth, the sense of right and wrong, and having it socially accepted, respected and sanctioned.

Berčič and George (2009) proposed a data, information, knowledge, rules (DIKR) hierarchy in the context of relational database theory. This hierarchy consisted of data, information, knowledge, deduced knowledge and induced knowledge. Bernstein (2009) used the DIKW Hierarchy as a background and starting point for reflecting on the opposite of knowledge, looking at whether an analogous sequences of stages surrounds ignorance, a domain referred to by Bernstein as 'non-knowledge' which posits opposites of each of the terms in the hierarchy.

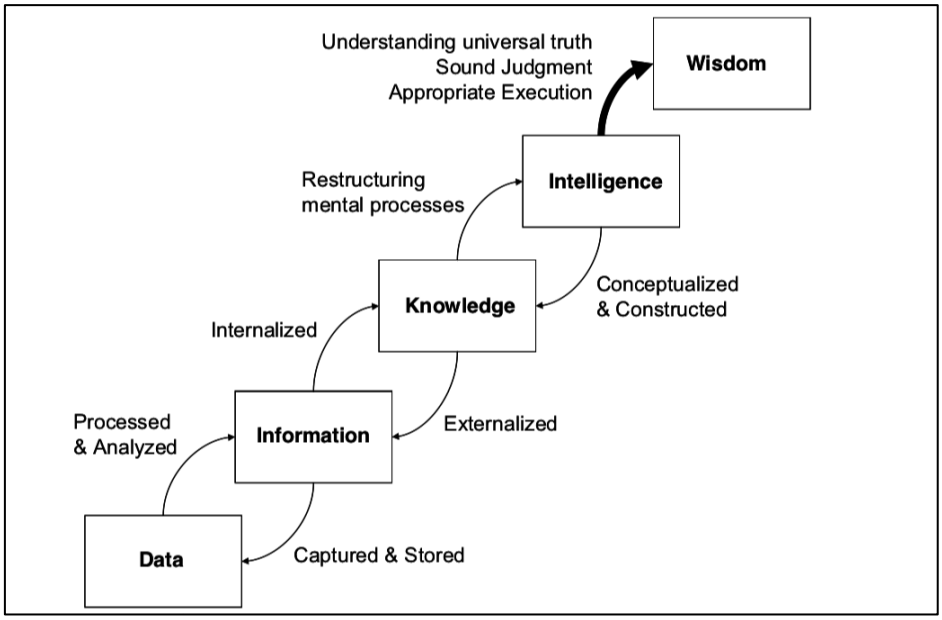


Figure 12 Liew's (2013) Version of the DIKW Hierarchy

Liew (2013) suggests a DIKW model which included 'intelligence' (DIKIW) as depicted in Figure 12 since intelligence has inseparable relationships with knowledge and wisdom. there must be intelligence to turn knowledge into wisdom.

Yao (2019) presented, in accordance with the DIKW Hierarchy, a perception-cognition-action (PCA) conceptual model that is applicable to studying intelligent data analytics, intelligent systems, and human understanding. Because of the issues with providing clear-cut definitions between all the levels as depicted in Figure 13. Yao combined the information and knowledge levels into a single level. The result is a tri-level data-information/knowledge-wisdom (D-I/K- W) hierarchy. In Yao's hierarchy, the Information/ knowledge level is a mediator that connects the data level and the wisdom level. Its function is to facilitate data-driven wisdom (i.e., data-driven decision-making). Yao suggests that a progressive view of the DIKW Hierarchy is a knowledge hierarchy with four types of knowledge. This would mean that data are a kind of weakest knowledge, information is a type of weak knowledge, and wisdom is a high kind of knowledge.

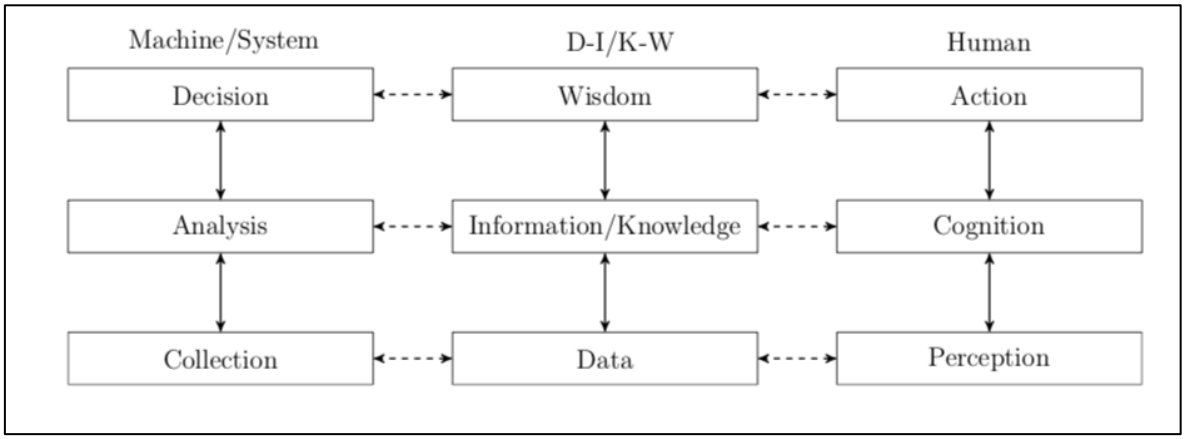


Figure 13 Yao's (2019) Version of the DIKW Hierarchy

In 2007, Rowley (2007) revisited the hierarchy by examining the articulation of it in a number of widely read textbooks. Rowley found that despite a number of articulations, typically all formulations share a common view that the key elements are data, information, knowledge and wisdom and that these entities were virtually always arranged in the same order. Indeed, Fricke (2009) points to Rowley on the fact that there are genuine and possibly substantive differences in view about DIKW and its properties, but that there is a core, and sufficient similarities, for a position to be extracted and scrutinized.

6.3.4 Data

Ackoff (1989) defined data as the product of observation, it is recorded, captured and stored symbols and signals that represent properties of objects, events and their environment. Symbols include words (text and/or verbal), numbers, diagrams, images (still

or video) that are the building blocks of communication. Signals include sensor and/or sensory readings of light, sound, smell, taste and touch Liew (2013). Ackoff (1989) states that information systems generate, store, retrieve and process data, thus, data might be an attribute of a data record (the contents of a field or item) in a relational database. A key element of data is that it has no use in and of itself because it is unorganised or unprocessed (Rowley, 2007). Data will only become usable once it is processed an in the relevant form, only then does it gain meaning and practical value. Data may be 'processed' by identifying relationships (Barlas, Ginart, & Dorrity, 2005) or by limiting it to only that which is relevant to the problem at hand (Carlisle, 2006).

6.3.5 Information

The definition of information has been considered by many disciplines, from communications theory, to library and information science, information systems, cognitive science and organisation science. This consideration has generated multiple perspectives on the essential nature of information (Rowley, 2007). Rowley (2007) found that both information systems and knowledge management literature defined information in terms of data, information is seen to be organised or structured data, Ackoff (1989) made it clear that the difference between data and information as functional, not structural. Structure is not just for structure's sake. It must create meaning for the recipient (Bierly III, Kessler & Christensen, 2000 citing Davis & Olson, 1985) for the result of structuring to be considered information instead of data. Data when connected to other data becomes information by making a relational connection, this is data with meaning obtained from the context (Berčlč and George, 2009). Thus, information can be considered a 'higher' form of data because it provides greater context and greater meaning (Johnson and Higgins, 2010), which makes it valuable, useful and relevant. For example, in relational database theory, information refers to data records, which are brought into context by their primary key (data with the identification of the entity it refers to) (Berčič and George, 2009).

In addition to organising data to create information, information can also be inferred from data, and does not need to be immediately available (Ackoff, 1989). Fricke (2009) uses the example of the enquiry of 'What is the average temperature for July?', where there may be individual temperatures explicitly recorded but perhaps not the average. Here, the average temperature can be calculated or inferred from the data about individual

temperatures. When structured or inferred appropriately, information is descriptive in nature (Batra, 2014) and answers questions that begin with such words as who, what, when and how many (Ackoff, 1989). Information is also of real or perceived value in current or prospective actions or decisions (Bierly III, Kessler & Christensen, (2000), although the action to which it will be put does not have to have been visualised for it to be considered information (Batra, 2014).

6.3.6 Knowledge

Plato first defined knowledge as 'justified true belief' (Mure and Ross, 1953), but the concept has been debated by many academics over the years, from Aristotle (Mure and Ross, 1953), to Descartes (1911), to Kant (1965) and Polanyi (1958). It is thought that the purpose of knowledge is to better our lives and in the context of business, to create or increase value for the enterprise and all its stakeholders (Liew, 2007).

Debates about the nature of knowledge have gained momentum over the last few decades with the development of knowledge management as a discipline (Rowley, 2007). In trying to find a common definition of knowledge, Rowley (2007) found that definitional statements on knowledge were often more complex than those for data or information. There is also a difference between the various definitions of knowledge in the world and the definition of knowledge used in the context of the DIKW Hierarchy. Thus, it is important when looking at definitions of knowledge to be mindful of the context in which they are being used.

In terms of how it is acquired, Ackoff (1989) posited that knowledge can be obtained by transmission from another who has it, or by extracting it from experience, which is done by acquainting oneself with facts, truths and principles, gained through study or investigation. The literature also contains different perspectives on whether masses of information alone constitute knowledge, or whether there must be an additional step that takes it beyond amassed information and turns it into knowledge.

BerčIč and George (2009) define knowledge as referring to a collection of data records such as a table, spreadsheet or entire database e.g. a fact database in an expert system. This falls down on the side of amassed information being enough to be considered knowledge,

although it is worth noting that this definition was in the context of relational database theory. To some extent this is in keeping with the definition of Ackoff (1989) who bases his definition of knowledge on collected masses of information. However, Ackoff also includes the entity of 'understanding' in his depiction of the DIKW hierarchy, which he defines as probabilistic or interpolative processes in order to answer "why"-questions that can be used to create new information or knowledge. For many academics, this separate notion of understanding forms part of their definitions of knowledge. Indeed, Ackoff himself acknowledges that understanding cannot exist on its own and requires knowledge and some kind of reasoning mechanism. Pearlson and Saunders (2004) assert that knowledge involves the synthesis of multiple sources of information over time. However, this is said in the context of the human mind and the word synthesis suggests something more than just the collection of this information.

Wan and Alagar (2014) are clear that they do not see stored information itself as knowledge. To them, the capacity to understand, analyse, and reason on the basis of stored information generates knowledge. Knowledge is the ability to make correct decisions ("what to do"), based on 'when to do", "why to do" and "how to do" (Wan and Alagar, 2014). Indeed, in the context of the hierarchy, there is a common emphasis in various definitions of knowledge as being something that is actionable. Ackoff (1989) defined knowledge as know-how, which makes it possible to transform information into instructions and Kakabadse et al. (2003) went a step further than this, suggesting that knowledge is not merely 'actionable' but 'actioned' information. They described knowledge as 'information put to productive use'.

Other perspectives define knowledge as the combination of data and information, but to which a combination of values, rules (Pearlson and Saunders, 2004) expert opinion, skills and experience are added, the result being a valuable asset for useful intent. The transformation of information into knowledge requires a quantum jump, a manual that describes how a car works is an example of information, but when used can be considered a knowledge base (Johnson and Higgins, 2010). Batra (2014) asserts that the superiority of knowledge over information is in its ability to guide action and aid decision-making, which can clearly be seen from the quality of questions that can be answered by knowledge, questions like how, how to and why (compared to information which answers questions like who, what, when and how many). Batra (2014) also distinguishes knowledge from

merely amassing information by describing it as more proactive, because the questions that get answered with knowledge suggest one or several courses of action, whereas whilst information may reduce uncertainty, it does not per se trigger decisions.

Fricke (2009) uses the example of temperature in a room to illustrate the relationship between data, information and knowledge. A room's temperature may turn from data into information when an agent queries 'what is the temperature?'. This information can then become instructions to turn an air conditioner on if the agent knows (a) the temperature the room needs to be, (b) that this temperature is above this and (c) that use of an air conditioner will lower the temperature. In this case, the information on room temperature transitions into knowledge on how to cool.

In addition to the differences between various definitions of knowledge in the world and the definition of knowledge used in the context of the DIKW Hierarchy, there are also differences between the definitions of 'knowledge' held in the human mind, and 'knowledge' that is held in information systems. Many academics disagree on the ability of information systems to contain knowledge, which may be reflected in the different definitions on whether massing information is enough to constitute knowledge.

Bocij et al. (2003) elaborated on the difference by separating knowledge into 'know-how' (tacit knowledge) and 'know-what' (explicit knowledge). Explicit knowledge (know-what) is knowledge that can be written down, transmitted and understood by a recipient e.g. the fact that London is in the United Kingdom. Whereas, tacit knowledge (know-how) is not always known explicitly, even by expert practitioners, and may be difficult or impossible to explicitly transfer to other people e.g. the ability to speak a language, ride a bike or play a musical instrument. Bocij et al. (2003) suggested that only explicit knowledge can be recorded in information systems and Rowley (2007) found that knowledge management texts were more likely than information systems texts to discuss the difference between explicit and tacit knowledge. Knowledge management texts described tacit knowledge as being embedded in the individual and explicit knowledge as residing in documents, databases and other recorded formats. Thus, Rowley's findings support Bocij et al.'s assertion because if tacit knowledge cannot be stored in information systems, it could be seen as pointless for texts in this discipline to devote much time to discussing such knowledge.

In addition to know-what and know-how, Liew (2007) offers another form of knowledge, 'Know-why'. Liew defines knowledge as the (1) cognition or recognition (know-what), (2) capacity to act (know-how), and (3) understanding (know-why)[9].

These different conceptions of knowledge can allow us to examine the extent to which knowledge can be held in information systems, a fact which may change over time with progress in machine learning and artificial intelligence. Indeed, Bocij et al. (2003) do suggest that the difference between explicit and tacit knowledge can be thought of as a continuum instead of a sharp distinction, which suggests that tacit knowledge may be able to be stored in an information system to some extent. Indeed, Fricke (2009) held that some know-how's might be articulated as procedural rules. These will normally be procedural rules e.g. 'if-then' rules such as knowing how to solve a quadratic equation, which can be written down and stored in a repository. Fricke does acknowledge that other know-hows do not seem to be of this kind.

Another important point about knowledge is that it is more subjective than information. The process of knowledge development is dependent on what an agent knows and their cognitive ability to assimilate know-how and know-why. This is true of both the human brain and of information systems. All knowledge is shaped by context because everyone relies on education and experience to make sense of things (Johnson and Higgins, 2010). Subjectivity is higher at the level of knowledge because the frame of reference of the decision maker guides the kind of questions to which answers are sought and choice of action once the answers to various questions start to become available (Batra, 2014).

6.3.7 Wisdom

In terms of the hierarchy, the definition of wisdom has been described as 'elusive' (Jashapara, 2005), When revisiting the notion of the DIKW Hierarchy, Rowley (2007) found that there was limited reference to wisdom and its definition, concluding that more work needs to be undertaken to develop an understanding of the applicability and relevance of the concept and how it is developed and managed. Indeed, whilst many ancient

[9] It is worth noting that Liew's definition of knowledge was in the context of the human mind.

philosophers have extensively discussed what wisdom means, modern literature on wisdom is limited to some extent (Sternberg, 2008).

Baltes and Staudinger (2000) asserted that wisdom, although difficult to achieve and to specify, is easily recognized when manifested. Hoppe et al., (2011) characterised wisdom as a fluctuating concept. They stated that it is not necessary to sharpen the meaning and indeed it may be counterproductive to try to sharpen the conceptual boundaries of vaguely bounded research objects while in operation. As long as objects are in flux, too, the corresponding concepts must remain in flux, too. Hoppe et al., (2011) concluded that the perspective the DIKW-model offers on the concept of wisdom seems likely on first sight – but shows some serious discrepancies with all common definitions of wisdom they found in the literature. Those definitions might differ in important points with the result that a holistic definition cannot be found, but they certainly agree on enough points that contradict its integration in the model as it is. Hoppe et al., (2011) therefore proposed to get back to the original version of the model, only containing data, information and knowledge as concepts building on each other.

Wan and Alagar (2014) discuss wisdom in the context of developing smart systems. They posited that the essential characteristics of wisdom are (1) knowing (understanding) facts (not just data), (2) understanding the procedures to extract information from data and facts and making it precise and valid, (3) having the knowledge to determine contexts that are relevant for initiating actions, (4) having the knowledge and skills to formally analyse the consequences of initiating actions, and (5) taking decisions based on ethical factors that affect the safety and privacy of every entity in its environment. The result is *Wisdom=Knowledge+Ethics+Action.*

From the literature that discusses wisdom, at its most basic it can understood as:

- the use of what is available
- the execution of this in the form of a decision; and
- the result being the right, correct or most appropriate decision.

In terms of using what is available in the decision-making process (Johnson and Higgins, 2010:30), generally sources agree that this includes information, data and knowledge, supporting the general consensus that each entity in the hierarchy builds on the ones below

it. But In addition, wisdom is thought to include the use of experience, good judgement (Oxford English Dictionary, 2011), personal values, desired outcomes (Johnson and Higgins, 2010:96), belief, abilities, skills (Sternberg, 1998), and professional values (Johnson and Higgins, 2010:96). Baltes and Kunzmann (2003) argue that wisdom is not primarily a cognitive phenomenon, but that it involves cognitive, emotional and motivational characteristics. Baltes and colleagues (1939–2006) also found that wisdom involves a rich repertoire of procedural knowledge about how to perform certain skills and routines such as complex decision-making about interpersonal problems and conflict resolution. In addition, it involves lifespan contextualisation, which is an appreciation of the many themes and contexts of life such as self, family, peer group, school, workplace, community, society and culture, and the variations and interrelationships among these across the lifespan.

Baltes and Staudinger (2000) define wisdom as representing a truly superior level of knowledge, judgment and advice with extraordinary scope, depth, measure and balance. Wisdom addresses important and difficult life questions and strategies about conduct and the meaning of life and is the perfect synergy of mind and character, an orchestration of knowledge and virtues (Baltes and Staudinger, 2000). Wisdom also includes knowledge about the limits of knowledge and the uncertainties of the world (Baltes and Staudinger, 2000).

The *Berlin Wisdom Paradigm* defines wisdom as "expert knowledge of the fundamental pragmatics of life" and narrows those pragmatics to a set of criteria: rich factual knowledge, rich procedural knowledge, life span contextualism, relativism and the ability to understand and manage uncertainty (Baltes and Smith, 1990). Indeed, a key part of Ardelt's (2003) theory is that wisdom cannot be learned out of books, it is based on personal experience and integrated application, wisdom is gained through self-reflection of experiences and formulation of deeper goals (Bierly III et al., 2000). A person may have encyclopedic knowledge of the facts and figures relating to the countries of the world; but that knowledge, of itself, will not make that person wise. The wide knowledge has to be applicable to tricky problems of an ethical and practical kind, of how to act (Fricke, 2009). Learning from one's own mistakes is also one of the important ways to gain wisdom. However, it should not be taken out of context by assuming that wisdom is gained only from making mistakes. Nevertheless, learning from mistakes is useful including vicarious

learning. Learning from other people's mistakes is a substitute of learning from own mistakes (Liew, 2007).

However, wisdom is about more than the ability to acquire these things, it is about the understanding of how to execute them, and wisdom is the vehicle used for integrating all of these things into our decision-making process Johnson and Higgins, 2010:96). Jessup and Valacich (2003) see wisdom as allowing you to understand how to apply concepts from one domain to new situations or problems and Awad and Ghaziri (2004) suggest that wisdom gives the ability to see beyond the horizon. Ackoff (1989) purports that wisdom is the ability to see long-term consequences of any act and to evaluate them relative to the ideal of total control (omnipotence), which to Ackoff is the ability to satisfy any and every desire (Ackoff, 1989,8). Ackoff, for example, defined wisdom as evaluated knowledge, i.e. he defined wisdom as a process that makes use of knowledge in order to answer "difficult" questions while considering human factors like moral or ethical codes. Compared to the definition of understanding given above, this means that wisdom requires knowledge and – possibly several different – reasoning mechanisms that are able to handle complex additional constraints implied by, e.g., ethical codes (Hoppe et al., 2011).

Baltes and colleagues (1939–2006) also assert that wisdom entails an appreciation of the relavitism of values and life priorities with a tolerance for differences in values and priorities help by individuals and society in the service of the common good. The wise person is respectful of the unique set of values that other people hold, since the common good can be achieved by many routes. Baltes and colleagues (1939–2006) also assert that wisdom entails a recognition and management of uncertainty and a tolerance for ambiguity. It involves an appreciation that when solving any problem, each of us has access to incomplete information about the past and present; uncertainty about the future; and limited information-processing capacity. Wisdom is therefore the use of practical intelligence in a way that balances ones own interests and those of others involved in the problem and the wider community to achieve a common good for all.

Wisdom is also about more than just making any decision, it is about assimilating knowledge in such a way that it increases the power to act and exponentially improves the result from acting (Johnson and Higgins, 2010:95). It is one thing to turn information into knowledge that makes things happen, but wisdom goes beyond this to make the 'best'

thing happen (Johnson and Higgins, 2010:95). Ackoff (1989) defines wisdom as allowing agents to use knowledge, understanding and judgment effectively to achieve a balance between individual and collective human values.

In terms of arriving at the most appropriate decision, Rowley summarises early debates concluding that this is the behaviour that does the most good in terms of ethical and social consideration (Rowley, 2007). Whereas, Ackoff (1989) and others (Johnson and Higgins, 2010:64) assert that the most appropriate decision is the one that uses the context which knowledge gives data and information to increase effectiveness. Ackoff (1989) contrasts intelligence, which is the ability to increase efficiency, with wisdom, as the ability to increase effectiveness. This effectiveness could arguably be related to evil or immoral outcomes.

Baltes and Staudinger (2000) defined wisdom as representing knowledge used for the good or well-being of oneself and that of others. Wisdom has also been described as a grasp of the overall situation (Barlas et al., 2005) to achieve goals (Bierly et al., 2000; Hastie, Tibshirani, & Friedman, 2001) e.g. the realisation that increasing profits (goal) can be obtained by cross-merchandising two products that have a relation in consumer buying habits.

