Data Categorisation and Classification: A Systematic Review[1]

Rahul Patil, Anjula Gurtoo

Centre for Society and Policy, Indian Institute of Science, Bangalore

Abstract

Indian government is taking bold steps to galvanise the data economy through open government data platforms, smart city initiatives, and legal means. Many players in the domain aspire to leverage the monetisation potential of untapped data silos. However, while industry practitioners and public authorities look to categorise their data assets for maximizing usefulness, they have limited know-how and limited understanding of the data flow ecosystem. The primary research question of interest in this paper, therefore, is to understand what the key existing data categories or taxonomies are to promote good data governance. To address the query, we systematically reviewed prior research from academia, public and legal administration, and industry reports. We found that data users classify their data assets to address five broad pre-requisites namely, regulatory / statutory compliance, technical requirements, sociocultural and operational responsibilities, risk mitigation, and corporate policies or voluntary criteria. As the next step we will investigate examples from the industry to illustrate and expand the five board archetypes in practice.

Keywords

Data classification, Data categorisation, Principles of classification; Existing criteria and standards.

[1] [1] This work is supported by a grant from Omidyar Network. We gratefully acknowledge the discussions held with the Omidyar Network team and other academicians in shaping this paper.

1. Introduction

This article adopts systematic review approach to formalise a categorisation of varieties of data and the underlying criteria to segregate them into different buckets. The proposed archetypes for data classification criteria will facilitate the future research agenda for data classification and categorisation in the context of practice. Section 2 introduces the concept and relevancy of data, data classification, and data categorisation. The gaps in the literature are defined with the outline of past research to highlight the need for categorisation of the data classification criteria, the identification of data varieties, and the cognitive approach behind the data classification efforts. Section 3 illustrates the methodology adopted to develop the archetypes building upon the scholarly literature and practitioners' perspectives. Section 4 discusses the five proposed archetypes on data classification criteria: regulatory and statutory compliance, addressing technical requirements, fulfilling sociocultural and operational responsibilities, risk mitigation, and satisfying corporate policies or voluntary criteria. Section 5 discusses interdependencies, potential, and limitations of the proposed archetypes.

What are data and information?

Defining data is genuinely challenging, considering its multidimensionality, contextualisation, or territorial and sectoral applications. The standardised definitions state data as "the reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing" (ISO/IEC 2382 -1:1993)[i] or "information in a specific representation, usually as a sequence of symbols that have meaning" (CNSSI 4009, 2015)[ii]. Information referred herein can be either facts and ideas that can be represented as various forms of data, or as knowledge in any medium or form that can be communicated between system entities" (IETF RFC 4949, 2007)[iii]. Basically, information turns into data if it is represented as formally suitable for communication and processing. Data can be generated through human activities, from machines or sensors, or as a by-product of other processes (Banterle, F., 2020)[iv]. Entities made up of such data are called data assets like, documents, databases, websites, or any information-based resource or service (CNSSI 4009, 2015).

Difference between data classification and data categorisation

Both in practice and literature, data classification and categorisation are used interchangeably. However, research dive shows significant differences between the two. Broader understanding of categorisation by Jacob (2004)[v] highlights categorization of data as the process of dividing the data into groups of datasets whose members are in some way similar to each other. The data classification, on the other hand, can be seen as the process of orderly and systematic assignment of datasets to one and only class within a system of mutually exclusive and non-overlapping classes (Jacob, 2004)[vi]. For example, information categorisation from the purview of security and privacy controls involves characterisation of information based on a potential impact on organisational operations or assets, individuals, other organisations, and the overall country after loss of confidentiality, integrity, or availability of the information (NIST FIPS 199, 2004[vii]; OMB A-130, 2016[viii]).