Wisdom supports the notion that the interest of the community and greater good outweighs individual self-interest. However, this is not to say that greater good should be achieved at the expense of self-interest. On the contrary, if practical wisdom was to have it, it would be a win-win situation where all interests are served (Liew 2013). Liew (2013) uses the example of James E. Burke's (former CEO of Johnson and Johnson) decision to remove $100 million worth of Tylenol painkillers worldwide off the shelves in response to seven people being reportedly killed after ingesting cyanide-laced Tylenol capsules in Chicago as an example of wisdom. Liew (2013) asserts that this decision illustrates action-oriented wisdom reflecting the knowledge of what is right and the courage to do it. In the aftermath, the company's reputation strengthened, and the overall corporate value increased instead.

As with the other entities in the hierarchy, there is much debate in the literature on the extent to which information systems have the ability to possess wisdom. Many have argued

that computers do not have the ability to possess wisdom, indeed Bellinger et al. (2004) assert that computers do not and never will have the ability to do so due to its inherent ethical aspect, which means that it 'resides as much in the heart as in the mind'. Rowley (2007) also posits that wisdom may have more to do with human intuition, understanding, interpretation and actions, than with systems. Carlisle (2006) also posited that wisdom rests in the capabilities of cognition and human understanding, and a computational wisdom base is currently difficult to imagine (Barlas et al., 2005). Schumaker (2011) posits that it is this incorporation of understanding that currently sets the divide between man and the machine and Pearlson and Saunders (2004) suggest that human input goes up in the higher levels of the hierarchy, whilst computer input goes down. Ackoff (1989) also alluded to the fact that information systems could not possess wisdom. In stating that wisdom is the ability to increase effectiveness and intelligence is the ability to increase efficiency, Ackoff writes that evaluations of efficiency are all based on a logic that, in principle, can be programmed into a computer and automated but does not say the same for wisdom. Hoppe et al. (2011) also argued that whilst data, information and knowledge are considered to be storable from a computer science standpoint, wisdom cannot and nor can it be transferred from one being to another. However, Chen (2014) argues that because of the fusion of humans, computers, and things in the hyper-world, developing human-level intelligence becomes a tangible goal of the DIKW related research. Thus, whether or not information systems can currently possess wisdom, as information systems develop this could be the goal of research in this area. Indeed, the concept of wisdom in computer science appears mostly related with the search for criteria that might permit the design of a computational system that shows the ability to provide real or at least simulated wisdom.

Bellinger et al. suggest that moving from data to information involves 'understanding relations', moving from information to knowledge involves 'understanding patterns', and moving from knowledge to wisdom involves 'understanding principles'. Thus, the extent to which information systems can possess wisdom could be evaluated by the extent to which they have the ability to understand principles.

The ability of information systems to possess wisdom will not only be dependent on the definition of wisdom that is being used but also the continuing research in computer science, artificial intelligence and information systems which may make the ability of information systems to possess wisdom a reality.

6.3.8 Criticisms of the hierarchy

Despite its wide acknowledgement, the DIKW model has been heavily criticised.

The hierarchy has often been criticised for lacking rigour (Fricke, 2009) because of the lack of agreement on the entities it includes, their definitions and the processes that convert the entities into each other (Rowley, 2007). Rowley (2007) found that it was unclear from the literature whether there is a sharp divide between the entities or whether they lie on a continuum, with different levels of meaning, structure and actionability. For example, there is general agreement that information is organised or structured data, but all data has some structure, either to enable it to be coded into an information system or for our minds to be able to 'make sense' of it, so structure alone cannot separate data and information. Additionally, whether an item of data in a database has any meaning to an individual, team or organisation, depends on the alignment between the data structure and the cognitive schema of the individual. This makes it difficult to objectively say whether something is data or information. Boddy et. al (2005) highlighted a similar difficulty in using actionability to differentiate between information and knowledge, particularly in the case of explicit knowledge. Yao (2019) argues that there may not exist a clear-cut boundary between any two adjacent levels. Because of this difficulty of defining the different entities, there is a there is a common phenomenon of defining the entities in terms of each other. Liew (2007) asserts that this is a circular definition and logical fallacy as defining their interrelationships does not constitute a definition of what they are as an entity.

The hierarchy has also been criticised for being too simple. Muller and Maasdorp (2011) assert that the possible explanation for the dominance of the model in Knowledge Management is because of its simplicity. However, this simplicity sometimes means that the hierarchy does not reflect real-world phenomenon. Mutongi (2016) posited that according to social constructivists, knowledge results from a far more complex process that is social, goal-driven and culturally-bound than often described in the context of the hierarchy. Other academics argue that proponents of the model only focus on elements that work within it, omitting other important aspects. (Mutongi, 2016). For example, Bernstein (2009) asserts that Ackoff focused only on specific modes of data, information, knowledge, and wisdom, and neglected important distinctions observed by information scientists such as Buckland (1991), Machlup (1980) and Soergel (1985). Fricke (2009) also

criticises Ackoff from omitting the question of 'why?' from the examples of information seeking questions such as 'who?', 'what?', 'where?', 'when?' and 'how many?'. Fricke (2009) argues that to answer a why-question you have to penetrate beneath the surface, to go beyond the 'data'; and that is exactly what the hierarchy approach forbids. As an example, Fricke argues that inspectors of an airplane crash would want to search for information to know why the crash occurred.

Another criticism of the hierarchy is that there is a flaw in the relationships between the entities, because this is not linear and there is no step by step through the lifecycle from data to wisdom (Johnson and Higgins, 2010:96). Instead, of a ground-up evolution of data through the information lifecycle, Wan and Alagar assert that Wisdom is often obtained by a repeated cycling through the DIK layers and that thus the pyramid structure for DIKW is somewhat misleading. Fricke argues that this focus on a linear model is uninspired and undesirable methodologically because it encourages the mindless and meaningless collection of data in the hope that one day it can be turned into information. Wan and Alagar concur with the doubly linked chain structure proposed by Ahsan and Shah (2011). This repeated cycle through DIK layers, denoted by $D \Leftrightarrow I \Leftrightarrow K$, is quite typical in the development of large computerized systems. To Mutongi (2016), the linear model and hierarchy also does not work as knowledge always comes first and is not really determined by information. To allow one to come up with the right data and information and analyse it, one has to be knowledgeable about that data/information. Thus, to Mutongi (2016) knowledge cannot be put in a hierarchy as for everything to function there is need for knowledge.

Another criticism of the hierarchy is that its intellectual backdrop is positivism (the thoroughly discredited methodological viewpoint of the 1930s) which believes that concepts which cannot be defined or measured by instruments are meaningless and that data and information must be rock solid true (Fricke, 2009). This focus on truth and certainty in the hierarchy aligns with Plato's definition of knowledge as justified true belief, but Mutongi (2016) argues that regarding knowledge as infallible might not be true in some cases. Knowledge changes with time and circumstances, what used to be true ten years ago may no longer be true today. Fricke also argues that this positivist approach means that the hierarchy may not account for statistical generalisations e.g. that most rattlesnakes are dangerous. If data must be true, a positivist application of the hierarchy would mean

that inferences must lead to true information and true conclusions. Yet, inductively derived data and information can be false, which a proponent of DIKW may or may not be content with. Either way, to allow for statistical generalisations would be a complete departure from building a pyramid on a solid base, yet not including them excludes a large amount of information and knowledge available and utilised in the world. For Fricke, either DIKW does not permit inductive, or similar, inference, in which case statements like 'most rattlesnakes are dangerous' cannot be information or it does, in which case it abandons its core faith that data and information have to be rock solid true.

Batra (2014) has also questioned the applicability of the hierarchy in the advent of big data analytics, suggesting that there is now a need to re-examine this. Big data is the collection and processing of large amounts of data which is generated daily as 'an exhaust or by-product' of innumerable activities in daily operations and modern technologies (Batra, 2014). Analytics are used to process this data, carry out predictive modelling and develop algorithms, which aim to find patterns and derive insights. These then help guide the future actions of enterprises and allow them to make strategic decisions in real time (Batra, 2014).

An example of big data analytics in practice is the prediction by Google of the spread of the H1N1 flu virus in specific regions of the US. Google analysed three billion search queries per day for 50 million most common search items related to flu to develop a large number of mathematical models and to identify a narrow set of models. These were then used for real time prediction of the spread (Mayor-Schonberger, 2013).

Batra (2014) argues that big data analytics challenge the DIKW hierarchy by:

1. Demolishing the conceptual boundaries between data, information and knowledge;
2. Replacing the need for human decision-making; and
3. Removing the need to search for causality

Batra (2014) argues that the conceptual boundaries between data, information, knowledge and wisdom are practically demolished in the advent of big data analytics because it is highly 'application and practitioner orientated' instead and does not need to dwell on the conceptual or semantic differences between the entities. The analytics create actionable information from algorithms, which Batra (2014) argues creates a jump straight from data

into action, ignoring the intermittent steps of information and knowledge of the hierarchy. Batra (2014) argues that this makes irrelevant the already fuzzy and fragile distinctions between data, information and knowledge. However, it is possible to break these processes down and link them to the hierarchy and Batra concedes that the predictive models and the results derived from them become the information and/or knowledge generated become a proxy for actionable information or knowledge in the context of the Hierarchy. Batra (2014) also notes that the development of big data analytics is the same as the DIKW's notion of giving meaning to data to make it information and knowledge. Batra (2014) also argues that big data relies on data being the key resource, whereas knowledge was the key resource in the knowledge management era and enterprises now take pride in being data-driven. Batra conceded that a subset of the entire data serves as test data and the learnings derived from it are then applied to the remaining data. These learnings are then effectively knowledge, and thus knowledge is still as important in the big data era. It is also difficult to maintain an argument that either data or knowledge is more important than each other whilst maintaining an argument that the boundaries between these two entities are effectively demolished. Thus, it seems that in the advent of big data analytics, there may be less purpose in distinguishing between the entities and applying the hierarchy for big data to produce a result, yet it can be done and there are use cases where applying the hierarchy helps in understanding the phenomenon of big data analytics.

Batra (2014) also argued that because algorithms generate actionable information, the decision-maker does not have to search for causality. Predictive models are essentially based on correlation analysis among variables identified by sifting through the entire data population to discover patterns and correlations (Mayor-Schonberger, 2013). As such, the focus is on arriving at the most likely probability for a prediction to be true, rather than on predicting with certainty. Yet, despite less focus on causality, the advent of big data analytics does not mean that causality never matters. Indeed, Batra (2014) points out that although the result of big data analytics is actionable information or knowledge, if a human reviews it or wants to create wisdom from this, causation will need to be examined.

Human judgement may also be required to take the final call on whether a specific action should be taken and Siegel (2013) and Mayor-Schonberger (2013) have devoted considerable attention to whether it would be ethical to take certain actions based only on the predicted outcome of a situation. It is also worth noting that if the big data analytics

result in an automated-decision is about an individual, then under the European Union Data Protection Framework meaningful information about the logic involved in the decision will be required if it produces legal effects or similarly significant effects for the individual (Article 13(2)(f) and Article 22 GDPR) which means at least in the context of personal data this causality will be required.

Finally, Batra (2014) also argues that big data changes the way decisions are made, replacing the need for human decision-making. Where previously a given set of hypotheses generated through careful understanding of the interdependence of various factors would be applied, decision-making is increasingly a function of predictive modelling, by learning from data to predict the future behaviour of individuals. It is this that then drives decisions. Yet, as discussed in relation to causation, humans may still want to review the output before a decision is made and potentially other sources, not just the output of big data analytics. Thus, the difference here is that the careful understanding of various factors is done by the analytics in an information system (often in real time) rather than the human mind. Decision-makers are still interested in data, information, knowledge and wisdom, they just arrive at it a different way.

Despite Batra's comments what emerges is simply a different application of the hierarchy, with information systems collecting and processing knowledge and wisdom where humans once did. The key difference between the processing of big data and traditional data seems to be where the information and knowledge is held and produced, how quickly this can be done and how much data can be used. Mayor-Schonberger (2013) points out that because of data's vast size, decisions are made by machines and not by humans (though humans as always have the authority to override machine-made decisions).

6.3.9 Summary

Although the model has received a large amount of criticism, many academics still recognise the value it has and the theoretical gap it would be left were it to be abandoned completely. Fricke (2009) questioned whether the model should no longer be part of the canon of information science and other disciplines but concluded that abandoning it would leave an intellectual and theoretical vacuum over the nature of data, information, knowledge and wisdom, and their interrelationships.

Baškarada and Koronios (2013) also noted that as most information systems literature relies on these concepts, abandoning the hierarchy does not seem feasible and that instead, developing clear, consistent, and unambiguous definitions of the terms, their relationships, and their quality dimensions is imperative. Others argue that the hierarchy works better in some contexts than others. Bernstein (2009) conceded that despite the model glossing over various kinds of data, information, knowledge and wisdom, it captures insights useful not only by personnel and organisational management but also to the organisation of library resources and Mutongi (2016) argues that the hierarchy makes the study of information Science and Information and Communication Technology (ICT) easier, but narrows the study of knowledge and Knowledge Management.

Thus, despite the criticisms it received there were a number of reasons why the DIKW hierarchy seemed to be an appropriate model upon which to base the proposed categorisation of personal data.

Firstly, because the purpose of the concept of 'personal data' is to define a whole ecosphere, it seemed appropriate to find a model that was capable of conceptualising all information and acknowledging levels within this.

Secondly, a review of the literature on the model made clear that the criticisms the model has received would not affect its adequacy as a basis for categorizing personal data. The criticism that there are many different versions of the steps in the hierarchy and much debate about the definitions of each level (Frické, 2009) is not problematic here, because the model would be adapted and applied to personal data, meaning the specific definitions provided in this chapter and adopted in relation to the new approach to categorizing personal are the only ones which matter. Thus, the various different versions are actually a positive, because from these the most appropriate steps and definitions for the purpose of a model for categorising 'personal data' can be used. This is also true of the criticism that there is a lack of clear definition for what is meant by 'wisdom', as an appropriate interpretation of this can be defined for the new model.

Thirdly, Frické's (2009) criticism of the DIKW Hierarchy, that it ignores the huge domain of the unobservable for which no instruments of measurement exist is also not relevant here. This criticism rests on the fact that for the purposes of information science and knowledge

management, information must always rest on data measured by instrument, which must always be true. Whereas, under the data protection framework, 'personal data' need not be true or proven (European Commission, 2007:6). This will therefore not be an issue for the application presented in this thesis; neither will the criticism that the hierarchy does not accommodate universal statements or statistical generalisations.

Therefore, the DIKW still proved to be a useful model upon which to base a new approach to categorizing personal data and in fact it has been used as a basis for this before, as discussed in the next section.

6.4 The DIKW Hierarchy and Personal Data

6.4.1 Berčič and George's Application

As discussed, the DIKW hierarchy provides a lot of promise as a basis for a new model and categorization of personal data. However, the model cannot readily be applied to the legal framework. Whilst the fact that there are various versions in the hierarchy allows for a flexible application of it in domains beyond information science and knowledge management, it does mean that connections between data protection law and computer science need to be made.

This thesis is not the first to make the link between data protection and the DIKW Hierarchy, as Berčič and George (2009) attempted this connection, by investigating how these semantic forms in information science mapped to correlative concepts under the DPD, and how each of them were legally protected. Given the many different versions of the DIKW hierarchy, their paper began by proposing a 'data, information, knowledge, rules, hierarchy' (DIKR - depicted in Figure 15).

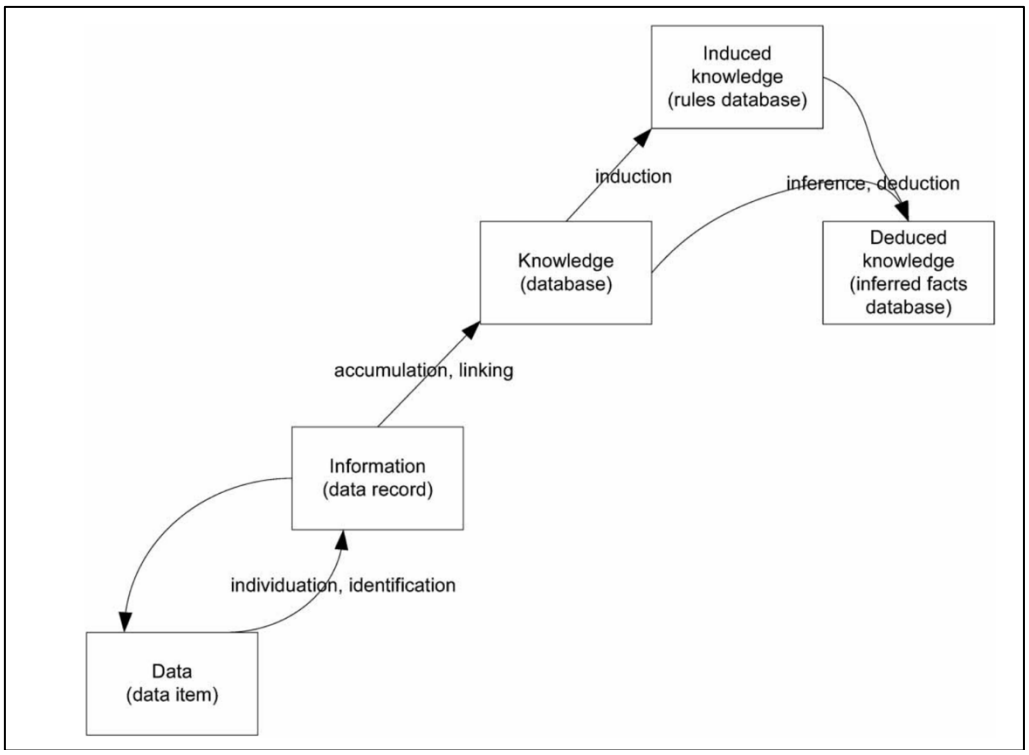


Figure 14 Berčič and George's (2009) Version of the DIKW Hierarchy

Berčič and George (2009) suggested that moving up the hierarchy means adding something new to a higher level and extracting something from a lower level. Instead of wisdom, Berčič and George refer to 'understanding' and describe it as cognitive and analytical and the process by which 'new knowledge is synthesised from previously held knowledge' (Berčič and George, 2009). *Understanding* refers to rules derived from knowledge by induction (e.g. a rules database used in expert systems and artificial intelligence). it is called *induced knowledge*. There is another standard part of expert systems, apart from the facts and rules database, called the 'derived or inferred facts database', which is obtained by inference from the facts database using the rules database. This is called *deduced knowledge*. Table 17 depicts the semantic concepts and their links to the DPD.

Berčič and George posit that data refers to anonymous information that is data about an unidentified individual or anonymized data, it contains facts about unidentified entities and that they are not yet connected to the entities that they refer to.

Table 17 Berčič and George's Protected concepts in the processing of personal data

Semantic Concept	**…Is incarnated in…**	**…And Contains…**	**... and is exemplified by ...**	**... and mentioned in the Directive**	**Typical data operation**
Data	unidentified data record	fact about unidentified individual	anonymous DNA analysis	anonymized personal data (Article 26 of the Recitals)	anonymization
Information	identified data record	fact about identified individual	analysis of DNA fragment of a known individual	personal data (Article 2 of the Directive)	individuation
Knowledge	collection of personal data records (database)	many facts about identified individual; individual's profile, linking databases	DNA analysis of a known individual in its entirety	data filing system (Article 2 of the Directive)	accumulation and linking of facts
Deduced Knowledge	collection of inferred facts from the facts and rules database (deducing new facts with rules from the theory or expert system)	inferred facts from the facts and rules databases	deducing possible health problems from DNA analysis	(a) automated processing of data intended to evaluate certain personal aspects relating to individual, such as his performance at work, creditworthiness, reliability, conduct (Article 15 of Data Directive, (b) processing for statistical purposes (paragraph 2 of the Article 11 Directive) processing for the purposes of historical or scientific research (Article 29 of the Recital, Article 6 of the Directive, paragraph 2 of the Article 11 of the Directive)	inference, deduction, application of statistical and other scientific methods
Induced Knowledge	rules database (inducing new theories about individual)	rules about individual	testing individual responses to drugs (pharmacogenomics); inducing buyers' buying habits on basis of their buying history		induction, theory formation, hypothesis testing, learning rules

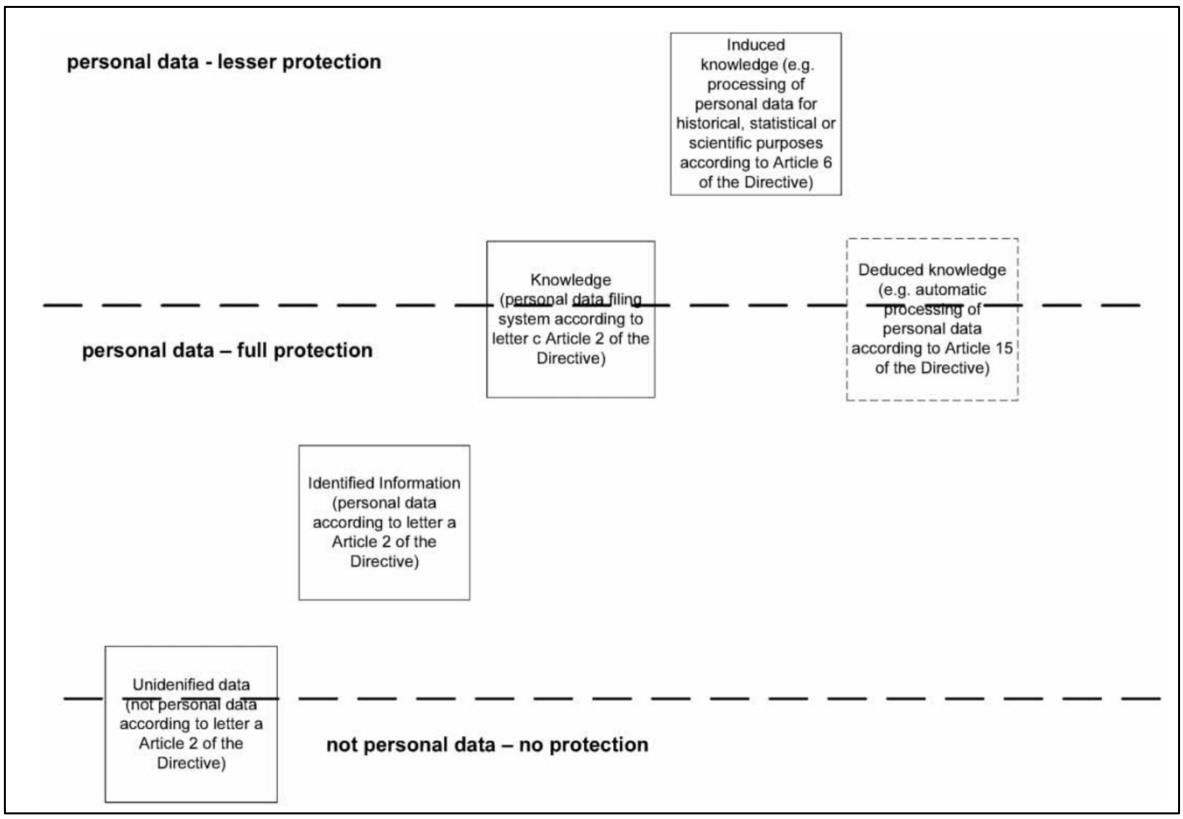


Figure 15 Interpretation of the actual protection of the DIKR Hierarchy under the DPD

Information refers to personal data (i.e. data about an identifiable individual), this can be an identified data record in a database or in a spreadsheet. While the Directive did not protect data that does not identify an individual, it does protect identifiable data (data, according to the DIKR hierarchy) as personal data (information according to DIKR hierarchy). That is, in this case it gives the same protection to data that it usually gives to information.