Hjørland (1997) distinguishes classification and categorisation based on the level of ambition in the scheme for aggregating or segregating data. *Ad hoc* classification can be termed as categorization and helps to organise data in a useful way with a low level of ambition. However, pragmatic classification is a purpose-driven and more ambitious than just categorisation, and scientific classification, is highly systematic, backed by research, and useful for knowledge management. For example, when data is arranged either region-wise or state-wise at any urban data exchange platform, then the practice can be referred as categorisation of data. Data classification is referred as a pragmatic, if it can respond to the higher purpose, say, retrieval of a particular dataset of interest and do not just club them together based on the likeness. Data classification addressing hierarchical features or varied dimensions such as origin, sensitivity, and use model of data, simultaneously, can be helpful in multiple scenarios, e.g., data access control, monetisation, or pricing. Such structures can be designated as scientific classification.

Jacob (2004)[ix] defines data categorization as the process of data organization comprises of identification of resemblance across datasets to aggregate them in buckets, called categories. For example, companies in manufacturing sector categorise their data across three major sections, design and development, logistics and production, and the after-sales services (Mordinyi & Biffl, 2015)[x]. The data categorised across these sections are not strictly mutually exclusive and are categorised for ease of access across sections. Such categories are flexible enough to respond to new patterns of similarity between datasets.

Data classification abides to the systematic and consistent application of predefined principles governing the structure and interrelationship of classes (Jacob, 2004)[xi]. For example, personal data can be a class of any information that relates to an identified or identifiable natural person either directly or indirectly (Council of Europe, 1981)[xii] and the data other than personal data can classified as non-personal data (FFD Regulation, 2018)[xiii]. These two classes of data are mutually exclusive and non-overlapping.

Data classification techniques

The classification-as-scaffolding is built upon the notion of cognition scaffolding proposed by Clark (1997)[xiv] and Engelbart (1967) [xv]. This approach equates the classification system as a function of retrieval methods and also as knowledge storage or teaching devices that support cognitive economy using external structures of hierarchical relationships and standardised or patterned responses. The casting of a data classification structure as a knowledge storage device points to the collated understanding associated with the label and the definition of the particular class in a hierarchical relationships that helps practitioner to include or exclude particular dataset in a respective class. This understanding also aids reducing the burden of information associated with related subordinate or superordinate classes in any hierarchical setting. For example, a company decides to classify its data (i) as customer (iia) and employee data (iib), and each of them into personal data (iiia) and non-personal data (iiib). They also classify some of their datasets under personal data class as sensitive personal data (iiia1) in order to comply with the privacy regulations.

Classification-as-scaffolding reduces the strain on the class (e.g., personal data) to adhere knowledge aiding data classification as some of the knowledge is inherited from the superordinate classes (e.g. employee data) and will be inherited to its subordinate classes (e.g. sensitive personal data). Jacob (2001) reports that classification-as-scaffolding approach is a closed system with its relatively rigid or inflexible structure that is reluctant to adopt internal changes when practitioners apply it across domains.

Contrary to the prior approach, the classification-as-infrastructure is considered as an open system that is inherently flexible, receptive to internal modification, adaptive to conventional practices and biases across domains (Bowker and Star, 1999[xvi]; Jacob, 2001). This approach is based on the perspective of classification system depicted by Bowker and Star (1999) which is in contrast to the conventional approach of segregating entities across mutually exclusive classes with patently distinctive boundaries. They define classification as "a spatial, temporal, or spatio-temporal segmentation of the world" and classification system as "a set of boxes (metaphorical or literal) into which things can be put to then do some kind of work". They argue that, ideally, classification systems have consistent and unique classificatory principles such as sorting entities by their origin and decent or as per their temporal or functional order. However, in practice, classification systems may face the conceptual contradictions or people's disagreement, ignorance, or misunderstanding. For example, differences, disagreements, or contradictions over class labels, definitions, or inclusion-exclusion criteria might be present between the data classification structures proposed by guiding legal instruments like treaties and their enacted variants like laws in the member countries. With this contention of Bowker and Star, Jacob (2001) further illustrates the role of classification-as-infrastructure as a social conventions associated with technologies and organisational practices to support knowledge management in practice. It can be inferred that this approach views data classification as an infrastructure that is deeply hybridised with the technological developments and organisation's practices and is not adopted as a distinctive physical construct.