Knowledge refers to a personal data filing system (i.e. a database) according to Article 2(c) of the Directive. Berčič and George posit that a typical operation on the third level of the DIKR hierarchy is the linking of existing databases. Deduced knowledge forms part of yet another data filing system and that Induced knowledge (rules database) does not have a specific correlative in the data protection parlance. This is depicted in Figure 16.

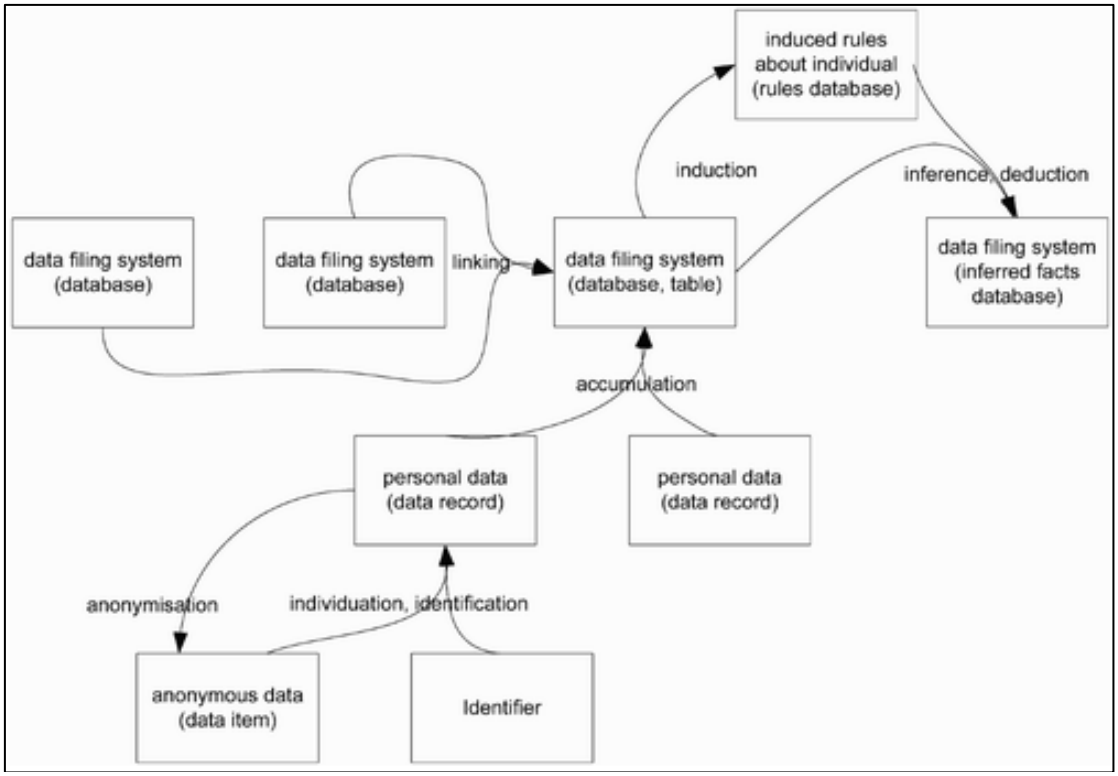


Figure 16 Berčič and George Interpretation of the DIKR Hierarchy in terms of Data Protection

Berčič and George also explain that there are several ways of adding something new to a level. One way is *individuation*, the change of data into information by way of identification (hence forming a personal data record). Another is the *accumulation* of information (i.e. different items of personal data) about an individual to form a database table or an entire database. A third way is the *linking of database tables* to create a new database. A fourth is *linking of disparate databases*. A fifth is the *application of common knowledge and methods* (rules, categorization schemes, statistical methods) to information about an individual thus obtaining new, derived facts about the individual. The most advanced example of the latter is the making of hypotheses (induction) about an individual on the basis of knowledge contained in a database. Berčič and George (2009) link these activities to the legal concept of 'processing' under Article 2(b) of the Data Protection Directive.

According to Berčič and George (2009), deduced knowledge is knowledge extracted from a knowledge database according to some predetermined procedure (e.g. application of rules from a rules database or by application of statistical procedures). It is inferring new facts from induced rules and present knowledge. In expert systems terminology in addition to having a facts database, and a rules database, a typical expert system will also have a

database with derived or inferred facts, which is obtained by the application of inference procedures from the facts database using the rules database as depicted in Figure 16.

Berčič and George argue that since derived facts do not differ from otherwise obtained data, they can be deemed to be information (when alone) or knowledge (when collected in a database). Whilst from a systems perspective this may be true, from the perspective of the individual there is a big difference between the information that the individual has provided and information that is derived or inferred by the data controller.

Berčič and George link deduced knowledge to Article 15 (Automated Decision Making) and Article 12 (Right of Access, including to the logic of such automated decisions) of the then Directive (now Article's 22 and 15 GDPR). The logic is the content of the rules database. Yet, these are not the only scenarios where information will be deduced, and these Article's will not always apply as there will need to be a legal effect etc. for them to come into play.

Berčič and George believe that the application of general knowledge to an individual's personal data from their profile merits special attention and should either be mentioned separately in the Directive or interpreted as automated decisions which fall under the scope of Article 15. The problem with this is that it would make Article 15 too wide in what it catches and Berčič and George do not specify how it should have special protection, you also cannot change the law easily. and George's theory is that an individual should have the right to obtain knowledge of the logic involved to create the database as they argue that the right does not arise if such data are not construed as automated decisions. Does access extend to derived or inferred data?

Berčič and George argue that personal data processing is limited by the DPD to some extent by the principle of purpose limitation under Article 6 (now Article 5 GDPR), and that if the purposes do not involve diagnostics etc. using personal data for other purposes than that for which it was collected should not be done. Yet, this offers little protection, as purposes are wide, as is the definition of compatible and purposes do not necessarily indicate whether further information will be derived and used for this or just the personal data that has been provided. Berčič and George use the example of an individual paying an institution to decipher his DNA for the individual to store for their own use meaning that they have not given the institution the permission to also run a diagnostic test on it, then this should

not be done. Yet there are many ways in which this can be done compliantly and in an opaque way to the individual. Including for historic, scientific purposes etc.

Berčič and George define induced knowledge as the scientific discovery of rules on the basis of present knowledge. They use the example of testing individual DNA responses to drugs or inducing individuals buying habits on the basis of their buying history as examples of induced knowledge. Berčič and George assert that this is not specifically mentioned in the DPD and that if anything, such processing would warrant a lesser degree of protection than ordinary processing of personal data because they would classify it as archiving purposes in the public interest, scientific or historical research purposes or statistical purposes which received lesser protection under the DPD and now GDPR. In the opinion of Berčič and George, this type of processing should merit special treatment too or at least the same treatment as that of automated decision-making.

Berčič and George made two recommendations regarding the scope of protection of the various forms of data processing:

- That individuals should be protected with respect to the application of general knowledge to personal data related to them (deduction) in the same way they are protected with respect to automated processing of data defined in Art.15 of the DPD (irrespective of whether such application of general knowledge to data related to them includes automated decisions or just automated inference).
- Induced knowledge about an individual should merit more (not less presently) protection as ordinary knowledge – again, one of the possibilities would be to warrant the same treatment as is given to automated decisions.

6.4.2 Why can't Berčič and George's Model Be Adopted?

Although Berčič and George have already linked the DIKW to data protection law, there are various reasons why their model cannot be adopted here and this thesis takes a different approach.

First, they linked the DIKW Hierarchy to the whole of the data protection ecosphere, including personal data and anonymous data, whereas the categorisation this thesis proposes focuses just on using it within the definition of personal data.

Second, in describing how the DIKW concepts correlated to concepts under the DPD Berčič and George found that some concepts overlap, such as their definitions of deduced and induced knowledge. However, as discussed in Section 6.2.4, the categorization required will need to have mutually exclusive categories in order to be legally certain as different obligations and rights will be attached to them.

Third, their mapping needs to be updated in light of the GDPR. Particularly, the fact that they argue that individuals should be protected with respect to the application of general knowledge to personal data related to them (deduction) in the same way as they are protected with respect to the automated processing of data under Article 15 DPD, irrespective of whether such application includes automated decisions or just automated inference (Berčič and George, 2009:201). Under Article 4(4) GDPR, the concept of 'profiling' is introduced, which warrants this further protection.

It is for these reasons that the categorization presented in the next section of this thesis goes beyond this work.

6.5 The New Categorisation of Personal Data

6.5.1 Introduction

As discussed, the research goal of this thesis is to understand deficiencies in the current approaches to the transparency of personal data processing in the context of the obligation to inform. Its goal is to propose an improvement to the way organisations can be transparent about their personal data processing. This Chapter proposes that the DIKW Hierarchy is adapted and used as the superordinate level lens through which personal data should be categorised. It proposes that the obligation to inform individuals of the categories of personal data that are processed about them should be to inform individuals whether they process data, information, knowledge or wisdom about them and whether this includes personal data or non-special category personal data, with a link to the specific data types which are processed in these categories.

The reason for this new approach is because, as discussed in Chapter 5, despite the term 'category' being used to describe the legal requirement for providing information about the personal data being processed, a consistent approach to categorising personal data has not organically occurred. Additionally, none of the individual current approaches are sufficient in making the processing of personal data transparent to a level that equates the information available to data subjects to that which is possessed by data controllers.

From the categorisation approaches examined in Chapter 4, it was found that as a whole, categorisations of personal data focused on four themes: the identifiability of the data subject, the sensitivity of the personal data, what the data is (e.g. the data types processed) and the source of the data. Chapter 4 also highlighted four benefits for categorising personal data for transparency, which were: enabling an assessment of privacy risk, reducing the overall information that needs to be provided to increase transparency, allowing for the contextualisation of other information under the obligation to inform and allowing for the attachment of different levels of protection or obligations in relation to the categories.

In analysing which of the current categorisations take these approaches and enable these benefits, it was found that none combined all four themes, and therefore one single current approach did not meet all the highlighted benefits for transparency. It was because of these findings that the new approach presented in this Chapter is proposed, to attempt to incorporate as many of the four themes found in the current approaches to categorising personal data as possible. The purpose of the new approach is also to achieve as many, if not all, of the benefits of providing information on the categorisation of personal data for transparency.

The purpose of this new categorisation is not just to focus on what personal data is held by a controller, but also the outcome of any operations on it that will produce further personal data, given that almost any personal data can reveal other personal data and even sensitive or special category personal data. Because of this ability, it is difficult for controllers to fully anticipate the personal data that they will be able to infer or derive from a dataset, let alone expect individuals to be able to anticipate this, and the impact of it. Current approaches focus on the personal data that is held, rather than incorporating an element of categorisation which indicates whether personal data will be used to further infer and derive personal data and what this is, which the new approach seeks to address.

It is envisaged that this new approach to categorisation of personal data can be used by various parties to increase transparency, including but not limited to, system owners, data engineers, legal experts and data protection regulators, in addition to individuals themselves. It can be used to overcome the long-recognised issue of designing legally compliant technologies (Goodman and Flaxman, 2016; Greengard, 2018) by providing a mediatory object which both data engineers and legal experts can use to discuss legal requirements and system capabilities/data processing initiatives. This can empower those working with the data to understand the extent to which the data that they are processing is personal or special category personal data under the EU Data Protection Framework and whether how they process it (or would like to) would categorise it as data, information, knowledge or wisdom. This new approach can support an understanding of what legal requirements apply and when engagement with a legal expert in this area for further advice is required.

The new approach can also support in the transparency of personal data processing for data subjects, because legal experts can better understand how personal data is being processed from those processing personal data within the business and ensure this information is provided to data subjects in order to comply with the obligation to inform. Furthermore, by simplifying the information that is required by data subjects to understand the extent to which controllers will be inferring and deriving further personal data about them, this can increase the transparency of personal data processing for individuals and help them to distinguish between controllers that will be inferring and deriving more personal data and those that will not. Indeed, in addition to being transparent about what controllers do, they can also use this new approach as a basis to convey what they do NOT do e.g. '*We do not process wisdom or knowledge about you, so whilst we capture your sexuality we do not derive anything further from it by using it to create knowledge or wisdom about you'*. Because the new approach has been designed to fit within the current laws of the EU Data Protection Framework, it could also currently be used by Regulators or those tasked with interpreting the law to clarify the scope of existing rights of data subjects and obligations of controllers in the context of personal data. However, it could also be used to shape any future rights and obligations which may be introduced under the Framework.

In presenting the new approach in this Chapter, Table 17 is not intended to be a depiction of how this information should be presented to data subjects but is the flow chart that should be worked through in order to categorise personal data using the proposed definitions in Table 18 of data, information, knowledge and wisdom. Table 18 also provides an example in practice of the type of processing that would indicate whether the controller was processing data, information, knowledge or wisdom in addition to how they could then communicate the categories of personal data they are processing to an individual under the obligation to inform in this example. In particular, whether these elements are special category personal data or non-special category personal data, before the data types are presented, as this supports and enables the individual to make a personal assessment of risk about these categories of personal data being processed.

.

The proposed new approach to categorising personal data at the superordinate level is depicted in Figure 17. The starting point is that all personal data will either be processed by controllers as data, information, knowledge or wisdom level and then within these categories it will either be special category or non-special category personal data that is processed (or produced at the wisdom level). Within the category of special category personal data, the categories provided by the framework are included and then below this, the specific data types of personal data that belong to these categories would sit e.g. 'caucasian' is a data type the category of 'processing of personal data that reveals racial or ethnic origin'. Because the law does not provide the equivalent categories for non-special category personal data, this is not included in Figure 17. Arguably a categorisation of personal data at this level is not required as once the data type is provided, information about the category will not increase transparency which is why it is not recommended that the individual be informed of this level of the categorisation. For example, if a controller informs an individual that you use their Facebook likes to generate a probabilistic prediction of whether they are heterosexual, homosexual or bisexual, it does not increase transparency by also informing them that the controller processes data concerning the individual's sexual orientation.

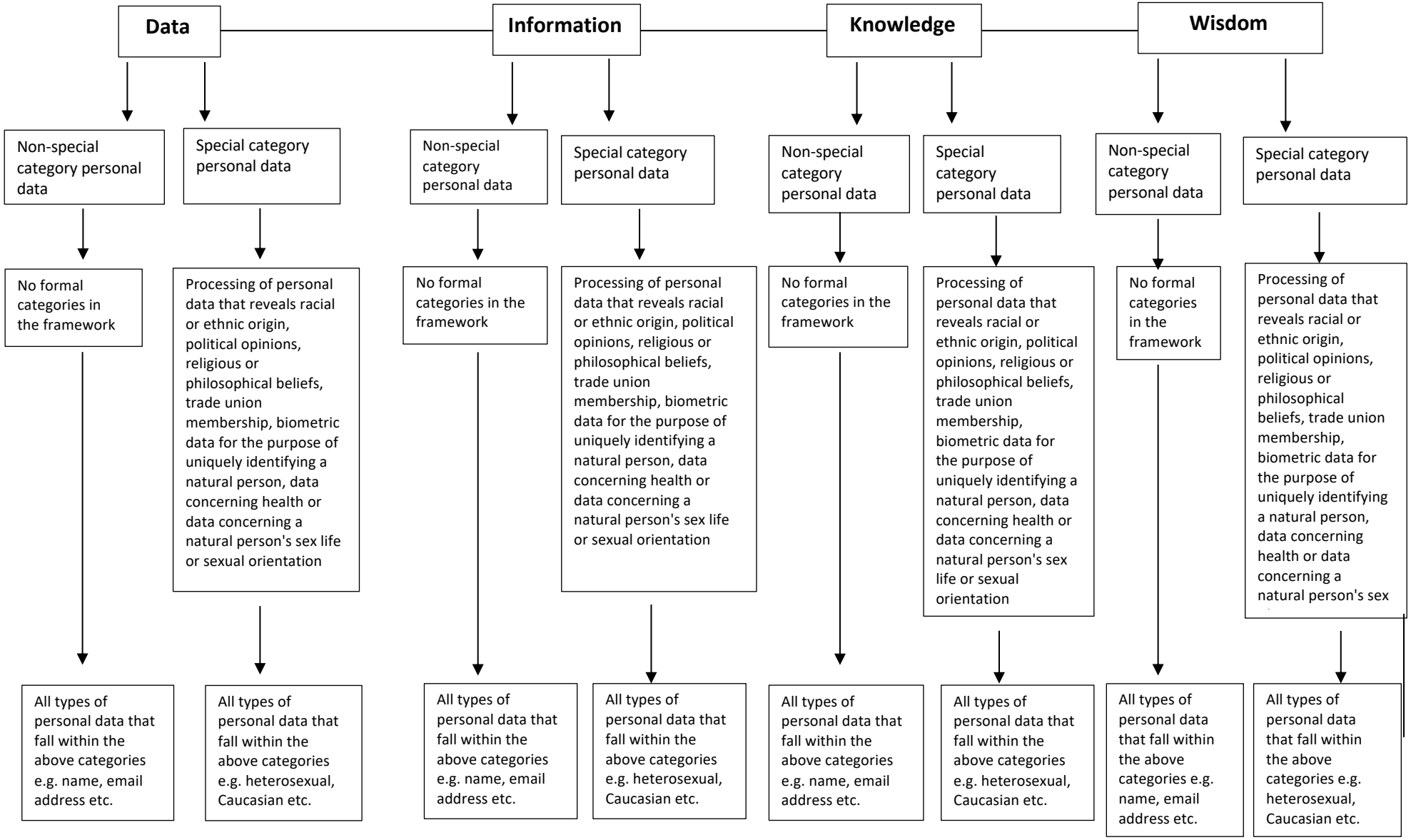


Figure 17 Overview of New Approach

Table 18 New Approach to Categorising Personal Data

Category	Definition	Example in Practice	Example of Informing on the Categories of Personal Data
Personal Data as Wisdom	Persona Data which relates to an identified or identifiable individual which is provided by them or obtained from elsewhere which is used to infer or derive further information about the individual or to which probability-based analytics are used to find correlations which provide new ways to infer personal data about individuals which either uses or produces special category personal data.	A controller combines all the types of personal data it has about an individual including their click-through rate with all the types of personal data it has about its other customers and their click-through rate to identify correlations between these which will improve its ability to predict people's click-through rate.	*"We process the non-special category personal data we have about you at the wisdom level because we use it to understand new ways to infer personal data about individuals. For a full list of the data types of personal data we process in this category click here. We update this list of data types as we discover new data types about you".*
Personal Data as Knowledge	Personal data which relates to an identified or identifiable individual which is provided by them or obtained from elsewhere and from action beyond consultation of it in a system will be taken to it but that will not be used to generate more personal data about the individual.	An individual's click through rate which is calculated by the ratio of ads the individual clicks on to the total number of ads they are shown.	*"We process the non-special category personal data we have about you at the knowledge level because we use the data types that we have about you to derive other personal data about you. For a full list of the data types of personal data that we process in this category click here".*
Personal Data as Information	Data which relates to an identified or identifiable individual and that has been provided by them or obtained from elsewhere that is consulted by a human; or computer in answer to a question but is not used to generate additional personal data about the individual and does not physically leave the relevant filing system in which it is stored other than for backup purposes other than for backup purposes.	The content that is posted to a SNS which may be viewed by humans or computers to make sure that it does not contain any illegal or offensive content.	*"We process the non-special category personal data we have about you at the information level because we access it but we do not use it to create more personal data about you. For a full list of the data types of personal data that we process in this category click here".*
Personal Data as Data	Data which relates to an identified individual and has been provided by them or obtained from elsewhere that is simply stored and not accessed or used for any additional purposes other than providing the service.	Data generated by use of a connected product in order for it to work e.g. a connected heating system which when an individual wants to remotely change the temperature at home, sends a message from the app to a company's cloud storage and then to the heating	*"We process the non-special category personal data we have about you at the information level because we only store it and do not access it or use it for any other purpose other than storage. For a full list of the data*

Category	Definition	Example in Practice	Example of Informing on the Categories of Personal Data
	Personal data which relates to an identifiable individual and has been provided by them or obtained from elsewhere, which is only consulted by the controller, does not physically leave the primary relevant filing system in which it is stored other than for backup purposes and is not used to generate additional personal data about the individual.	system to action the request. This personal data as messages is simply stored within the company's cloud storage by default but is never accessed or utilised by humans or computers beyond providing the service.	*types of personal data that we process in this category click here".*
Special Category Personal Data as Wisdom	Personal Data which relates to an identified or identifiable individual which is provided by them or obtained from elsewhere which is used without the individual's awareness to infer or derive further information about the individual or to which probability-based analytics are used to find correlations which provide new ways to infer personal data about individuals which either uses or produces special category personal data.	A controller combines all the types of personal data it has about an individual including special category personal data type of their sexuality and their click through rate with all the types of personal data it has about its other customers and their click through rate to identify correlations between these which will improve its ability to predict people's click through rate.	*"We process the special category personal data we have about you at the wisdom level because we use it to understand new ways to infer personal data about individuals. For a full list of the data types of personal data we process in this category click here. We update this list of data types as we discover new data types about you".*
Special Category Personal Data as Knowledge	Special Category Personal Data which relates to an identified or identifiable individual which is provided by them or obtained from elsewhere and action beyond consultation of it in a system will be taken to it but that will not be used to generate more personal data about the individual.	A controller captures individuals' weight and height to calculate their Body Mass Index.	*"We process the special category personal data we have about you at the knowledge level because we use the data types that we have about you to derive other personal data about you. For a full list of the data types of personal data that we process in this category click here".*
Special Category Personal Data as Information	Special category personal data which relates to an identified individual and that has been provided by them or obtained from elsewhere that is consulted by a human or computer in answer to a question but is not used to generate additional personal data about the individual and does not physically leave the primary	A SNS that captures an individual's sexuality but only for the purpose of generating anonymous statistics about the users of their service.	*"We process the special category personal data we have about you at the information level because we access it but we do not use it to create more personal data about you. For a full list of the data types of personal data that we process in this category click here".*

Category	Definition	Example in Practice	Example of Informing on the Categories of Personal Data
	relevant filing system in which it is stored other than for backup purposes.		
Special Category Personal Data as Data	Special Category personal data which relates to an identified individual and has been provided by them or obtained from elsewhere but is simply stored and not accessed or used for any additional purposes other than providing the service. Special Category personal data which relates to an identifiable individual and has been provided by them or obtained from elsewhere, which is only consulted by the controller, does not physically leave the primary relevant filing system in which it is stored other than for backup purposes and is not used to generate additional personal data about the individual.	An individual uploads a copy of their health insurance claim form to a cloud storage facility. The controller does not access the document but technically processes the personal data in order to store it.	*"We process the special category personal data we have about you at the data level because we only store it and do not access it or use it for any other purpose other than storage. For a full list of the data types of personal data that we process in this category click here"*

Table 18 shows how data, information, knowledge and wisdom are defined in the context of the framework. It provides an example of personal data that would fall within this category and then what information would be provided by the controller in the privacy policy or the method they use to provide information to the individual on this.

This approach to categorisation and the requirement to specify the categories of personal data being processed under the obligation to inform asserts that it does not make a difference whether the personal data is obtained from the individual or elsewhere in the actual categorisation in line with the UK DPA 2018 in terms of the information that should be provided under the obligation to inform and data at all levels can be provided by individuals or sourced from third parties. The specific data types could be divided into the sources which they are from to provide this information.

The approach also does not distinguish between 'derived' and 'inferred' personal data as being separate in the way the OECD did for the purposes of the obligation to inform as both generating new personal data about individuals or know ways to process personal data about individuals is viewed at the wisdom level because these both seek to know more about individuals than they are intentionally giving away.

The approach also distinguishes between merely consulting the data in a database to answer a specific question, without using this personal data to generate any other personal data and actioning the personal data beyond this e.g. using it to send direct marketing to individuals or selling it. This is because actions beyond consultation will often result in the personal data being stored in multiple places and increased risk for the processing of personal data.

6.6 Validation

As discussed in Section 5.1.2, a combination of requirements analysis and case study methodology was chosen as the methods that would be used to validate whether the new approach to informing individuals of the categories of personal data being processed about them under the obligation to inform increases the transparency of personal data processing.

Requirements Analysis is an integral part of information systems design and is critical to the success of interactive systems, it is used to create the functional requirements of a system that scenario-based testing subsequently tests. The requirements should be documented, actionable, measurable, testable, traceable, related to identified business needs or opportunities, and defined to a level of detail sufficient for system design. Requirements Analysis generally happens in three stages (Hammond, Gross and Wesson, 2002):

- Requirements Discovery: requirements can be discovered from various different sources and using various different methods e.g. business process documentation, stakeholder interviews
- Recording Requirements: requirements can be documented in various different forms, for example in a summary list, in use cases, user stories and process specifications
- Analysing Requirements: requirements are analysed to make sure that they are clear, complete, unduplicated, concise, valid, consistent and unambiguous

In order to validate the proposed new approach the three steps of requirements analysis were taken to identify the requirements for transparency which the new approach would be tested against, these are listed in Table 19 and are gathered from the previous work of this thesis which effectively formed a set of requirements for an appropriate categorisation of personal data under the obligation to inform. In particular the requirements come from understanding what the benefits of categorising personal data are (Section 5.3.1), what the themes of current categorisations of personal data are (Section 5.5.1) and the requirements for a classic categorisation of personal data (Section 6.2) and are used to form a list of requirements to indicate an appropriate categorisation of personal data under the obligation to inform, against which the new approach can be validated. In relation to Requirement 2(d) the categorisation must indicate source, the OECD'S version of this was used (provided, observed, inferred and derived) opposed to source in terms of 'from your device' etc. because this was the most comprehensive example of source.

The first step and second step, of discovering the requirements and recording them has been undertaken throughout this thesis when: investigating what the benefits of an

appropriate categorisation of personal data to increase the transparency of personal data processing under the obligation to inform would be in Section 5.4; what the current approaches to categorisation in practice are in Section 5.5 and 5.6, and what the requirements for a classic categorisation approach are in Section 6.2. To complete step two, these requirements were brought together as a summary list in Table 19 and analysed to make sure that they are clear, complete, unduplicated, concise, valid, consistent and unambiguous. These are the requirements against which the new approach will be tested to see if it satisfies them which is presented in Section 6.6.1. Table 19 and Section 6.6.1 also contrasts the new approach with the other approaches to categorisation of personal data

Case studies are:

> *"An empirical inquiry about a contemporary phenomenon (e.g." a case"), set within its real-world context – especially when the boundaries between phenomenon and context are not clearly evident".*
> **Yin (2009a, p.18)**

There are three common steps to designing a case study. The first is to design a 'case', which is a bounded entity e.g. a person, organization, behavioural condition, event, or other social phenomenon, that you are going to study. The second is to decide whether your case study will consist of single or multiple cases, the third is to decide whether or not to use theory to help complete your essential methodological steps e.g. developing research questions and selecting your cases (Yin, 2012). Sources of evidence in case studies can come from multiple sources, from direct observations and interviews to archival records, documents and physical artifacts.

Multiple cases were chosen for the purpose of this thesis, these were the two hypothetical cases of a how a Social Networking Site and a Connected Heating System may process personal data. The reason for choosing the case of a Social Networking Site was because this was the main case that has been studied at various points throughout this thesis. The reason for choosing a second case was to contrast with the application of the new approach to SNS to understand whether the results of the case study are more generalisable and whether this approach has use beyond SNS, to another internet-based service.

Because each SNS and Connected Product will process different types of information and for different purposes, and because of the current lack of transparency on exactly how these services are processing personal data which this thesis seeks to improve, two hypothetical case studies were constructed based on information available and the researcher's knowledge of how these systems process personal data. These are presented below in Section 6.6.2 and then the application of the new approach to categorising personal data is discussed in relation to them, including how it would vary if aspects of the case study changed.

6.6.1 Validation against requirements

In relation to requirement 1(a), that the categorisation will allow for a subjective assessment of any risk involved in the processing, this requirement is met by the new approach. Whilst arguably categorisation of superordinate level of data, information, knowledge and wisdom only allows for an objective assessment of risk, because the lower the level, the less risk is involved in the processing, the provision of data types in the context of this allows the individual to make a subjective decision about whether there is a risk to their privacy from processing of this category. For example, if an individual is informed by a controller that they are processing special category personal data at the wisdom level and that this will result in the data types of heterosexual, homosexual or bisexual bring predicted about them and they have not yet informed anyone else of their sexuality they can make an assessment of the risk to their privacy of this being processed.

As shown in Table 14, in relation to the other approaches to categorisation, categorising in relation to identifiability can be useful for understanding the ability of the data controller to link the data to the individual, which arguably allows for an assessment of risk. Categorising in relation to sensitivity can also help a data subject to assess the risk of the personal data processing by knowing whether it was sensitive personal data or non-special category personal data that is being processed. However, as discussed previously, this is quite a binary distinction and so whilst it allows for an objective assessment of risk, it does not allow for a subjective assessment as the processing of data that is classed as sensitive will increase the risk for everyone, compared to the processing of non-special category personal data which is why this requirement is partially met.

In relation to categorising the data by what it is, this approach can allow for an objective and subjective assessment of risk. Categorising in relation to source only partially meets this requirement because, whilst it helps individuals to understand whether it is just the information they provide that the controller will process about them which indicates some level or risk, more granular information about what this personal data is would be required for a true assessment of risk.

In relation to requirement 1(b), that the categorisation reduces the amount of information that needs to be provided under the obligation to inform, this requirement is partially met because whilst it has the potential to be met, although further validation will be required to confirm this. Arguably, the recommendation in this thesis increases the amount of information that needs to be provided initially because the categories need to be listed and then all of the data types that are processed within this. However, once the categories of personal data of data, information, knowledge and wisdom are understood by individuals then the amount of information that is required to be provided to individuals well be reduced. This could also allow for the potential for visualisations for the different categories could be provided. For example, if a controller informs the individual that it processes personal data at the data level they understand what this means in practice then they will not also need to be informed of the purposes which the personal data is used for as they will know that the only purpose is to provide the service to the individual which they have requested.

In terms of the other current approaches to categorisation, none of the current approaches met this either. Categorisation in relation to what the data is would definitely not meet this requirement because of the almost infinite types of personal data that a controller could process about individuals, especially in the context of online services would mean endorsing this approach would require a lot of additional information to be provided to individuals. In relation to the other current approaches to categorising personal data, they all partially meet this requirement. Categorising in relation to the degree of identifiability does not increase the amount of information that would be provided under the obligation to inform, but it does not reduce it by conveying other information just from the categories, this is also true of categorising in relation to sensitivity and source.

In relation to requirement 1(c), that the categories can be used as an anchor to which further information can be attached to contextualise the processing, this requirement is met by the new approach. Various information, such as the storage period for different data types and categories processed can be attached to them as well as the different purposes the categories are used for.

In relation to the other current approaches to categorising personal data, the approach of categorising in relation to what the data is allows for this contextualisation but the other approaches only partially meet this requirement. Categorising in relation to identifiability does allow controllers to contextualise other information, but it is arguable what benefit this give individuals in terms of transparency beyond the benefits in relation to risk analysis. Categorising in relation to sensitivity also allows other information to be attached to the categories but does not really contextualise it, because it is such a binary categorisation between special category and not special category, and whilst special categories have further detail of the categories that fall within this, non-special category personal data does not. The same is true of categorising in relation to source, where further information can be attached to these categories, it would not contextualise and increase transparency in the same way that specifying the specific personal data being processed does.

In relation to 1(d), that the categories can be used to attach different levels of rights for data subjects and obligations for data controllers. The new approach achieves this in practice and the requirement is met, both in assigning some of the rights and obligations under the current framework but also if a future framework were to be envisaged it could be organised around the risk of processing these different categories of personal data. For example, in relation to the current framework processing personal data as wisdom would be the only category that would involve the 'profiling' of individuals which is defined under Article 4(4) GDPR as:

> *"Any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements".*
>
> **General Data Protection Act (2018)**

It is outside the scope of this thesis to assign all applicable rights and obligations under the GDPR to the proposed categorisation, but an example which highlights this is the obligations and risks that the activity of Profiling attracts various additional responsibilities under the GDPR, including being one of the mandatory requirements for a Data Protection Impact Assessment (DPIA) (Article 35(3)(a) GDPR). Profiling also requires controllers to provide meaningful information about the logic involved in the automated decision-making as well as the significance and envisaged consequences of such processing for the data subject under the obligation to inform (Article 13(2)(h) and Article 14(2)(g) GDPR) and the right of access of the data subject (Article 15(1)(h)). The individual also has the right to object to profiling (Article 21(1) GDPR) where processing is based on the lawful basis's of processing that is necessary for the performance of tasks carried out in the public interest or in the exercise of an official authority vested in the controller (Article 6(1)(e) GDPR) or where processing is based on the legitimate interests of the data controller (Article 6(1)(f) GDPR). These obligations of the data controller and rights of the individual will only be in play when personal data is processed as wisdom as this is the only category where personal data is generated from the personal data.

A future version of the Framework could also organise rights and obligations under the Framework around this categorisation of personal data. As a new framework would likely change the rights and obligations that data subjects have and controllers are under which it would be impossible to predict what this may contain, so in consideration of how to re-assign the rights of data subjects and data controller's obligations under the current framework if a controller processes at the data level and puts technical measures in place to ensure that this data cannot be accessed then it could be deemed that data subjects do not have a right of access over the data either because it would require access to the data.

In terms of the other approaches to categorising personal data, all approaches meet this requirement as they allow divide personal data up and so rights and obligations can be attached to them.

In relation to requirement 2(a), which requires the categorisation to inform the individual of all of the specific personal data types that are processed about them, this requirement is met because there would be a requirement to list all the specific personal data types that

would be processed in a category. In relation to wisdom, where new data types may be generated which the controller cannot foresee, there would be an expectation that the controller updates the list of data types they process as they discover new data types. This would be in line with the current obligation to inform which when personal data is not obtained from the data subject, requires the controller to inform the individual of the information required under the obligation to inform within a reasonable period after obtaining the personal data but at the latest within one month (Article 14(3)(a)). In terms of the current approaches to categorisation, only the approach of categorising in relation to what the data is meets this requirement although arguably in practice the forms of it in practice do not meet this requirement as it is impossible to create a complete taxonomy of personal data. Categorisation in relation to sensitivity, identifiability and source does not detail all the specific personal data types that will be processed. Even in relation to sensitivity which under the GDPR, differentiates between categories of special category personal data, it does not provide the equivalent for non-special category personal data and the categories provided are still categories, rather than data types.

In relation to requirement 2(b), which requires the categorisation to distinguish between the identifiability of the individual, this requirement is partially met by the new approach. Although the new approach distinguishes between identified and identifiable individuals at some levels, it does not do this at all levels, in a way that allows an individual to know from the category of personal data alone whether it is identified or identifiable personal data that is processed about them.

The new approach also only differentiates between data belonging to ‘identified’ and ‘identifiable’ individuals, as these are the concepts of identifiability within the definition personal data under the GDPR. Arguably, as the new approach within this thesis is focused on personal data, anonymous data (as a category of data) is also an implicit within the framework because it is excluded from protection is not discussed here

All layers in the new approach include both identified and identifiable data, because the GDPR does not distinguish between situations where the data subject is ‘identified’ and where they are ‘identifiable’ in deciding whether the data relating to them is classed as personal data and thus whether processing of it falls within the scope of the GDPR.

However, the GDPR does introduce the concept of 'pseudonymisation' in Article 4(5) of the GDPR as the:

> *"…processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person".*
> **Article 4(5) General Data Protection Act (2018)**

The GDPR acknowledges that the application of pseudonymisation to personal data can reduce the risks to individuals (Recital 28) which is why, in the new approach, the consultation of pseudonymised personal data and by the controller only is within the same category as the storage of identified personal data that is not consulted by the controller. This is because in the GDPR Recital 29 states that:

> *"In order to create incentives to apply pseudonymisation when processing personal data, measures of pseudonymisation should, whilst allowing general analysis, be possible within the same controller when that controller has taken technical and organisational measures necessary to ensure, for the processing concerned, that this Regulation is implemented, and that additional information for attributing the personal data to a specific data subject is kept separately. The controller processing the personal data should indicate the authorised persons within the same controller".*
> **Article 4(5) General Data Protection Act (2018)**

The reason they are categorised as the same is because to be truly pseudonymised, personal data will have technical and organisational measures in place to prevent re-identification of the individual, and therefore within the controller's organisation the personal data of identified individuals which is not accessed and the pseudonymised personal data of individuals both present the same level of risk where technical and organisational measures are properly implemented.

If the new approach were to be used to tailor different rights and obligations and there would be more rights and obligations at the information level than at the data level, this distinction would create an incentive for controllers to pseudonymise personal data within their organisation. In particular, Article 11(1) and (2) of the GDPR state that:

> *"If the purposes for which a controller processes personal data do not or do no longer require the identification of a data subject by the controller, the controller shall not be obliged to maintain, acquire or process additional information in order to process additional information in order to identify the data subject for the sole purpose of complying with this Regulation".*
>
> **Article 11(1) General Data Protection Act (2018)**

and;

> *"Where, in cases referred to in paragraph 1 of this Article, the controller is able to demonstrate that it is not in a position to identify the data subject, the controller shall inform the data subject that it is not in a position to identify the data subject, the controller shall inform the data subject accordingly, if possible. In such cases, Article 15 to 20 shall not apply except where the data subject, for the purpose of exercising his or her rights under those articles, provides additional information enabling his or her identification"*
>
> **Article 11(2) General Data Protection Act (2018)**

The GDPR also states that:

> *"In order to create incentives to apply pseudonymisation when processing personal data, measures of pseudonymisation when processing personal data, measures of pseudonymisation should, whilst allowing general analysis, be possible within the same controller when that controller has taken technical and organisational measures necessary to ensure, for the processing concerned, that this Regulation is implemented, and that additional information for attributing the personal data to a specific data subject is kept separately. The controller processing the personal should indicate the authorised persons within the same controller'*

Recital 29 General Data Protection Act (2018)

This means that the regulation intends that analysis can be formed on pseudonymised personal data if it is done within the same controller, this is reflected in the requirement for the data level that identifiable personal data does not leave the system but also requires that only anonymous statistics (not personal data) to be produced. In the new approach, even analysis on pseudonymised would count as wisdom if it produced pseudonymised or personal data. This is because it would need to be transparent to individuals that controllers are learning more things about them, and how, to be made clear.

In tailoring different rights and obligations to the different categories presented in this thesis, it could be argued that when processing personal data as 'data', the controller would not be required to respond to data subject rights requests under Articles 15 to 20 GDPR because this would require them to move from the data level to the information level in order to identify the individual access their personal data category to action the request and defy the technical and organisational measures that they have put in place to prevent this identification. It could also be argued that this may only apply to certain rights too. Whether this reasoning would work in practice would depend on the extent of the technical and organisational measures the controller had taken to make the personal data 'identifiable' rather than 'identified'.

The category of information also distinguishes between identifiability because it only concerns 'identified' personal data. The categories of knowledge and wisdom do not differentiate between identified and identifiable personal data because both categories involve taking the data out of the primary system in which it is stored and processing it beyond consultation which increases the risk that individuals may be re-identified and be used for purposes beyond those which it was intended to be processed for, this categories reduce the control and increase the risk which reduces the distinction between identified and identifiable personal data.

In terms of the current approaches to categorising personal data, only categorising in relation to identifiability meets this requirement, the other approaches do not because none of them allow for a distinction between the identifiability of the individual.

In relation to requirement 2(c), which requires the categorisation to inform individuals of whether special category personal data is being processed about them, this requirement is met by the new approach. Whilst this is not done at the superordinate level, the approach does require that there is a link to the specific types of personal data that are processed in these categories.

In terms of the current approaches to categorising personal data, categorising in relation to sensitivity and what the personal data is (in addition to the new approach) are the only approaches that meet this requirement. Categorising in relation to identifiability does not tell the individual whether it will be special category or non-special category personal data that are processed about them. Categorising in relation to source partially meets this requirement because there may be implicit examples of when this may be the case, for example if you know that the source of personal data is that it is provided by you, you can know whether you have provided any special category personal data, however at the other levels informing of source alone does not indicate this.

In relation to requirement 2(d), which requires the categorisation to inform individuals of how a controller receives personal data in terms of whether it is provided by them or observed, derived or inferred about them, this requirement is partially met by the categorisation of the new approach. The fact that a controller operates at the wisdom level indicates that they are deriving and inferring personal data. This not met by any of the current approaches to categorising personal data other than source as depicted in Table 14.

In relation to requirement 2(e), which requires the categorisation to be exhaustive, this requirement is met by the new approach. As discussed previously in this thesis, the issue with categorising personal data by type is that you cannot be exhaustive because new types of personal data are being generated and created all the time. The new approach overcomes this, by creating the exhaustiveness at the subordinate level for certainty and then data types do not need to be exhaustive. Future work could look to create this taxonomy of personal data, but as discussed it is debatable how useful this effort would be given how new types of personal data are generated all the time.

In relation to the current approaches to categorising personal data, this requirement is only met by approaches of categorising in relation to identifiability, sensitivity and source because these approaches attempt to divide all possible personal data into categories, the approaches of categorising in relation to what the personal data is cannot categorise all personal data because of the reasons discussed in this thesis, new data types are created all the time.

In relation to requirement 2(f), which requires the categorisation to have consistent granularity, this requirement is also met by the new approach. Whilst granularity is not consistent across the whole approach, granularity is consistent at each level of the new approach.

In terms of the current approaches to transparency, this requirement is also met by categorising in relation to sensitivity and identifiability and source but is only partially met by the approaches which categorise what the personal data is because these approaches often have different granularities at the same level. For example, at the level of 'item of personal data', the MyDex White Paper '*The Case for Personal Information Empowerment: The rise of the personal data store'* (2010) in Table 12 has at the same level of 'type of personal data', 'passport number' and 'search and research results' which have different levels of granularity and specificity. Some approaches to categorising in relation to what the personal data is are more consistent, for example the approach to special category personal data under the GDPR.

In relation to requirement 3(a), which requires the categories within the categorisation to be mutually exclusive, this is met by the new approach. Whilst data types may exist in each can fall into any category at the superordinate level, when processed in relation to an individual only one category can be true. Although there may be different categories for different individuals for example, a controller may only process data in relation to some individuals but processes information or knowledge about others, yet it cannot process the same data type as data, information, knowledge and wisdom in the specific way that it is processing it.

In terms of the current approaches to categorising personal data, this is met in relation to identifiability, it is only partially met in relation to sensitivity because there is some overlap

between the categories which are used. It is not met in relation to what the personal data is because quite frequently in the categories described data types can fall into multiple ones, for example in Table 12 at the same level there are the examples of 'search and research results' and 'my plans, preferences' to which a search result of 'how much does a train to London cost next Wednesday' could be categorised as both, meaning that there are overlaps between the categories. This requirement is also only partially met in relation to source because as discussed, the distinction between the categories is often not clear and so whilst they attempt to be mutually exclusive it is not always clear which category certain personal data belongs to and so they could be categorised in relation to both at the same time.

In relation to requirement 3(b), which requires the categories to be clearly defined, including a list of the necessary and sufficient features for category membership or characteristics, this requirement is met by the new approach. The definitions provided in Table 18 allow individuals to understand which category of personal data is being processed. This requirement is also partially met by the approaches of categorising in relation to sensitivity, identifiability and source where definitions are provided but there are often inferred from what the categories are labelled as or definitions are provided but they are not clear enough to understand the distinction between the categories. The requirement is not met by categorising in relation to what the personal data is because often definitions are not provided and inferences about what the category includes from the labels alone are often not clear. For example, as Table 12 shows that the P3P Specification included the categories of 'Purchase Information' and 'Financial Information' without specifying further what would be included in these other than to provide examples of data types that would fall within these of 'Method of Payment' for 'Purchase Information' and 'Credit or Debit Card Info' for 'Financial Information' which are arguably the same types meaning that the lack of definition makes it unclear how to categorise data types.