2. Research methodology

Literature review is used to understand the primary research question of interest in this paper, that is, to understand what the key existing data categories or taxonomies are to promote good data governance. Content analysis with a deductive approach is applied to the reviewed literature to increase the reliability of the coding scheme (Elo & Kyngäs, 2008[xvii]; Kyngäs et al., 2020[xviii]). The primary materials analysed include research articles, literature, and practice reviews, founding legal documents, international treaties, and law propositions to the American, European, and Indian legislative bodies (Table 1).

Categorisation matrix is created based on the asset management and information security management system guidelines in international standards, ISO 55000:2014, ISO 55001:2014, ISO 27000:2018, and FIPS 199. Post content analysis, the papers are organised as coded text into pre-determined categories in the categorisation matrix. In the final reporting phase, all the results are consolidated under each category.

The coding scheme is operationalised in two stages. The first-level codes represent five major categories (namely, regulatory & statutory compliance, technical requirements, socio-cultural & operational responsibilities, risk mitigation, and corporate policies or voluntary criteria) and subsequent sub-codes are created based on the categorisation criteria adopted in the relevant literature.

Table 1. Structured data analysis matrix and coding scheme

What are the key existing data categories or taxonomies to promote good data governance?	Regulatory & statutory compliance	Technical requirements	Socio-cultural & operational responsibilities	Risk mitigation	Corporate policies or voluntary criteria
	• Privacy & security • Origin of data • Nature of content • Level of access • Purpose of use	• Origin of data • Sensitivity of data • Use model • Nature & need of data collection • Format & structure of data • Complexity of content	• Business operations and intelligence • Assignment of ownership • Criticality for operational compliance • Business units	• Secrecy of data • Nature of producer • Nature of subject accessing data and assigned clearance	• Sensitivity as per company policy • Granularity of data under assessment • Data storage preference
Additional categorisation parameters	Focus on data-related fundamental rights and transparency	Focus on technical characteristic of data	Focus on non-technical and non-regulatory data-related practices	Focus on data security and data access related regulatory and non-regulatory aspects	Focus on parameters adopted under company policies and self-adopted constraints or criteria

3. Results: Archetypes of data classification criteria

The purpose of data classification influences the selection of the classification criteria, for example, proposed use of data, nature of data users, technicalities of data, and statutory compliances in relation to the application of data. Data classification can be carried out by an individual in personal capacities or as an organisation's employee either manually or with the help of programs, tools, and techniques. An individual or an organisation may propose to classify the data considering the several viewpoints including: to protect the fundamental rights and freedoms of the concerned to whom data belongs to; to fulfil the needs of risk management programs; to maintain the transparency in data management; to formulate the policies and managerial decisions; to optimise the data access; to monetise the implicit data value; to manage the solutions at either working directory/repository or data warehouse; to support and strengthen engineering production systems, units, and processes; to measure the properties; to manage or mine the knowledge from the data; etc. We cite a range of use cases and practical examples under these viewpoints establishing prevalence of the archetypes in the upcoming sections.

Data stewardship in designging classification criteria

Data ownership is one of the primary dimensions in devising data classification criteria. EU Commission's report (2016; p. 2)[xix] under the Digital Single Market initiative has affirmed the legal uncertainty around data ownership as the barrier for the free flow data. On the other hand, Banterle, F (2020; p. 216)[xx] alludes to the data ownership complexities by elaborating that a data ownership regime determined by contract, factual control, intellectual property like copyright and database rights, trade secrets, and data protection laws already results in a strong protection mechanism for data. Also, the gaps in law have been filled through contractual schemes and technological access restrictions. The non-rivalrous nature of data or overlapping intellectual property rights on data limit the control of data owners over their data to prevent the third party-reuse of data.