In relation to requirement 3(c), which is that there must be a success evaluation procedure to assess whether a data type has been categorised correctly, this requirement is met by the new approach. This is provided for in the validation criteria provided for in Table 18. If the categorisation meets these criteria, then they will be confirmed to be a member of the category. This is only partially met by the current approaches to categorising personal data

because whilst a success evaluation procedure can be inferred from some of the category descriptions there, the lack of clear definitions means that this cannot be done consistently for all categories.

6.6.2 Summary

Thus, in analysing whether the new approach meets the requirements for an appropriate categorisation of personal data under the obligation to inform, this section finds that ten of the thirteen requirements identified for an appropriate categorisation in this thesis are met by the new approach as shown in Table 19. Whilst the new approach does not meet them all in practice, its meets more of them than any of the current approaches identified with categorising in relation to identifiability being the next closes.

There are three requirements which are only partially met by the new approach: that the new approach decreases the amount of information which needs to be provided under the obligation to inform, that the categorisation should differentiate between the identifiability of individuals and that the categorisation should inform individuals of how they receive the personal data. The first requirement is only partially met because further validation would be required to see whether this new approach reduces the amount of information that is provided in practice in the way it proposes; and second because the new approach does not make it clear at every level whether the personal data processed is that of identified or identifiable individuals, but does at some levels which is why this requirement is partially met. The new approach does not meet the requirement of informing individuals of the source of the personal data which is processed about them. Although operating at the wisdom level does indicate this it does not do so at all levels.

Thus, in the first stage of validation of this approach it can be concluded that the new approach meets more requirements which have been identified for an appropriate categorisation of personal data for transparency under the obligation to inform than the current approaches in practice and thus is a more appropriate approach to categorising personal data which a regulator could adopt under the obligation to inform. To support this conclusion and to explain further about how the new approach would work in practice the thesis also uses two hypothetical case studies to demonstrate the new approach in practice to support the argument that this works. These are discussed in the next subsection.

Table 19 Requirements for a Categorisation of Personal Data

Key

Met	
Partially Met	
Not Met	

Requirement Source	Requirement	Validation Criteria	New Approach	Identifiability	Sensitivity	What the Personal Data is	Source
1.Benefits of Categorisation	1(a). The categorisation enables a data subject to make a subjective assessment of any risk to their privacy involved in the processing.	This requirement enables a data subject to understand the category of personal data that is being processed and to make a subjective decision about whether there is a risk to their privacy when this category of personal data is processed. This differs from an objective assessment of risk where there would be consistent agreement among all parties as whether the processing of the category of personal data about them would present a privacy risk.					
	1(b). The categorisation reduces the amount of information that needs to be provided to individuals to increase transparency of the personal that is processed about them under the obligation to inform.	This requirement means that by identifying the category of personal data processed, other information that is required to satisfy the obligation to inform is evident and therefore less information is required to be provided to make what is known about the processing of personal data equal between the data subject and the controller.					
	1(c). The categories can be used as an anchor to which further information about	This requirement allows the controller to attach other information that is required to be provided under the obligation to inform to the category of personal data e.g. the purpose of processing the category					

Requirement Source	Requirement	Validation Criteria	New Approach	Identifiability	Sensitivity	What the Personal Data is	Source
	processing can be attached, allowing controllers to contextualise other information that they are required to provide under the obligation to inform.	of personal data, how long the personal data will be kept for and who will have access to the personal data etc. which increases the transparency of personal data processing.					
	1(d). The categories can be used to attach different levels of protection in terms of rights and obligations to them informing individuals of the categories of personal data being processed allows for the attachment of different levels of protection or obligations and rights.	Different rights of the data subjects and different obligations for data controllers that are provided for in the data protection framework can be attached to the different personal data categories					
2.Current approaches to categorisation	2(a). It informs the individual of the specific personal data types that will be processed	The categorisation should allow for a description of the specific items of personal data that are processed e.g. name, username etc.					
	2(b). The categories distinguish between the identifiability of the individual	The framework should differentiate between the identifiability of the data subject.					
	2(c). The categories allow the individual to understand whether the personal data being processed is special category personal data.	The framework should have different approaches for data that is considered sensitive because it belongs to a special category of personal data.					
	2(d). The categories inform individuals of how they receive the personal data	The categorisation should provide individuals with information about whether the personal data is provided by them or observed, derived or inferred about them.					
	2(e). Exhaustive	The categories should be exhaustive so that they cover all personal data and no future categories will need to be created.					

Requirement Source	Requirement	Validation Criteria	New Approach	Identifiability	Sensitivity	What the Personal Data is	Source
	2(f). Consistent Granularity	The categories should have the same consistent level of granularity.					
3.Classic Categorisation Theory	3(a). Mutually Exclusive	Categories should be independently different so that only one category can be identified in relation to a processing activity.					
	3(b). Clearly Defined	There is a clear definition of the category which allows controllers to easily categorise the personal data that they process and for data subjects to understand what this means. The definition should include all necessary and sufficient features or a list of characteristics which make category membership clear.					
	3(c). Is there a success evaluation procedure?	A procedure is provided which allows for a check on whether personal data has been categorised correctly and belongs to that category					

Table 20 Total No. of Requirements Met, Partially Met and Not Met

	New Approach	Identifiability	Sensitivity	What the Personal Data Is	Source
No. of Requirements Met	**10**	**6**	**4**	**5**	**3**
No. of Requirements Partially Met	**3**	**5**	**7**	**2**	**8**
No. of Requirements Not Met	**0**	**2**	**2**	**6**	**2**

6.6.3 Validation using case studies

Case Study 1 is of a hypothetical Social Networking Site and some of the ways in which it processes an individual's personal data.

Case Study 1: A SNS requires an individual to provide their name, email address and a password for the purpose of creating an account and allowing a user to log in, a user ID is also then generated. The SNS also uses cookies and similar technologies to collect the user's IP address, cookie ID, user ID, browser type, device type, location, time zone, and language of the individual. The individual can upload photos to the site which include the metadata of longitude, latitude, date and time the photo was taken only the photo itself is reviewed by SNS to make sure it does not contain offensive content. Individuals can like each other's photos and can also send messages to each other which are encrypted end to end so that only the sender and receiver can understand the data sent. Because the social networking site is provided for free, it makes money to sustain it by allowing third parties to purchase advertising space on its pages. Advertisers are often keen to direct ads based on people's sexuality which user's are often reluctant to provide and so the SNS takes the likes of individuals that have let the SNS know their sexuality to look for correlations between likes and users that have identified as homosexual, heterosexual or bi-sexual to find likely predictors of sexuality and uses this to classify the user's that have not voluntarily provided their sexuality. The SNS it shares the user's IP address, cookie ID, user ID, browser type, device type, location, time zone, language and sexuality with third parties in the form of a 'bid request' via a third party platform and the highest bidding advertiser's advert is then served to the individual on the SNS.

In relation to Case Study 1, the SNS would need to inform individuals that it processes non-special category personal data as data, information and that it processes Special Category Personal Data as Wisdom.

It processes name, email address and password as data because these are stored in the system but are only used to authenticate the individual to log in, the messages the individual sends are also only processed as data because although they are stored on the

SNS cloud storage facility but they are encrypted end to end to the SNS cannot view them. If the individual contacts the SNS then they may view the name and email address to confirm the individual's identity in which case they will be processing it as information but they do not do this by default.

The SNS processes IP address, cookie ID, user ID, browser type, device type, location, time zone, and language as knowledge because this is not used to look for correlations for inferences of sexuality but it is not just consulted by the controller as it leaves the primary system of storage and is shared with third parties.

If the SNS was not using the likes of the photos as wisdom then they would be processing them, and the metadata associated with them as information or data depending on whether they just consulted the photo of the individual as part of a review for offensive content and whether this was done in a pseudonymised way or not. However, because the SNS uses the photos and the likes on them to infer an individual's sexuality they will be using processing Special Category Personal Data as wisdom and would likely process the metadata in this because of the potential for more accurate inferences with more data points included in the search for correlations.

Case Study 2: An individual uses an internet connected heating system which allows them to remotely control their heating at home. They purchase the system and register their warranty with the company by filling out an online form who creates a record of the customer which includes their name, address, email address, phone number and product ID. A User ID is also generated which links this record to the individual's product in the company's cloud storage system. The heating system comes with a controller which can be used at home but the individual also downloads an app to their smart phone to remotely control the system. The app collects data some data that would be classed as personal data it were were handled by the controller, but is used purely for the functionality of the app, is stored on the smart phone only and is not accessed by the controller. When an individual wants to control the heating remotely, they use the app and a message is sent from the app to the company's cloud storage facility, in which the message is processed and then sent to the individual's product to complete the request. For example, the individual is at work and is leaving later than normal and so they use the app on their phone to change the time that the heating will come on from 18:00 to 19:00. This sends a message to the cloud

including the individual's IP Address, User ID, timestamp of the request and the action which the product needs to perform which is processed and sent to the heating system and changes the time which the heating will come on. The company stores the messages in the cloud and only uses them to generate anonymous statistics for trouble shooting and market research purposes. The only time the messages of an individual's machine will be looked at in isolation is if they contact the company by phone or by email because there is an issue with their product. A customer service agent will then ask the individual to confirm three personal data types to validate their identity and then look at the messages they have received in relation to the product to advise the individual how to repair it. Notes in relation to the call, internet chat or email exchange are then stored on the customers record and the call, chat history or emails are also stored but only ever consulted and do not leave the system on which they are stored. The company sells various connected products and the business model of the company is to make money from the cost of the products that it sells. The marketing department requests access to the customer record and data stored in the cloud so that they can understand new things about customers. They want to use the addresses of the individuals and specifically their postcode to profile them by appending the average house price of their postcode area to their customer record to categorise them as rich and poor so that they can directly market to those classified as 'rich'. They also want to analyse all of the information in the cloud from all their different connected products to look for correlations in terms of use which indicate whether an owner will be classed as rich or poor but also if there are any other interesting correlations between users of different products which may allow them to predict more things about users in the future.

In relation to Case Study 2, the way the company is currently operating they would need to advise under the obligation to inform that they process non-special category personal data as data about individuals and that they will only process this as information about where the individual contacts them and requests that they do so e.g. when they call to speak to a customer service agent. The controller will also need to provide a link to the data types which will be name, address, email address, phone number product ID, user ID, IP Address, timestamp of the request and the list of specific actions which the individual can make the product perform remotely e.g. "change heating on time".

The controller operates at this level because the company does not access the personal data generated by the app, the data generated by use of the product is sent to the cloud

or the customer's data record as a matter of routine, and only anonymous data is produced from this (opposed to personal data that would put it in the wisdom category).

Whilst the generation of notes in relation to the individual's call or email exchange could be seen as wisdom because it is the generation of new personal data, because this is provided by the individual instead of inferred or derived by the controller and they are aware of this it falls outside this category and would still be classed as information or special category information if it included special category personal data.

If the request from the marketing department were granted then the company would need to inform individuals that it processes data as data, information, knowledge and wisdom. The act of using the average house price of a customer's postcode area to categorise them as rich and poor would take them to the wisdom level as it is using basic maths of an average house price being above or below a certain threshold to classify this about an individual and whether it is true or not will not matter. The company would also need to provide a link to say that they process name, postcode, average house price of your postcode area and whether you are rich or poor at this level.

The act of analysing all of the information in the cloud from all their different connected products to look for correlations in terms of use which indicate whether an owner will be classed as rich or poor but also if there are any other interesting correlations between users of different products which may allow them to predict more things about users in the future which would also take them to the wisdom level and they would need to advise that and provide a link to the data types they put into the analysis and the personal data that they derive or infer. As the controller derives or infers new information it should update this link with the types that they find and if any of these data types are special category personal data they will need to advise individuals that they now process this category too.

6.6.4 Summary of case study validation

These two hypothetical case studies have demonstrated how the new approach to categorising personal data can be applied to the complexities of personal data processing in the context of SNS and online services and validates the ability of the new approach to be used to inform individuals about the personal data that is being processed. The case

studies deliberately included a range of different ways the services process personal data to strengthen the conclusion that the new approach can be used in relation to personal data processing beyond just these examples here and at least to all online services.

6.7 Conclusion and Limitations

The research goal of this thesis was to understand deficiencies in the current approaches to the transparency of personal data processing in the context of the obligation to inform in practice and to propose an improvement to the way organisations can be transparent about their personal data processing. This chapter presented the proposed improvement to the way organisations can be transparent about their personal data processing by providing a new approach to categorising personal data which controllers can use to provide individuals with information about the personal data that it processes. The ability of the new approach to increase the transparency of personal processing was validated through two methods, requirements analysis and case study methodology.

Analysis from various points within this thesis are used to provide a list of requirements for an appropriate categorisation of personal data for the purposes of transparency against which the new approach is validated. These include the benefits for transparency that an appropriate categorisation can have, the themes in the current approaches to categorising personal data (under the assumption that these are good traits to have) and the requirements for a classical categorisation approach which is the only approach which would lead to a legally certain categorisation of personal data. The Chapter found that whilst the new approach does not meet all the requirements, it does meet more than any of the other approaches that are there in practice because it has had the benefit of learning the strengths and weaknesses of the current approaches.

Two case studies were then used to validate the new approach by explaining how it would apply in practice and navigate the complexities of personal data processing in the real world. The two cases were a hypothetical SNS and a connected product. SNS was chosen because this was the example that has used throughout this thesis, a connected product was also chosen as another example of a technology that has created serious concerns for privacy and to contrast with the findings in relation to SNS to support the conclusion that this approach could be applied beyond SNS alone. These case studies validate that the approach could be applied to different processing scenarios.

There are some limitations in this approach. First, there may be more recommendations that are required for an appropriate categorisation of personal data in relation to transparency. The requirements used here were not intended to be exhaustive and constructing an exhaustive list of requirements for increasing transparency under the obligation to inform is outside the scope of this thesis and this could form part of future work that would be required before the approach is formally adopted by a regulator . However, despite not being exhaustive the recommendations do provide a good starting point for indicating the value of this approach, especially in the context of other current approaches to categorising personal data.

A second limitation is that there may be other approaches to categorising personal data which have not been examined here and which may also meet the requirements which the new approach does. It was outside the scope of this thesis to systematically look for all the approaches to categorising personal data and instead only included the categorisations which were found in examining the literature on privacy and the obligation to inform. Future work could look to do more of a systematic search for categorisations of personal data to validate that this approach is the best one.

A third limitation of this study is that it only validated that SNS and connected products which does not validate that it can be used in relation to all services. To overcome this future work could look to apply this to more, and more detailed case studies which would likely be required as part of a guide to explain how the approach works which would be needed if the categorisation were to be adopted by a regulator.

A fourth limitation is the fact that the approach only uses theoretical approaches to validate the success of the new approach in increasing the transparency of personal data under the obligation to inform. Prior to adoption some empirical methods should be used to validate that this does increase the transparency of personal data processing in practice and this could also form part of future work. For example, a laboratory study could be used to validate that compared to current approaches or different approaches to categorising personal data, the new approach presented within this thesis increases the ability of individuals to understand what personal data a controller is processing about them and how they are doing this.

Chapter 7 Conclusion

Increasing advancements in technology have created a threat to the privacy of individuals. To some extent data protection laws have been introduced to counter this threat, of which transparency is a key principle so that individuals understand how their personal data is being processed by organisations.

One of the obligations individuals have in relation to transparency is to provide individuals with certain information about how they process their personal data, which has led to the adoption of privacy policies as the de facto method of compliance with the obligation. These have been long criticised for their inability to make the processing of personal data transparent for individuals which is why there is a large amount of research in the Human Computer Interaction Community of Computer Science on how to improve them.

Despite this research, not a lot has changed in practice and generally organisations still rely on privacy policies to inform individuals of this information. This interdisciplinary thesis uses the knowledge, strengths and methods of Computer Science and Law to approach the research goal of understanding the deficiencies in the current approaches to the transparency of personal data processing in the context of the obligation to inform in practice and to propose an improvement to the way organisations can be transparent about their personal data processing. This research aimed to bridge these gaps by looking at the phenomenon of privacy policies in more detail to try and understand what may be preventing the ability of privacy policies to make personal data processing transparent and to propose an improvement to this. To do this, three research questions were created:

RQ1: Are the privacy policies of Social Networking Sites (SNS) similar enough in the information they provide about their personal data processing for the standardization of Privacy Policies to be possible?

RQ2: When is there a legal requirement in the EU and UK to provide information about the specific personal data being processed and in the context of the obligation to inform and what is the requirement for this?

RQ3: Does various current approaches to categorising personal data and informing individuals of these achieve the aims of transparency under the European Union Data Protection Framework?

RQ4: Can the DIKW model be used to increase the transparency of personal data processing in relation to the obligation to inform under the European Union Data Protection Framework?

This chapter presents a summary of the investigations performed as part of this research together with the findings that emerged. The contributions made by this thesis are charted, leading to recommendations and proposals for future work.

7.1 Findings Summary

As discussed, there were four distinct phases of this research and the findings from each phase fed into and shaped the subsequent ones.

7.1.1 Understanding More About Privacy Policies

Firstly, it was important to understand more about privacy policies in practice. To do this a preliminary study was designed to understand more about privacy policies themselves. It did this through the lens of investigating whether the privacy policies of SNS were similar enough for the creation of a standardised privacy policy to be possible. This led to the creation of RQ1, which led to five Research Sub Questions:

RSQ 1. What is the similarity between the privacy policies of the top six SNS globally, in the clauses that they use?

RSQ 2. What is the similarity between the privacy policies of the top six SNS globally, in their coverage of forty recommendations of information to include in a privacy policy made by the UK Information Commissioners Office (ICO) in their Code of Practice (ICO Code)?

RSQ 3. Are there any recommendations of the ICO Code, which all privacy policies do not address?

RSQ 4. Are there any themes of information addressed in all of the privacy policies that were not included in the forty recommendations from the ICO Code?

RSQ 5. To what extent is standardization possible between the privacy policies of SNS?

The preliminary study discussed in this chapter had two aims. First, to understand whether privacy policies are similar enough in practice for standardisation of them to be possible, and second, to understand more about the phenomena that are 'privacy policies', with a focus on the information that they contain. In relation to the first aim, the study found that the privacy policies of SNS demonstrated homogeneity and promising potential for standardization, albeit at a thematic, rather than clause level. Five recommendations were then made to support achieving this in practice. In relation to the second aim, the study found that there was a number of ICO Code Recommendations for transparency which were not addressed by the privacy policies. It also found that there were a number of themes of information that all of the privacy policies included, but that went beyond the recommendations of the ICO Code. Given the overall research goal of this thesis, there were many ways the findings from this study could have been taken forward.

In relation to the second aim, the results from the study highlighted something of interest and in particular questioned two implicit assumptions that had been made going into it. First, was that ICO's recommendations for a compliant privacy policy would provide a full list of the information required to make data processing transparent; and second, that that the problem with increasing the transparency of processing lies simply in improving the communication of the information that privacy policies contain. Therefore, whilst the findings in relation to either aim of this preliminary study could be pursued further, if the problems raised by second aim were left unresolved, they could stand in the way of the creation of a transparent, standardized, and legally compliant privacy policy. To research all of the problems raised would be a gargantuan task, far beyond the scope of a single PhD and therefore the most surprising one was chosen. The work undertaken in the next chapter was devised to continue to meet the research goal of the thesis. In particular, it focuses on whether there is a legal requirement in the EU and UK, for organisations to provide information about the specific personal data that they are processing and whether this increases the transparency of personal data processing in practice.

7.1.2 The Legal Requirement to Provide Information on the Personal Data Processed

Given the unexpected findings of the preliminary study, it was important to devise an investigation to look into these further which led to the creation of RQ2:

When is there a legal requirement in the EU and UK to provide information about the specific personal data being processed under the obligation to inform and what is the requirement for this?

In answer to RQ2, this investigation found that, when referred to, the requirement to provide information about personal data is to specify the 'categories of data' that are processed. It also found that whilst specifying what the categories of 'special category personal data' are, the law does not specify what a categorisation of 'non-special category personal data' should be. Furthermore, it found that both the previous and current execution of the obligation to inform in the EU and UK make it difficult for data controllers to understand exactly when they are under an obligation to inform individuals of the categories of data that they process about them. This lack of clarity in the law under the Data Protection Directive (DPD) meant that it was not necessarily a deficiency in the ICO Code that it did not mention providing this information as a recommendation, as it was not stated within the DPA 1998 and was not necessarily an obligation in all scenarios. Yet, as ICO advised Google that in 2015 they needed to provide individuals with 'an exhaustive list of the types of data processed', it should have been included, even as an example of something that may be required for processing to be fair.

The investigation found that whilst the introduction of the GDPR provided some clarity, by making this requirement mandatory when personal data is not obtained from the individual, it did not clarify when organisations are under an obligation to do this when they do obtain the personal data from them. Unfortunately, guidance from European data protection bodies has also not provided clarity, and in particular, the Article 29 Working Party has been inconsistent in the terminology it has used to describe what exactly what needs to be provided under this obligation, without clarifying what these terms mean, and whether they are equivalent.

In contrast to the GDPR, the introduction of the DPA 2018 in the UK has provided clarity on when controllers are under an obligation to provide this information, by requiring it in all circumstances. However, guidance from the UK Regulator contradicts this position by advising that it is not always a requirement. This lack of clarity means that it is unclear for controllers when they are under a requirement to provide this information.

Therefore, this thesis found in answer the RQ2 that the law in the UK and the EU is unclear on when data controllers are under an obligation to provide individuals with information on the personal data they process under the obligation to inform and even when it is clear, it is unclear what this requirement entails.

7.1.3 Categorisation in Practice

Given the RQ2 found that the law was unclear on when a controller is under a requirement to provide information on personal data they process and what this requirement is, it seemed obvious that regulators needed to provide clarity on the requirement. However, it was important to understand what the recommendation should be both in terms of when and how. This led to RQ3:

RQ3: Do various current approaches to categorising personal data and informing individuals of these achieve the aims of transparency under the European Union Data Protection Framework?

Because of the breadth of RQ3, four RSQ's were created:

RSQ1: What are the benefits of categorising personal data and does informing individuals of this increase the transparency of personal data processing?

RSQ2: Is there a consistent approach to categorising 'non- special category personal data' in practice?

RSQ3: Are any of the current approaches to categorising personal data in practice sufficient in making the processing of personal data transparent?

RSQ4: How are SNS categorising personal data and informing individuals of this in practice and does this achieve the benefits for transparency?

In answering RSQ1, the investigation found that this lack of clarity reduces the transparency of personal data processing, because there are many benefits for transparency by categorising personal data and informing individuals of which ones of these are processed. This suggests that European Regulators should confirm that there is a requirement for controllers to always specify the categories of personal data that they are processing and amend their guidance to reflect this. However, to do so they will also need to provide guidance on how to categorise 'non-special category personal data' which is why it was important to understand whether a consistent approach to categorising personal data has emerged.

In answering RSQ2, the investigation found that despite the lack of direction from the law on what a category of non-special category personal data is, different parties have interpreted their own approaches to categorising personal data, but in different ways. This means that a consistent approach to categorising non-special category personal data has not organically emerged under the obligation to inform. RSQ3 found that none of these approaches discovered are sufficient in practice at making the processing of personal data transparent to a level that equates the information available to subjects to that possessed by data controllers and achieves the benefits of transparency which an appropriate categorisation can create.

As it was the behaviour of the SNS that were investigated in the preliminary study of discussing the specific personal data they processed, RSQ4 looked to understand how SNS were doing this in practice, as this may inform the creation of a new categorisation of non-special category personal data. However, it found that the approaches of the SNS also did not achieve the benefits for transparency of an appropriate categorisation of personal data. This finding increases the importance for regulators to confirm how to categorise 'non-special category personal data', because even if they do not confirm that they must **always** be provided, it will be a requirement in some situations to do so, and controllers do not have the guidance they need to understand how to do this. This lack of clarity reduces the transparency of personal data processing and the answers to RSQ4 and RSQ5 mean that

there is not an approach to categorising personal data that can be readily adopted in practice by the regulator.

These findings suggested that this is an area where improvements can be made to the way data controllers are transparent about their personal data processing in accordance with the goal of this thesis. In particular, they suggested that a new categorisation of personal data is required in order to increase the transparency of processing which led to the final phase of this research.

7.1.4 A New Approach to Categorising Personal Data

Given the need for an appropriate categorisation of personal data to increase transparency under the obligation to inform, this thesis aims to bridge that gap and provide an approach which can be used. To do this, it created RQ4:

Can the DIKW model be used to increase the transparency of personal data processing in relation to the obligation to inform under the European Union Data Protection Framework?

To answer this question, this thesis reviewed the literature on how to create a good categorisation of something in general and looked at how both humans and computers do this to understand the requirements for a good categorisation. It also reviewed the literature on the DIKW Hierarchy and how it has been linked to personal data previously to understand how it can be adapted to create a categorisation of personal data. The findings were that the DIKW hierarchy could be adapted and used to categorise personal data to increase the transparency of personal data processing under the obligation to inform. This was validated using two methods of requirement analysis and case study methodology which evidenced that the new approach meets more of the requirements for an appropriate categorisation of personal data than the current approaches identified.

7.2 Contributions

In view of the findings presented in this work, this thesis makes four key contributions to the field of privacy and transparency of personal data processing.

1. It makes a contribution to computer science by highlighting the current assumption of HCI research on transparency, that transparency is a problem solved by improving communication of the information alone. If work continues to be based on this assumption the transparency of data processing without more research into the actual information individuals need for processing to be transparent, the research on it will eventually reach a glass ceiling.

2. It makes a contribution to law by highlighting the fact that the law in the EU and UK is unclear on whether (and how) organisations need to inform individuals about the categories of personal data that they process, but that this should be a legal requirement for data processing to be transparent

3. It makes a contribution to computer science and law by proposing a new approach to categorising personal data under the obligation to inform, using knowledge from both computer science and law. Not only does this new approach support the creation of an appropriate 'obligation to inform' requirement that makes data processing more transparent, it also provides an approach upon which various system designs can be based e.g. autonomous consent agents, personal data stores and privacy policy generators. This supports other efforts in HCI on improving the transparency of personal data processing.

4. Whilst all of these contributions can also be seen as web science contributions, the overall impact in making online data processing more transparent contributes to web science by contributing to the appropriate governance of the web which supports and ensures its continued use.

7.3 Limitations, Recommendations and Future Work

Due to the scope of this research and the time available, there are naturally a number of limitations to this work.

First, this did not focus on all legislation within the Framework that look at Data Protection, but chose to focus on the most prominent ones. This may mean that there is guidance in the Framework in other legislation that supports this but that was not viewed. The reason

for this is that the thesis was focused on the obligation to inform and thus chose to focus on the legislation in which this obligation/right is created.

Second, a systematic review of all of the different proposals/approaches for categorising personal data was not conducted and therefore the sample reviewed is smaller than the researcher would have liked to indicate that a new approach was required. Further work could be done to look at the approaches being taken in practice and to incorporate the benefits of their approaches into the new approach.

Third, this categorisation of personal data focuses on the obligation to inform but specifying the personal data being processed underpins various compliance requirements within the Framework. Therefore, further work could be done on understanding how the new approach would apply to and Data Controllers and other stakeholders in complying with these compliance requirements.

Fourth, the case studies to validate the new approach had to be very limited and arguably simplistic due to the time available in which to complete this stage of the research. Therefore, the case studies are not intended to be comprehensive but more to be an illustrative example of how the new approach could be applied. With more time I would have liked to have created more complex case studies and future work could be used to create these and elaborate on how the new approach would apply to these more complex, and potentially more diverse examples further.

Research into understanding how to make the processing of personal data transparent is a broad and complex area of research, from the findings of this thesis a few recommendations are made to support increasing the transparency of personal data processing.

1: Data Controllers should be obligated to inform individuals of the categories and types of personal data within these that they process about them. Initially guidance should be provided by regulators that this is a requirement, and at some point the GDPR should be amended to require this is or a law that replaces it should require this. Case law could also

be used to make this requirement in certain circumstances. ICO should amend their guidance to make sure that it is consistent with this.

2: Further work should be done to validate the new approach. This includes a wider review of the different categorisations of personal data in practice to make sure the new approach presented is still the most appropriate, a search for any more requirements that categorisations should be tested against to conclude that they are appropriate categorisations for increasing the transparency of personal data processing under the obligation to inform and empirical validation of the ability of the approach to increase the transparency of personal data processing in practice.

3: The production of guidance by a regulator for controllers and data subjects on how to categorise personal data and inform on these. In confirming that this is a requirement, regulators should also produce accompanying guidance on how to categorise personal data and provide information on this in practice so that it is clear for controllers what obligation they are under and individuals are clear what they are expected to provide.

This thesis has indicated a number of possible directions where future work could build upon this exploratory research. The work on and recommendations made in relation to the standardisation of privacy policies in the preliminary study discussed in Chapter 3 could be pursued. As discussed, it was not pursued in this thesis, but the recommendations made in Section 3.5.2 could be implemented to work towards having standardised privacy policies. Further validation on the new approach could be used to justify its potential to increase transparency and overcome some of the limitations with the approach of this thesis. This could include empirical validation and the application to more hypothetical cases. The work could also be built upon by examining other information that needs to be provided under the obligation to inform to increase the transparency of personal data processing, in addition to the categories of personal data.

4: Further work could look at how this new approach to categorisation could support in the design of systems that comply with policies and regulations. The issue of designing technologies that comply with regulations and laws has long been recognised in the HCI field (Goodman and Flaxman, 2016; Greengard, 2018) and Krebs, et. al (2019) suggests that the intersection of law and HCI proposes a formulation of adequate design guidelines to

comply with regulatory policies such as GDPR requirements in terms of transparency, accountability and data protection by design while maximizing usability and user experience. In particular in the understanding of the extent of the right that individuals have to access personal data about them, and information and explanation about the algorithms that are used to generate this information, an ongoing debate in policy, industry and academia (Selbst and Powles, 2018). Therefore, further work could be done to understand how this categorisation of personal data could be used as a mediatory object between system designers and legal experts that allows for the understanding of how different categories and types of personal data should be classed and used within a system.

5: Further work could look at how this new approach could be used as a mediatory object in law. As discussed previously, there is an ongoing debate in policy, industry and academia (Selbst and Powles, 2018) about the extent to which individuals are entitled to access to the personal data that is derived and inferred about them and the logic of the algorithms that produce this. Whilst it was outside the scope of this thesis to suggest what the legal approach should be or how this categorisation applies to this debate, this categorisation approach could be used as a way of making it clear to Data Contollers the personal data that they are expected to provide under the Obligation to Inform.

7.4 Concluding Statement

Increasing advancements in technology have the potential to threaten the privacy of individuals by reducing the control individuals have over the personal data that is processed about them. Making sure individuals understand what personal data organisations process about them (and how) is key to individuals being able to control how personal data about them is processed and to increase the transparency of personal data processing. Adopting the approach to categorising personal data presented in this thesis and requiring controllers to specify the categories and types of personal data that they process can increase the transparency of personal data processing and thus support countering the threat to privacy which increasing technologization poses.

Appendix A

Themes and Codes for Thematic Analysis

A.1 Table of ICO Code Recommendations, Their Definition for Thematic Coding and an Example of a Coded Clause

Theme	ICO Code Recommendation	Definition for Thematic Coding	Example of Clause
1	Tell people who you are/ The identity of the organisation in control of the processing/ who is collecting the information (5, 8, 9, 10)	The privacy policy states its name of the company and provides a physical contact address.	If you have questions or complaints regarding our Data Use Policy or practices, please contact us by mail at 1601 Willow Road, Menlo Park, CA 94025 if you reside in the U.S. or Canada, or at Facebook Ireland Ltd., Hanover Reach, 5-7 Hanover Quay, Dublin 2 Ireland if you live outside the U.S. or Canada **Facebook**
2	The purpose for obtaining the information/ Be clear why you need the information/ why you're collecting the information (8,9)*	The specific purpose for obtaining the personal data is described in the privacy policy.	**"phone number"** For example, if you add a phone number as a recovery option, if you forget your password Google can send you a text message with a code to enable you to reset it. **Google+**
3	Tell people the purpose for using the information.	The privacy policy tells the user the purpose for using the information.	If you email us, we may keep your message, email address, and contact information) to respond to your request **Twitter**
4	The purpose for disclosing the information/ details of how the organisations they pass it onto will use the information (8, 10)*	The privacy policy describes when they will pass their data to third parties and what the third party will do with this.	If the delivery of incentives requires your contact information, you may be asked to provide personal information to the third party fulfilling the incentive offer, which will be used only for the

Theme	ICO Code Recommendation	Definition for Thematic Coding	Example of Clause
			purpose of delivering incentives and verifying your contact information. **LinkedIn**
5	Tell people who their information will be shared with/disclosed to	The privacy policy describes which third parties users information will be shared with/disclosed to.	"Secret boards are visible to you and other participants in the board, and any participant may choose to make the contents of the board available to anyone else." **Pinterest**
6	Provide people with information about people's rights of access to their data/ About their rights and how they can exercise them *e.g. obtain a copy of their personal information or object to direct marketing* (5, 10)	The privacy policy describes the rights that individuals have over their personal data.	You have a right to (1) access, modify, correct, or delete your personal information controlled by LinkedIn regarding your profile, **LinkedIn**
7	Your arrangements for keeping the data secure/ What you are going to do to ensure the security of personal information (5, 10)	The privacy policy describes how the organisation will keep the user's personal data secure.	We take privacy and security seriously and have enabled HTTPS access to our site (turn on HTTPS), in addition to existing SSL access over mobile devices. **LinkedIn**
8	Try and predict whether you will be likely to do things with it in future without drawing up a long list of future possible uses if you are unlikely to use it for those purposes (9)	The privacy policy describes future purposes it may use the personal data it has collected for.	N/A – Not addressed by any of the privacy policies.
9	If people have a choice over whether	The privacy policy describes where there is a	Most mobile devices allow you to prevent real

Theme	ICO Code Recommendation	Definition for Thematic Coding	Example of Clause
	to provide information it should be properly explained to them/ they are given a proper opportunity to exercise it (9, 11)	choice of whether to provide personal data and individuals can exercise this.	time location data being sent to LinkedIn, and of course LinkedIn will honor your settings. **LinkedIn**
10	Where people have a choice over the use of information, it should be properly explained to them/ they are given a proper opportunity to exercise it (9, 11)	The privacy policy describes where there is a choice over use of personal data and individuals can exercise this.	If you wish to not receive targeted ads from most third party companies, you may opt-out by, as applicable, clicking on the AdChoice icon or "Ads by LinkedIn" link in or next to the ad or by visiting http://preferences-mgr.truste.com/ and http://www.networkadvertising.org/managing/opt_out.asp. **LinkedIn**
11	The implications/effects of collecting/ using/ disclosing the information (9, 4)	The privacy policy describes the effects on the individual of how they process personal data.	We also get technical information when you use our products or use websites or apps that have Pinterest features. These days, whenever you use a website, mobile application, or other Internet service, there's certain information that almost always gets created and recorded automatically. The same is true when you use our products. In addition to log data, we may also collect information about the device you're using Pinterest on, including what type of device it is, what operating system you're using, device settings, unique device identifiers, and crash data. **Pinterest**

Appendix A

Theme	ICO Code Recommendation	Definition for Thematic Coding	Example of Clause
12	Tell people how long you or other organisations intend to keep the data.	The privacy policy refers to how long it (or other organisations it shares the data with) intend to keep data for.	Typically, information associated with your account will be kept until your account is deleted". **Facebook**
13	Whether the information will be transferred overseas (10)	The privacy policy advises where in the world personal data will be transferred.	Yahoo may transfer your personal information for the general purposes set out above to any Yahoo group company worldwide, and they may use your personal information as set forth below. **Flickr**
14	Who to contact if they want to complain or know about how their information will be used (10)	The privacy policy provide contact details for the organiser processing the details.	If you have questions or concerns regarding this Policy, you should first contact LinkedIn. If contacting us does not resolve your complaint, you may raise your complaint with TRUSTe by Internet, by fax at 415-520-3420, or mail to TRUSTe Safe Harbor Compliance Dept. **LinkedIn**
15	About the right to complain to the Information Commissioner if there is a problem (10)	The privacy policies advise individuals that they have the right to complain to the UK Regulator, the Information Commissioner's Office.	N/A – Not addressed by any of the privacy policies
16	Different notices aimed at different groups of people it deals with (10)	There is more than one type privacy policy.	N/A – Not addressed by any of the privacy policies
17	Positive agreement required if previously collected information is used in a significantly	The privacy policy states that it will notify and capture a positive agreement from individuals if it will use	We will ask for your consent before using information for a purpose other than those that are

Theme	ICO Code Recommendation	Definition for Thematic Coding	Example of Clause
	different way/ If using information you already hold in a new way, you should actively seek their consent (11, 13)	previously collected personal data in a significantly different way.	set out in this Privacy Policy. **Google+**
18	Where the collection and use of information is essential to provide the service or carry out the transaction, even if individuals have no real choice, the collection of information about them still needs to be fair and transparent (11).	Privacy policies describe in detail what personal data is mandatory for use of a particular service that is offered.	To create an account on LinkedIn, you must provide us with at least your name, email address and/or mobile number, and a password. **LinkedIn**
19	Where individuals are required by law to provide their personal details be open with people and explain clearly why their information is being collected and what it will be used for (11)	The privacy policy explains the situations in which individuals will be specifically required to provide their personal data for a legal reason and details what this legal reason is.	N/A – Not addressed by any of the privacy policies
20	If you intend to market people by electronic means (email, SMS, fax or telephone etc) special rules may apply and you may need their permission before doing so *Electronic Communications Regulations 2003* (11)	The privacy policy explains when it will capture permission for before marketing to people.	These companies may use such information to help Yahoo communicate with you (to the extent consented by you) about offers from Yahoo and our marketing partners. **Flickr**

Theme	ICO Code Recommendation	Definition for Thematic Coding	Example of Clause
21	If you are collecting sensitive information you should actively tell them about it and gain positive agreement. *To actively communicate means to take positive action to provide a privacy notice, which differs from having a privacy notice available for members of the public who want to see it i.e. having to click on a web link* (11, 13, 4).	The privacy policy details when it will collect sensitive data and confirms that it will require an affirmative action for this processing to occur.	We require opt in consent for the sharing of any sensitive personal information. **Google+**
22	If providing or not providing personal information will have a significant effect on the individual you should actively tell them about it/ Consequences of not providing information *e.g. non-receipt of a benefit. To actively communicate means to take positive action to provide a privacy notice, which differs from having a privacy notice available for members of the public who want to see it i.e. having to click on a web link* (10, 13)	The privacy policy describes actions that will impact the ability to use of the service.	When your account is deactivated, it is not viewable on Twitter.com. **Twitter**

Theme	ICO Code Recommendation	Definition for Thematic Coding	Example of Clause
23	Obtain assurances (in form of written agreements) from any organisations you share personal information with about what they will do with the information and what the effect on people is likely to be (14)	The privacy policy confirms that it has robust agreements in place with third parties it shares information with.	These third-party developers have either negotiated an agreement to use LinkedIn platform technology or have agreed to our self-service API and Plugin terms in order to build applications ("Platform Applications")'. **LinkedIn**
24	Organisations that intend to combine information from different sources should explain this (14)	The privacy policy explains how they combine information.	Yahoo may combine information (including personally identifiable information) about you that we have with information we obtain from business partners or other companies. **Flickr**
25	and the likely consequences (14)	The privacy policy describes what will happen because information is combined.	We may also put together data about you to serve you ads or other content that might be more relevant to you. **LinkedIn**
26	In marketing contexts, when organisations ask for permission to share customer information with third parties e.g. companies in the same group, this should be backed up with more detail information such as the names of the companies involved for those who want it (14)	The privacy policy names the specific third parties that it shares personal data with.	N/A – Not addressed by any of the privacy policies
27	If an organisation intends to collect personal	The privacy policy confirms whether it sells or rents information and	N/A – Not addressed by any of the privacy policies

Theme	ICO Code Recommendation	Definition for Thematic Coding	Example of Clause
	information with the intention of selling or renting it you should make it clear to individuals that the information they provide could be supplied to anyone and used for any purpose and tell them this when they provide their details (14)	makes it clear that the information they provide could be sold or rented to anyone for any purpose.	
28	Individuals are told that their information may be passed onto other organisations. When a business is insolvent, bankrupt, being closed down or sold the database may be sold on (14)	The privacy policy states that if the business is insolvent, bankrupt, being closed down or sold the database may be sold on.	In the event that Twitter is involved in a bankruptcy, merger, acquisition, reorganization or sale of assets, your information may be sold or transferred as part of that transaction. **Twitter**
29	That if their information is rented, individuals are told that if the business is insolvent, bankrupt, being closed down or sold that their information will be returned to its owner (14)	The privacy policy confirms whether it rents information and that if the business is insolvent, bankrupt, being closed down or sold that their information will be returned to its owner.	N/A – Not addressed by any of the privacy policies
30	That if either of the above are true, the seller will make sure the information will only be used for the same or similar purpose and consent will be required for any new purposes (14)	The privacy confirms that if the business is insolvent, bankrupt, being closed down or sold that the personal data of the individual will only be used for the same or similar purposes and that consent will be required to use it for any new purposes.	But they will still have to honor the commitments we have made in this Data Use Policy. **Facebook**

Theme	ICO Code Recommendation	Definition for Thematic Coding	Example of Clause
31	Avoid using confusing terminology e.g. technical language (16, 4)	The privacy policy does not contain any technical terms unless it is strictly necessary to do so and an explanation is provided.	N/A – Not addressed by any of the privacy policies
32	Avoid using legalistic language (16)	The privacy policy does not contain any legal terms unless it is strictly necessary to do so and an explanation is provided.	If the ownership of our business changes, we may transfer your information to the new owner so they can continue to operate the service. **Facebook**
33	Don't give the impression that people have a choice when in reality they do not (11, 16)	The privacy policy is clear when an individual does not have a choice about processing.	*(LinkedIn may retain your personal information even after you have closed your account if retention is reasonably necessary to)* meet regulatory requirements, resolve disputes between Members, prevent fraud and abuse, or enforce this Privacy Policy and our User Agreement. **LinkedIn**
34	Make sure they comply with any relevant sectoral rules (17)	The privacy policy confirms that the organisation complies with relevant sectoral rules.	We also adhere to several self-regulatory frameworks. **Google+**
35	Good practice to provide the privacy notice in the same medium you use to collect the data (17)	The privacy policy is accessible online.	N/A
36	Is a layered notice used? Because the privacy notice is online, the organisation should make full use of the technology available to them *(13)* (18)	The privacy policy is provided in layers.	N/A

Appendix A

Theme	ICO Code Recommendation	Definition for Thematic Coding	Example of Clause
37	If you collect information from vulnerable individuals (such as children) have appropriate privacy notice to their level of understanding – *would they understand the consequences? (*18,19)	There are different versions of the privacy policy for vulnerable individuals.	N/A – Not addressed by any of the privacy policies
38	Good practice to provide your privacy notice in the language that your intended audience is most likely to understand, even if not required by law (19)	The privacy policy is provided in different languages.	N/A
39	Good practice to keep your privacy notice under regular review (19)	The privacy policy confirms that it will be reviewed regularly.	Yahoo may amend this policy from time to time. **Flickr**
40	Good practice to review the effectiveness of your notice by analysing complaints from the general public about your information use in general and your privacy notice in particular (19)	The privacy policy confirms that they accept feedback on the privacy policy.	If you have questions or suggestions complete a feedback form. **Flickr**

Appendix B Results of Thematic Analysis of ICO Code Recommendations

Key

	ICO Code Recommendation addressed
	ICO Code Recommendation not addressed

Theme	ICO Code Recommendation	Facebook	Pinterest	Twitter	Flickr	LinkedIn	Google +
1	**Tell people who you are/ The identity of the organisation in control of the processing/ who is collecting the information (5, 8, 9, 10)**						
2	**The purpose for obtaining the information/ Be clear why you need the information/ why you're collecting the information (8,9)***						
3	**The purpose for using the information/ how it will be used/ What you are going to do with their information (8, 4, 5)***						
4	**The purpose for disclosing the information/ details of how the organisations they pass it onto will use the information (8, 10)***						
5	**Who their information will be shared with/ disclosed to/ If they intend to pass the information on, the name of the organisations involved (5, 10)**						
6	**Provide people with information about people's rights of access to their data/ About their rights and how they can exercise them *e.g. obtain a copy of their personal information or object to direct marketing* (5, 10)**						
7	**Your arrangements for keeping the data secure/ What you are going to do to ensure the security of personal information (5, 10)**						
8	**Try and predict whether you will be likely to do things with it in future without drawing up a long list of future possible uses if you are unlikely to use it for those purposes (9)**						
9	**If people have a choice over <u>whether to provide</u> information it should be properly explained to them/ they are given a proper opportunity to exercise it (9, 11)**						
10	**Where people have a choice over the <u>use of information</u>, it should be properly explained to them/ they are given a proper opportunity to exercise it (9, 11)**						
11	**The implications/effects of collecting/ using/ disclosing the information (9, 4)**						
12	**How long you or other organisations intend to keep the information (10)**						
13	**Whether the information will be transferred overseas (10)**						
14	**Who to contact if they want to complain or know about how their information will be used (10)**						

Appendix B

Theme	ICO Code Recommendation	Facebook	Pinterest	Twitter	Flickr	LinkedIn	Google +
15	About the right to complain to the Information Commissioner if there is a problem (10)						
16	Different notices aimed at different groups of people it deals with (10)						
17	Positive agreement required if previously collected information is used in a significantly different way/ If using information you already hold in a new way, you should actively seek their consent (11, 13)						
18	Where the collection and use of information is essential to provide the service or carry out the transaction, even if individuals have no real choice, the collection of information about them still needs to be fair and transparent (11).						
19	Where individuals are required by law to provide their personal details be open with people and explain clearly why their information is being collected and what it will be used for (11)						
20	If you intend to market people by electronic means (email, SMS, fax or telephone etc) special rules may apply and you may need their permission before doing so *Electronic Communications Regulations 2003* (11)						
21	If you are collecting sensitive information you should actively tell them about it and gain positive agreement. *To actively communicate means to take positive action to provide a privacy notice, which differs from having a privacy notice available for members of the public who want to see it i.e. having to click on a web link* (11, 13, 4).						
22	If providing or not providing personal information will have a significant effect on the individual you should actively tell them about it/ Consequences of not providing information *e.g. non-receipt of a benefit. To actively communicate means to take positive action to provide a privacy notice, which differs from having a privacy notice available for members of the public who want to see it i.e. having to click on a web link* (10, 13)						
23	Obtain assurances (in form of written agreements) from any organisations you share personal information with about what they will do with the information and what the effect on people is likely to be (14)						
24	Organisations that intend to combine information from different sources should explain this (14)						
25	and the likely consequences (14)						
26	In marketing contexts, when organisations ask for permission to share customer information with third parties e.g. companies in the same group, this should be backed up with more detail information such as the names of the companies involved for those who want it (14)						
27	If an organisation intends to collect personal information with the intention of selling or						

Theme	ICO Code Recommendation	Facebook	Pinterest	Twitter	Flickr	LinkedIn	Google +
	renting it you should make it clear to individuals that the information they provide could be supplied to anyone and used for any purpose and tell them this when they provide their details (14)						
28	**Individuals are told that their information may be passed onto other organisations. When a business is insolvent, bankrupt, being closed down or sold the database may be sold on (14)**						
29	**That if their information is rented, individuals are told that if the business is insolvent, bankrupt, being closed down or sold that their information will be returned to its owner (14)**						
30	**That if either of the above are true, the seller will make sure the information will only be used for the same or similar purpose and consent will be required for any new purposes (14)**						
31	**Avoid using confusing terminology e.g. technical language (16, 4)**						
32	**Avoid using legalistic language (16)**						
33	**Don't give the impression that people have a choice when in reality they do not (11, 16)**						
34	**Make sure they comply with any relevant sectoral rules (17)**						
35	**Good practice to provide the privacy notice in the same medium you use to collect the data (17)**						
36	**Is a layered notice used? Because the privacy notice is online, the organisation should make full use of the technology available to them *(13)* (18)**						
37	**If you collect information from vulnerable individuals (such as children) have appropriate privacy notice to their level of understanding – *would they understand the consequences? (*18,19)**						
38	**Good practice to provide your privacy notice in the language that your intended audience is most likely to understand, even if not required by law (19)**						
39	**Good practice to keep your privacy notice under regular review (19)**						
40	**Good practice to review the effectiveness of your notice by analysing complaints from the general public about your information use in general and your privacy notice in particular (19)**						

List of References

6, P. (1998). The future of privacy: Volume 1 Private life and public policy London: Demos

Ackoff, R. (1989). From data to wisdom, *Journal of Applied Systems Analysis* 16, 3–9

Ahsan, S. and Shah, A (2011). Data, Information, Knowledge, Wisdom: a doubly linked chain? [Online] Available at: http://citeseerx.ist.psu.edu/viewdoc, [Accessed: 9th May 2016]

Alexa. (2014). Actionable Analytics for the Web. [Online] Available: http://www.alexa.com [Accessed: 1st August 2014]

Alexio, P. and Pardo, T.A.S. (2008). Finding Related Sentences in Multiple Documents for Multidocument Discourse Parsing of Brazilian Portuguese Texts. In *Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana* – TIL, pp. 298-303. Vila Velha, Espírito Santo. October, 26-28.

Allen & Overy, 'Binding Corporate Rules' <http://www.allenovery.com/SiteCollectionDocuments/BCRs.pdf > [Accessed: 9th May 2016]

Alpaydin, E. (2020) *Introduction to Machine Learning* (Fourth ed.). Cambridge, MA: MIT Press

Anderson, H. (2009). A privacy wake-up call for social networking sites. *Ent. L.R. 20(7), 245-248*

Anderson, H. (2009) A privacy wake-up call for social networking sites *Ent. L.R. 20(7), 245-248*

ANTON, A., EARP, J., HE, Q., STUFFLEBEAM, W., BOLCHINI, D., AND JENSEN, C. Financial privacy policies and the need for standardization. *IEEE Security & Privacy 2*, 2 (Mar-Apr 2004), 36–45.

Ardelt., M (2003) Empirical assessment of a three- dimensional wisdom scale. *Research on Aging*, 25(3):275, 2003.

Aristotle (1928) Posterior analytics. In: G.R.G. Mure and W.D Ross (transl.), *The Oxford Translation of Aristotle, Vol. 1* (OUP, Oxford, 1928), 83–94.

Armstrong, S.L., Gleitman, L.R. and Gleitman, H. (1983) What some concepts might not be. *Cognition,* 13(3), 263-308

Awad, E.M. and Ghaziri, H.M. (2004) *Knowledge Management* (Pearson Education International, Upper Saddle River, NJ, 2004).

Bailey, K. D. (1984) A three-level measurement model. *Quality and Quantity*, *18*(3), 225-245.

Bailey, K.D. (1994) Typologies and Taxonomies – an introduction to Classification Techniques. Sage: Thousand Oaks, CA

Baltes, P.B and Kunzmann, U. (2003) Wisdom, *The Psychologist* 16(3) (2003) 131–2.

Baltes, P.B. and Smith, J. (1990) Toward a psychology of wisdom and its ontogenesis.

Baltes, P.B. and Staudinger, U.M. (2000) A meta- heuristic (pragmatic) to orchestrate mind and virtue toward excellence. *American Psychologist*, 55(1):122–136,

Barker, K., Askari, M., Banerjee, M., Ghazinour, K., Mackas, B., Majedi, M., Pun, S. and Williams, A., 2009. Data privacy taxonomy. In *Dataspace: The Final Frontier* (pp. 42-54). Springer Berlin Heidelberg.

Barlas, I., & Ginart, A., & Dorrity, J. L. (2005). Self-evolution in knowledge bases. In J. Romania & L. Batchler (Eds.), Proceedings of IEEE Autotestcon '05 Conference (pp. 325-331). Orlando, FL.

Barsalou, L.W. (1993a) Flexibility, structure and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A. Collins and S. Gathercole (Eds.) *Theories of memory* (pp. 29-101). Hillsdale, NJ: Lawrence Erlbaum Associates.

Barsalou, L.W. (1993b) Challenging assumptions about concepts. *Cognitive Development, 8(2),* 169-180

Barsalou, L.W. (2005) Abstraction as Dynamic Interpretation in Perceptual Symbol Systems. in *Building Object Categories in Developmental Time* ed. By Gershkoff-Stowe, L. and Rakison, D.H. Lawrence Erlbaum Associates: London and Mahwah, New Jersey

Baskarada, S. and Koronios, A., (2013). Data, information, knowledge, wisdom (DIKW): a semiotic theoretical and empirical exploration of the hierarchy and its quality dimension. *Australasian Journal of Information Systems*, *18*(1).

Batra, S., (2014). Big data analytics and its reflections on DIKW hierarchy. *Review of Management*, *4*(1/2), p.5.

Becker, J., Heddier, M., Oksuz, A. and Knackstedt, R. (2014). The Effect of Providing Visualizations in Privacy Policies on Trust in Data Privacy and Security. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on* (pp. 3224-3233). IEEE

Beigi, G., & Liu, H. (2018). Privacy in social media: Identification, mitigation and applications. *arXiv preprint arXiv:1808.02191*.

Bellinger, G., Castro, D. and Mills, A. (2004). *Data, Information, Knowledge, and Wisdom.* Available at: www.systems-thinking.org/dikw/dikw.htm (accessed: 5 February 2016).

Berčič, B and George, C (2009) Investigating the legal protection of data, information and knowledge under the EU data protection regime *International Review of Law, Computers & Technology,* 23:3, 189-201

Bernstein, J. H. (2009). The data-information-knowledge-wisdom hierarchy and its antithesis. In Jacob, E. K. and Kwasnik, B. (Eds.). (2009). Proceedings North American Symposium on Knowledge Organization Vol. 2, Syracuse, NY, pp. 68-75.

BEUC (2011) A Comprehensive Approach on Personal Data Protection in the European Union, European Commission's Communication: The European Consumers' Organisation's response [Online] Available at: http://ec.europa.eu/justice/news/consulting_public/0006/contributions/organisations/beuc_en.pdf [Accessed 20 April 2016]

Bierly III, P.E., Kessler, E.H., and Christensen, E.W. (2000). Organizational Leaning, Knowledge and Wisdom. Journal of Organizational Change, 13, 6, 595-618

Bocij, P., Chaffey. D., Greasley, A., And Hickie, S (2003) *Business Information Systems: Technology, Development and Management for the e-Business* (FT Prentice Hall: Harlow)

Boddy, D., Boonstra, A., and Kennedy, G. (2005) *Managing Information Systems: an Organizational Perspective*, 2nd edn (FT Prentice Hall: Harlow).

Bornstein, M. (1984) Descriptive Taxonomy of psychological Categories Used by Infants. *Origins of Cognitive Skills* ed. By Sophian, C. Hillsdale, N.J.: Erlbaum.

Boubacar A., and Niu Z. (2014) Conceptual Clustering. In: Park J., Pan Y., Kim CS., Yang Y. (eds) Future Information Technology. Lecture Notes in Electrical Engineering, vol 309. Springer, Berlin, Heidelberg.

Bowker, G. C. and Star, S.L. (1999) *Sorting Things Out: Classification and its Consequences*. MIT Press, Cambridge, MA

Boyatzis, R. E. (1998). Transforming qualitative information: Thematic analysis and code development. London: Sage Publications.

Boyd, D. and Hargittai, E. (2010). Facebook privacy settings: Who cares? *First Monday* 15(8).

Bransford, J. D., and Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of verbal learning and verbal behavior*, *11*(6), 717-726.

Braun, V., and Clarke, V. (2006) Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77- 101.

Buckland, M. (1991). *Information and information systems.* (New York: Greenwood Press).

Carlisle, J. P. (2006). Escaping the veil of Maya: Wisdom and the organization. Proceedings of the 39th Annual Hawaii International Conference on System Sciences, Koloa Kauai, HI. doi: 10.1109/HICSS.2006.160

Charlesworth, A. (2000) Clash of the data titans? US and EU data privacy regulation. *European Public Law*, 6(2), pp.253-274.

Chalton, S. (2004). The Court of Appeal's interpretation of “personal data” in Durant v FSA–a welcome clarification, or a cat amongst the data protection pigeons?. *Computer Law & Security Review*, *20*(3), 175-181.

Chen, J. et al. (2016) WaaS-Wisdom as a Service. In: Zhong, N., Ma, J., Liu, J., Huang, R,. Tao, X. (eds) Wisdom Web of Things. Web Information Systems Engineering and Internet Technologies Book Series. (Springer: Cham)

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power- law distributions in empirical data. *SIAM review*, *51*(4), 661- 703.

Cleveland, H (1982). Information as a Resource. *The Futurist* 16(6): 34-39.

Corrigan, R., Eckman, F.R. and Noonan, N. (1989) *Linguistic Categorization* John Benjamins Publishing Company

Cradock, E., Millard, D. And Stalla-Bourdillon, S. (2015). Investigating Similarity Between Privacy Policies of Social Networking Sites as a Precursor for Standardization. *Proceedings of the 24th International Conference on World Wide Web Companion* (pp. 283-289). International World Wide Web Conferences Steering Committee.

Cradock, E., Millard, D. And Stalla-Bourdillon, S. (2016) An Extended Investigation of the Similarity Between Privacy Policies of Social Networking Sites as a Precursor for Standardization *The Journal of Web Science 2(1)*

Cranor, L. F. (2003). P3P: Making privacy policies more useful. *IEEE Security & Privacy*, (6), pp.50-55.

Cradock, E., Stalla-Bourdillon, S., & Millard, D. (2017). Nobody puts data in a corner? Why a new approach to categorising personal data is required for the obligation to inform. *Computer law & security review*, *33*(2), 142-158.

Cranor, L. F. (2012). Necessary But Not Sufficient: Standardized Mechanisms for Privacy Notice and Choice Journal on Telecommunications and High Technology Law 10(2), 273-307.

Cranor, L.F., Idouchi, K., Leon, P.G., Sleeper, M. & Ur, B. (2013) Are They Actually Any Different? Comparing Thousands of Financial Institutions’ Privacy Practices. In *Proceedings of the 12th Workshop on the Economics of Information Security* (WEIS 2013), Jun 11-12, Washington, DC.

Descartes, R. Discourse on the method (etc.). In: E.S. Haldane and G.R.T. Ross (transl.), *The Philosophical Works of Descartes, Vol. 1* (Cambridge University Press, Cambridge, 1911), 217–24

Doty, G.H., Glick, W.H. and Huber, G.P. (1993) *Fit, equifinality, and organizational effectiveness: a test of two configurational theories*. Academy of Management Journal 36(6), 1195–1250

Eldredge, N. and Cracraft, J. (1980) *Phylogenetic Patterns and the Evolutionary Process.* Columbia University Press: New York

Electronic Privacy Information Center (2009). *Complaint, Request for Investigation,*

Injunction, and Other Relief [Online] Available: http://epic.org/privacy/inrefacebook/EPIC-FacebookComplaint.pdf [Accessed: 21st July 2015]

Eliot, T.S. (1934) *The Rock* (Faber & Faber: London)

Estes, W.K. (1986) Memory storage and retrieval processes in category learning, *Journal of Experimental Psychology General,* 115(2), 155-174

European Commission 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Online Platforms and the Digital Single Market Opportunities and Challenges for Europe (COM(2016)288)' <https://ec.europa.eu/digital-single-market/en/news/communication-online-platforms-and-digital- single-market-opportunities-and-challenges-europe> [accessed 31 May 2016]

European Commission Working Party on The Protection of Individuals With Regard To The Processing Of Personal Data. (1999). Recommendation 1/99 on Invisible and Automatic Processing of Personal Data on the Internet Performed by Software and Hardware [Online] Available at: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion- recommendation/files/1999/wp17_en.pdf [Accessed 20 April 2016]

European Commission Working Party On The Protection Of Individuals With Regard To The Processing Of Personal Data (2018). *Guidelines on Transparency under Regulation 2016/679* [Online] Available: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=622227 [Accessed: 9th February 2020].

European Commission Working Party On The Protection Of Individuals With Regard To The Processing Of Personal Data (2017) Guidelines on Auto-mated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679 17/EN WP 251

European Commission Working Party The Protection Of Individuals With Regard To The Processing Of Personal Data (2007) *Opinion 4/2007on the concept of personal data* [Online] Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf

European Commission. (2003). *Analysis and impact study on the implementation of Directive EC 95/46 in Member States* [Online] Available: http://ec.europa.eu/justice/data-protection/reform/index_en.htm [29th January 2016].

European Commission. (2015). European Commission – Fact Sheet: Questions and Answers – Data Protection reform [Online] Available at: http://europa.eu/rapid/press-release_MEMO-15-6385_en.htm [Accessed 20 April 2016]

European Commission. (2016). *Reform of EU data protection rules* [Online] Available: http://ec.europa.eu/ [Accessed: 29th January justice/data-protection/reform/index_en.htm 2016].

European Commission. Working Party On The Protection Of Individuals With Regard To The Processing Of Personal Data. (2011). Opinion 15/2011 on the definition of consent [Online] Available at: http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2011/wp187_en.pdf [Accessed

20 April 2016]

European Commission. Working Party On The Protection Of Individuals With Regard To The Processing Of Personal Data (2000). Privacy on the Internet - An integrated EU Approach to On-line Data Protection [Online] Available at: http://ec.europa.eu/justice/data-protection/article- 29/documentation/opinion-recommendation/files/2000/wp37_en.pdf [Accessed 20 April 2016]

European Commission. Working Party On The Protection Of Individuals With Regard To The Processing Of Personal Data (2014). Appendix: List of possible compliance measures' [Online] Available at: http://ec.europa.eu/justice/data-protection/article-29/documentation/other-document/files/2014/20140923_letter_on_google_privacy_policy_appendix.pdf [Accessed 20 April 2016]

European Commission. Working Party On The Protection Of Individuals With Regard To The Processing Of Personal Data (2007). Opinion 4/2007 on the concept of personal data [Online] Available at: http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf [Accessed 20 April 2016]

European Commission (2014) *How will the data protection reform affect social networks?* [Online] Available at: http://ec.europa.eu/justice/data-protection/document/review2012/factsheets/3_en.pdf [Accessed 4th September 2014]

European Parliament And Of The Council (2002) Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications) [Online] Available at: http://eur- lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:en:HTML [Accessed 20 April 2016]

European Parliament and of the Council. (1995). DIRECTIVE 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data [Online] Available: http://eur- lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML [Accessed 21st December 2014].

European Parliament And Of The Council. (2009). DIRECTIVE 2009/136/EC amending Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services, Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector and Regulation (EC) No 2006/2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws [Online] Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:337:0011:0036:en:PDF [Accessed 20 April 2016]

European Parliament and of the Council. (2016). REGULATION (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [Online] Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN [Accessed 9th February 2020].

European Parliament. Committee On Civil Liberties, Justice And Home Affairs. (2012).Draft Report on the proposal for a regulation of the European Parliament and of the Council on the protection of individual with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) [Online] Available at: http://www.europarl.europa.eu/meetdocs/2009_2014/documents/libe/pr/922/922387/9 22387en.pdf [Accessed 23 June 2015]

Facebook. (2013). *Data Use Policy.* [Online] Available: https://www.facebook.com/privacy/explanation [Accessed 1st August 2014]. facebooks-revised-policies-and-terms-v1-3.pdf [Accessed 2 February 2016]

Federal Trade Commission. (2000). *Privacy Online: Fair Information Practices in the Electronic Marketplace* [Online] Available: https://www.ftc.gov/sites/default/files/documents/reports/privacy-online-fair-information-practices-electronic-marketplace-federal-trade-commission-report/privacy2000.pdf [Accessed 29 January 2016].

Federal Trade Commission. (2010). *Protecting Consumer Privacy in an Era of Rapid Change: A Proposed Framework for Businesses and Policymakers* [Online] Available: https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-bureau-consumer-protection-preliminary-ftc-staff-report-protecting-consumer/101201 *Final_0. pdf*. [Accessed 27 January 2016]

Fisher, D. H. (1987) Knowledge acquisition via incremental conceptual clustering. *Machine learning*, *2*(2), 139-172.

Fleishman, E. A., Quaintance, M. K., & Broedling, L. A. (1984). *Taxonomies of human performance: The description of human tasks.* Academic Press.

Flickr. (2014). *Yahoo Privacy Center. Community* [Online] Available: https://policies.yahoo.com/us/en/ yahoo/privacy/index.htm [Accessed 1 August 2014]

Flickr. (2015). *Thank You, Flickr Community* [Online] Available: http://blog.flickr.net/en/2015/06/10/thank-you-flickr-community/ [Accessed 2 February 2016]

Frické, M. (2009). The knowledge pyramid: a critique of the DIKW hierarchy *Journal of information science*, *35*(2), 131-142

Friedman, B., Lin, P. And Miller, J.K. (2005). Informed Consent By Design. *Security And Usability*, 503-530.
GAVISON, R. (1980). Privacy and the Limits of Law. *Yale law journal*, 421-471

Gellman, R. (1998) Does Privacy Law Work?, in PE Agre and M Rotenberg (eds), *Technology and Privacy: The New Landscape* Cambridge, MA: MIT Press
Gelman, S.A. (2003) *The essential child: Origins of essentialism in everyday thought.* London: Oxford University Press

Gervais, A., Filios, A., Lenders, V., & Capkun, S. (2017, September). Quantifying web adblocker privacy. In *European Symposium on Research in Computer Security* (pp. 21-42). Springer, Cham.

Glaser, B. G. (1978). Theoretical sensitivity: Advances in the methodology of grounded theory (Vol. 2). Mill Valley, CA: Sociology Press.

GNIP. (2016). *GNIP: Unleash the Power of Social Data* [Online] Available: http://gnip.com [Accessed 2 February 2016]

Gollapudi, S. (2016) *Practical machine learning.* Packt Publishing Ltd.

Goodman, B., & Flaxman, S. 2016. EU regulations on algorithmic decision-making and a "right to explanation." 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), 1–9.

Google (2016) 'Google Privacy and Terms' <https://www.google.co.uk/intl/en/policies/privacy/?fg=1> [accessed 23 September 2016]

Google. (2014). *Privacy Policy. Community* [Online] Available: https://www.google.co.uk/intl/en/policies/privacy/?fg=1> [accessed 23 September 2016]

Gostin, L.O., Levit, L.A. and Nass, S..J. (ed[s]) (2009). *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. (Washington D.C.: National Academies Press).

Great Britain. (1998) *Data Protection Act 1998: Elizabeth II (1998)* (London: The Stationary Office)

Greengard, S. 2018. Weighing the impact of GDPR. Communications of the ACM, 61(11), 16–18. https://doi.org/10.1145/3276744

Grossklags, J., and ACQUISTI, A. (2007). When 25 Cents is Too Much: An Experiment on Willingness-To-Sell and Willingness-To-Protect Personal Information. In *WEIS*. [Online] Available at: http://weis2007.econinfosec.org/papers/66.pdf [Accessed: 21st August 2014].

GUEST, G., MACQUEEN, K. M., and NAMEY, E. E. (2012). Introduction To Applied Thematic Analysis. *Applied Thematic Analysis*, London: Sage Publications

Guo, W., Rodolitz, J., & Birrell, E. (2020, November). Poli-see: An Interactive Tool for Visualizing Privacy Policies. In *Proceedings of the 19th Workshop on Privacy in the Electronic Society* (pp. 57-71).

Hargittai, E. (2010). Facebook privacy settings: Who cares? First Monday. 15(8).

Hastie, T., Tibshirani, R., and Friedman, J. (2001). The elements of statistical learning: Data mining, inference, and prediction. (New York, NY: Springer-Verlag)

HL Deb (1993-1994) 549 col. 37

HL Deb (1997-1998 585 col. 445), 1998

Holmes, K. M., and O'Loughlin. N. (2014) The experiences of people with learning disabilities on social networking sites. *British Journal of Learning Disabilities* 42.1 (2014): 1-5.

Hoppe, A., Seising, r., Nürnberger, A., and Wenzel., C (2011) Wisdom - the blurry top of human cognition in the DIKW-model? Proceedings of the 7th conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-11)

Information Commissioner's Office (2016) Where should you deliver privacy information to individuals? <https://ico.org.uk/for-organisations/guide-to-data-protection/privacy-notices-transparency-and-control/where- should-you-deliver-privacy-information-to-individuals/> [Accessed 7 October 2016]

Information Commissioner's Office (2016). What should you include in your privacy notice?'<https://ico.org.uk/for-organisations/guide-to-data-protection/privacy-notices-transparency-and- control/what-should-you-include-in-your-privacy-notice/> [Accessed 7 October 2016]

Information Commissioner's Office. (2010). *Privacy notices code of practice* [Online] Available: http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Detailed_specialist_guides/PRIVACY_NOTICES_COP_FINAL.ashx [Accessed 21 July 2015]

Information Commissioner's Office. (2016). Consultation: Privacy notices, transparency and control – a code of practice on communicating privacy information to individuals [Online] Available at: https://ico.org.uk/media/about-the-ico/privacy-notices-transparency-and-control-0-0.pdf [Accessed 20 April 2016]

Information Commissioner's Office. (2016). *Use a layered approach* [Online] Available: https://ico.org.uk/about-the-ico/privacy-notices-transparency-and-control/use-a-layered-approach/ [Accessed 3rd February 2016]

IPSOS. (2013). Global Respondents Visit Three Types of Websites Most Frequently: Search Engines (74%), Social Networking (64%) and Email Portals (55%) [Online] Available at: http://www.ipsos-na.com/news-polls/pressrelease.aspx?id=6303 [Accessed 20 April 2016]

Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. *New phytologist*, *11*(2), 37-50.

Jashapara, A. (2005) *Knowledge Management: an Integrated Approach* (FT Prentice Hall, Harlow).

Jessup., L.M. and Valacich, J.S. (2003). *Information Systems Today* (Prentice Hall, Upper Saddle River, NJ).

Johnson, B. and Higgins, J. (2010). *Information Lifecycle Support* (The Stationary Office, London)

Joint cases C-141/12 and C- 372/12 *YS and M. and S. v Minister of Immigration, Integration and Asylum* [2016], ECLI:EU:C:2014:2081, Opinion of Advocate General Sharpston

Kant, I. (1965) *Critique of Pure Reason* (St Martin's Press, New York) [translated by N.K. Smith].

Kaufman K.A. (2012) Conceptual Clustering. In: Seel N.M. (eds) Encyclopedia of the Sciences of Learning. Springer, Boston, MA.

Keil, F.C. (1994) Explanation, association, and the acquisition of word meaning. In L. Gleitman & B Landau (Eds.) *The acquisition of the lexicon* (pp. 169-196). Cambridge, MA: MIT Press

Keil, F.C. (2003) Categorisation, causation and the limits of understanding: Conceptual representation [Special issue]. *Language & Cognitive processes, 18(5-6), 63-692*

Keil, F.C. (2005) Knowledge, Categorization and the Bliss of Ignorance in *Building Object Categories in Developmental Time* ed. By Gershkoff-Stowe, L. and Rakison, D.H. Lawrence Erlbaum Associates: London and Mahwah, New Jersey

Kelley, P. G., Cesca, L., Bresee, J., & Cranor, L. F. (2010). Standardizing privacy notices: an online study of the nutrition label approach. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems* (pp. 1573-1582). ACM.

Kokolakis, S. (2017). Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & security*, *64*, 122-134.

Korff, D. (2003). EC Study On Implementation Of Data Protection Directive: comparative summary of national laws [Online] Available at: http://194.242.234.211/documents/10160/10704/Stato+di+attuazione+della+Direttiva +95-46-CE [Accessed 20 April 2016]

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802-5805.

Krebs, L. M., Alvarado Rodriguez, O. L., Dewitte, P., Ausloos, J., Geerts, D., Naudts, L., & Verbert, K. (2019, May). Tell me what you know: GDPR implications on designing transparency and accountability for news recommender systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-6).

Leon, P.G., Ur, B., Wang, Y., Sleeper, M., Balebako, R., Shay, R., Bauer, l., Christodorescu,

M. and Cranor, L.F. (2013) What matters to users?: factors that affect users' willingness to share information with online advertisers. In *Proceedings of the ninth symposium on usable privacy and security* (p. 7). ACM library/data_protection/detailed_specialist_guides/review_of_eu_dp_directive.ashx

[Accessed 4th September 2014]

Liew, A., (2013). DIKIW: Data, information, knowledge, intelligence, wisdom and their interrelationships. *Business Management Dynamics*, *2*(10), p.49.

LinkedIn. (2014). *Your Privacy Matters.* [Online] Available: *https://www.linkedin.com/legal/privacy-policy?trk=uno-reg-guest-home-privacy-policy* [Accessed 1st August 2014]

Machlup, F. (1980). *Knowledge and knowledge production.* (Princeton, NJ: Princeton University Press).

Mai, J.E., (2016). Big data privacy: The datafication of personal information. *The Information Society*, *32*(3), pp.192-199.

Malgieri, G. (2016). Property and (Intellectual) ownership of consumers' information: a new taxonomy for personal data. *Privacy in Germany-PinG*, (4), 133.

Malt, B.C. (1994) Water is not H-Sub-20. *Cognitive Psychology,* 27(1), 41-70

Markman, E. (1983) Two Different Kinds of Hierarchical Organization. *New Trends in Conceptual Representation: Challenges to Piaget's Theory* ed. By Ellen R Scholnick. Hillsdale, N.J.: Erlbaum

Martin, K., & Murphy, P. (2017). The role of data privacy in marketing, Journal of the Academy of Marketing Science, 45(2), doi:10.1007/ s11747-016-0495-4.

Mayer-Schonberger, V & Cukier, K., (2013), "Big Data: A Revolution that will transform how we live, work and think" (John Murray, UK)

Mcarthur, R. L. (2001). Reasonable expectations of privacy. *Ethics and Information technology*, *3*(2), 123-128.

McDonald, A. M. and Cranor, L. F. (2008). The Cost of reading privacy policies *ISJLP*, *4*, 543.

McRae, K., Cree, G.S., Westmacott, R, and de Sa, V.R. (1999) Further evidence for feature correlations in semantic memory: Visual word recognition [Special issue] *Canadian Journal of Experimental Psychology, 53(4),* 360-373

McRobb, S., & Stahl, B. C. (2007). Privacy as a shared feature of the e-phenomenon: a comparison of privacy policies in e-government, e-commerce and e-teaching. *International journal of information technology and management*, *6*(2-4), 232-249

Michalski, R. S. (1980) Knowledge acquisition through conceptual clustering: A theoretical framework and algorithm for partitioning data into conjunctive concepts. *International Journal of Policy Analysis and Information Systems*, 4, 219-243.

Michalski, R. S., and Stepp, R. E. (1983) Learning from observation: Conceptual clustering. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (pp. 331–363). Palo Alto: Tioga.

Moore, A. D., (2008) Defining Privacy. Journal of Social Philosophy, Vol. 39, No. 3, pp. 411-428, Fall 2008. Available at SSRN: https://ssrn.com/abstract=1980849 [Accessed 1st August 2014]

Muller, H. P. and Maasdorp, C.H. (2011) Fifth IEEE International Conference on Research Challenges in Information Science Proceedings, Gosier, Guadeloupe, France, May 19-21 2011

Mutongi, C. (2016) IOSR Journal of Business and Management (IOSR-JBM) e-ISSN: 2278-487X, p-ISSN: 2319-7668. Volume 18, Issue 7 .Ver. II (July 2016), PP 66-71

Mydex (2010) The Case for Personal Information Empowerment: The rise of the personal data store [Online] Available at: https://mydex.org/wp- content/uploads/2010/09/The-Case-for-Personal-Information-Empowerment-The- rise-of- the-personal-data-store-A-Mydex-White-paper-September-2010-Final- web.pdf [Accessed 23 June 2015]

N.K. Kakabadse, A. Kakabadse and A. Kouzmin, Reviewing the knowledge management literature: towards a taxonomy, *Journal of Knowledge Management* 7(4) (2003) 75–91.

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013) A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, *22*(3), 336-359.

Nosofsky, R.M. (1984) Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory & Cognition, 10(1),* 104-114

Obar, J. A., & Oeldorf-Hirsch, A. (2020). The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, *23*(1), 128-147.

Ohm, P (2010) Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA law review*, *57*, 1701

Olurin, M., Adams, C., and Logrippo, L. (2012). Platform for privacy preferences (P3P): Current status and future directions. In *Privacy, Security and Trust (PST), 2012 Tenth Annual International Conference on* (pp. 217-220). IEEE.

Organisation for Economic Cooperation and Development Working Party On Security And Privacy In The Digital Economy. (2014). Protecting Privacy in a Data-driven Economy: Taking Stock of Current Thinking [Online] Available: http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=dsti/iccp/reg%282014%293&doclanguage=en [Accessed 23rd June 2015]

Organisation for Economic Cooperation and Development. (2013). OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data [Online] Available: https://www.oecd.org/internet/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm [Accessed 10 February 2020]

Organisation for Economic Cooperation and Development. (2013). *The OECD Privacy Framework* [Online] Available: http://www.oecd.org/sti/ieconomy/ oecd_privacy_framework.pdf [Accessed 27th January 2016]

Oxford University Press Oxford English Mini Dictionary. (2011). (New York: Oxford University Press)

Pearlson, K. and Saunders, C (2004) *Managing and Using Information Systems: a Strategic Approach* (Wiley: New York)

Pinterest. (2014). *Privacy Policy* [Online] Available: https://about.pinterest.com/en/privacy-policy [Accessed 1st August 2014]

Polanyi, M.E. (1958) *Personal Knowledge: towards a Post-Critical Philosophy* (Routledge and Kegan Paul, London).

Posner, M.I. (1969) Abstraction and the process of recognition. In G.H. Bower & J.T. Spence (Eds.), The psychology of learning and motivation (pp. 43–100). New York: Academic Press.privacy/archive/20140331/ [Accessed 2nd February 2016]

Purtova, N. (2018). The law of everything. Broad concept of personal data and future of EU data protection law. *Law, Innovation and Technology*, *10*(1), 40-81.

Rasmussen, C., And Dara, R. (2014). Recommender Systems for Privacy Management: A Framework. In *High-Assurance Systems Engineering (HASE), 2014 IEEE 15th International Symposium on* (pp. 243-244). IEEE.

Reagle, J., And Cranor, L. F. (1999) The platform for privacy preferences. *Communications of the ACM*, *42*(2), 48-55.

Richards, N. M., & Solove, D. J. (2007). Privacy's Other Path: Recovering the Law of Confidentiality. *Geo. LJ*, *96*, 123.

Rips, L.J. (1989) Similarity, typicality and categorization. In S Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21-59) New York: Cambridge University Press

Robinson, N., Graux, H., Botterman, M. and Valeri, L., 2009. Review of the European data protection directive. *Rand Europe*.

Rogers, K. (2011) *The Internet and The Law* London:Palgrave Macmillan

Rosch, E. (1973) On the Internal Structure of Perceptual and Semantic Categories. *Cognitive Development and the Acquisition of Language* ed. By Thomas Moore. New York: Academic Press.

Rosch, E., and Lloyd, B. B. (Eds.). (1978). Cognition and categorization. (Lawrence Erlbaum).

Rosch, E., and Mervis, C. (1975) Family Resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7.573-609

Rowland, D., Kohl, U. and Charlesworth, A. (2012). *Information Technology Law. 4th Ed.* (Oxon: Routledge Publishing)
Rowley, J. (2007). The Wisdom Hierarchy: representations of the DIKW Hierarchy, Journal of Information Science, 33 (2) 2007, pp. 163–180

Rowley, J. (2007). What is information? *Information Services & Use* 18 (1998) 243–54.

Ryan, G. W., and Bernard, H. R. (2003). Techniques to identify themes. *Field methods*, *15*(1), 85-109.

Ryser, J. and Glinz, M., (1999) A scenario-based approach to validating and testing software systems using statecharts. In *Proc. 12th International Conference on Software and Systems Engineering and their Applications*.

Sarne, D., Schler, J., Singer, A., Sela, A., & Bar Siman Tov, I. (2019). Unsupervised topic extraction from privacy policies. In *Companion Proceedings of The 2019 World Wide Web Conference* (pp. 563-568).

Schumaker. R.P. (2011). From Data to Wisdom: The Progression of Computational Learning in Text Mining, Communications of the IIMA, Volume 11 Issue 1, Article 4

Schwartz, A., (2009). Looking back at P3P: lessons for the future. *Center for Democracy & Technology, https://www. cdt. org/files/pdfs/P3P_Retro_Final_0. pdf*.

Schwartz, P.M. and Solove, D.J., (2012). Pii problem: Privacy and a new concept of personally identifiable information, the. *NYUL Rev.*, *86*, 1814

Schwarz, A., Mehta, M., Johnson, N. and Chin, W.W. (2007) *Understanding frameworks and reviews: a commentary to assist us in moving our field forward by analyzing our past*. The Database for Advances in Information Systems 38(3), 29–50.

Scribbins, K. (2001). Privacy@ net: an international comparative study of consumer privacy on the internet. Consumers International. [Online] Available at: http://www.consumersinternational.org/media/304817/privacy@net-%20an%20international%20comparative%20study%20of%20consumer%20privacy%20on%20the%20internet.pdf [Accessed 20 April 2016]

Securities and Exchange Commission. (2009). *Final Model Privacy Form Under the Gramm-Leach-Bliley Act* [Online] Available: https://www.sec.gov/divisions/marketreg/tmcompliance/modelprivacyform-secg.htm [Accessed 29th January 2016]

Selbst, A., & Powles, J. (2018, January). "Meaningful Information" and the Right to Explanation. In *Conference on Fairness, Accountability and Transparency* (pp. 48-48). PMLR.

Sellars, S. (2011). Online privacy: do we have it and do we want it? A review of the risks and UK case law. *European Intellectual Property Review*, *33*(1), 9-17.

Shalev-Shwartz, S. and Ben-David, S. (2014) *Understanding Machine Learning: From Theory to Algorthms*, Cambridge University Press: Cambridge

Simpson, G. G. (1961). Principles of animal taxonomy. Columbia University Press

Smith, E.E., and Medin, D.L. (1981). *Concepts and categories.* Cambridge, MA: Harvard University Press.

Smith, H. J., Dinev, T., & Xu, H. (2011). Information privacy research: an interdisciplinary review. *MIS quarterly*, 989-1015.

Smith, L.B. (2005) Emerging Ideas About Categories. in *Building Object Categories in Developmental Time* ed. By Gershkoff-Stowe, L. and Rakison, D.H. Lawrence Erlbaum Associates: London and Mahwah, New Jersey

Soergel, D (1985). *Organizing information: Principles of database and retrieval systems*. (Orlando, FL: Academic Press).

Sokal, R.R. and Sneath, P.H.A. (1963) *Principles of Numerical Taxonomy*. W.H. Freeman and Company, San Francisco, CA.

Solove, Daniel J., (2002). Conceptualizing Privacy. California Law Review, Vol. 90, p. 1087, 2002. Available at SSRN: https://ssrn.com/abstract=313103 [Accessed 29th January 2016]

Solove, D. J. (2016). A brief history of information privacy law. *Proskauer on privacy, PLI*.

Special Eurobarometer 359. (2011). *Attitudes on Data Protection and Electronic Identity in the European Union* [Online] Available: http://ec.europa.eu/public_opinion/archives/ebs/ebs_359_en. pdf [Accessed 4th September 2014]

statista. (2016). *Leading social networks worldwide as of January 2016, ranked by number of active users (in millions).* [Online] Available: http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ [Accessed 2nd February 2016]

Stenmark, D. (2002). Information vs. knowledge: The role of intranets in knowledge management. Proceedings of the 35th Annual Hawaii International Conference on System Sciences, Waikoloa, HI.

Sternberg, R.J. (1998) A Balance Theory of Wisdom. Review of General Psychology, Vol. 2, No. 4, 1998

Stewart, D. W. (2017). A comment on privacy. *Journal of the academy of marketing science*, *45*(2), 156-159.

Stone Temple Consulting. (2015). *Hard Numbers for Public Posting Activity on Google Plus.* [Online] Available: https://www.stonetemple.com/real-numbers-for-the-activity-on-google-plus/ [Accessed 2nd February 2016]

Strauss, A., and Corbin, J. M. (1990). Basics of qualitative research: Grounded theory procedures and techniques. (London: Sage)

Sweeney, L. (2000). Simple demographics often identify people uniquely. *Health (San Francisco)*, *671*(2000), 1-34.

Taylor, M (2015) Safeguarding the Right to Data Protection in the EU, 30th and 31st October 2014, Paris, France. *Utrecht J. Int'l & Eur. L.*, *31*, 145

Teltzrow, M., & Kobsa, A. (2004). Impacts of user privacy preferences on personalized systems. In *Designing personalized user experiences in eCommerce* (pp. 315-332). Springer, Dordrecht.

Tsai, J. Y., Egelman, S., Cranor, L. Acquisti, A. (2011). The Effect of Online Privacy Information on Purchasing Behaviour: An Experimental Study, Information Systems Research, 1047-7047, Vol. 22(2) 254-268.

Tuomi, I. (1999). Data Is More than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory, Journal of Management Information Systems, 16:3,103-117, DOI: 10.1080/07421222.1999.11518258 [Accessed 20 April 2016]

Twitter. (2013). *Twitter Privacy Policy* [Online] Available: https://twitter.com/privacy?lang=en [Accessed 1st August 2014]

United States. (1998) Children's Online Privacy Protection Act 1998. 15 U.S.C. 6501–6505

Usable Privacy. (2016). Explore Privacy Policies [Online] Available at: https://explore.usableprivacy.org [Accessed 20 April 2016]

Van Alsenoy, B., Verdoodt, V., Heyman, R., Wauters, E., Ausloos, J. and Acar, G., (2015). From social media service to advertising network: a critical analysis of Facebook's Revised Policies and Terms. [Online] Available: https://www.law.kuleuven.be/citip/en/news/item/

Van Hoecke, M. (ed.), (2011) Methodologies of Legal Research – Which Kind of Method for What Kind of Discipline? (Oxford and Portland, OR: Hart Publishing,)

Vibhute, K. and Aynalem, F. (2009) *Legal Research Methods* p. 16.

Vigilocco, G., Vinson, D., Lewis, W., and Garrett, M. (2004) Representing the meanings of object and action words: The Featural Unity and Semantic Space (FUSS) hypothesis. *Cognitive Psychology, 48,* 422-488

Wachter, S. (2018). Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR. *Computer law & security review*, *34*(3), 436-449.

Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. *Colum. Bus. L. Rev.*, 494.

Wan and Alagar (2014) Synthesizing Data-to-Wisdom Hierarchy for Developing Smart Systems, 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery

Warren, S. and Brandeis, L.D. (1890) "The Right to Privacy" *Harvard Law Review* 4

Weber., R.H, 'Internet of Things: Privacy Issues Revisited' (2015)31 Computer Law & Security Review 618, 625

Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Leon, P. G., ... & Norton, T. B. (2016, August). The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1330-1340).

Wisdom: Its nature, origins, and development, pages 87–120, 1990.

Wittgenstein, L. (1953) *Philosophical investigations* (G.E.M. Anscombe, Trans) New York: Macmillan

Wogalter, M. S. (1999). On the adequacy of legal documents: factors that influence informed consent. *Ergonomics*, *42*(4), 593-613.

World Economic Forum. (2014). 'Rethinking Personal Data: A New Lens for Strengthening Trust'
<http://www3.weforum.org/docs/WEF_RethinkingPersonalData_ANewLens_Report_2014.pdf> [Accessed 9 April 2016]

Wu, L., Majedi, M., Ghazinour, K. and Barker, K., (2010), March. Analysis of social networking privacy policies. In *Proceedings of the 2010 EDBT/ICDT Workshops* (p. 32). ACM.
Xu, D., and Tian, Y. (2015) A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*(2), 165-193.

Yao, Y., (2019) *Tri-level thinking: models of three-way decision* International Journal of Machine Learning and Cybernetics https://doi.org/10.1007/s13042-019-01040-2 [Accessed 10 February 2020]

Zeleny, M. (1987). Management Support Systems: Towards Integrated Knowledge Management. In *Human Systems Managemen*t 7(1): 59-70.

Zhang, Z., Otterbacher, J., and Radev, D. (2003). Learning cross-document structural relationships using boosting. In *Proceedings of the twelfth international conference on Information and knowledge management* (pp. 124-130). ACM.

Zwenne, G (2013). Diluted Privacy [Online] Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2488486 [Accessed 20 April 2016]

Yee, G., & Korba, L. (2005, May). Comparing and matching privacy policies using community consensus. In *Proceedings, 16th IRMA International Conference, San Diego, California*.