Some approaches directly address the concept of data control rather than data ownership (Poikola, et al, 2014; p. 9)[xxi]. Though it seems tempting to proclaim the data ownership to individuals, exclusive ownership rights are difficult to apply to data. In brief, legal ambiguity over the data ownership assignments or the presence of multiple tools controlling data ownership for diverse set of data-types add extra layer to the data classification perspectives.

In proposing the data classification criteria, whether at state-owned departments or private/non-government entities, regulatory and statutory compliances mandatory to data users, controllers, and processors contribute significantly. We embed and interpret these requirements in the first archetype (in the Section 3.1.1) below. The governments and administrators globally prescribe these requirements through standards (e.g. the NIST FIPS 199); laws focusing on information technology and information security management (e.g. the US's Information Technology Management Reform Act of 1996 i.e. Public Law 104-106, Federal Information Security Management Act of 2002 i.e. Public Law 107-347), data privacy (the US's Privacy Act of 1974 i.e. Public Law 93-579, the EU's General Data Protection Regulation, India's Personal Data Protection Bill), intellectual property (IP) management (e.g. copyright acts, database-related IP acts), regulations for promoting free flow of data (e.g. the FFD regulation i.e. EU 2018/1807); executive orders or ordinances (e.g. the US's EO12958 or EO13292); or guidelines and white papers by special task forces or plenary conferences of national or international agencies (e.g. the ECE's Conference of

European Statisticians report 2019, Report by the Committee of Experts on Non-Personal Data Governance Framework, 2020).

In summary, data classification drives public and private organisations to better manage data assets, monetise implicit data value, and aids businesses to enhance the market share or incentives. It helps to understand types of data available, the location of data, and the needs of regulatory or access controls to protect the data or achieve mission objectives. A key to achieving this is the need to adopt a clear data classification strategy backed by explicit data classification criteria. Each of the following archetypes discusses these classification criteria in greater detail.

3.1 Regulatory and statutory compliance

Regulatory and statutory requirements refer to laws and guidelines focusing on the data rights of the stakeholders and transparency in data governance. This archetype encompasses the data classification criteria designed to address the regulatory and statutory requirements prescribed by various national and international agencies, standard-setting bodies, special task force, or plenary conferences. The criteria categorised under this archetype particularly focus on fulfilment of the legal provisions protecting fundamental rights and freedoms of stakeholders, enhancing transparency in the data management across the data value chain, characterising the nature of data assets as per statutory distinction, and prevent or combat crime. This archetype is distinct from the 'risk mitigation' archetype (in Section 3.4) and underlying criteria whereby the later addresses data security and data access control needs from statutory viewpoint.

The key to protecting the fundamental rights of the stakeholder or data owners is to give them control over their ability to decide about the fate of their data. This archetype consolidates the classification criteria facilitating compliance of privacy regulations or the right of respect for private life, address the right of self-determination, and abides by the right of secrecy of correspondence. Even though these provisions and their interpretations seem similar, they are slightly different and not identical from the viewpoint of statutory frameworks (Warken, C., 2018)[xxii]. For example, the personal data classification aiming to minimise the threat to an individual's right to privacy strengthen right of respect for life (Mróz, K., 2020)[xxiii]. In practice, companies adopt 'policy-by-design' approach focusing on data classification of employee data to have transparency and grant control to the employees. Data-types in such use cases can use labels such as 'do not use', 'use for statistical purposes only (anonymise)', 'use with care - can be sensitive', 'use it, no problem', and 'I don't know' (Sahqani, W., 2021)[xxiv]. Similar data classification criteria respecting the right to self-determination facilitate an individual's control over data to decide grant of data-access parameters. The Nordic model, MyData, developed on this core idea equips an individual to be in control of their own data (Poikola, et al, 2014)[xxv]. On the other hand, the right to secrecy of correspondence can be achieved by framing data classification criteria abiding confidentiality.

The privacy acts and frameworks globally distinguish data based on the identifiability and sensitivity of the content thereunder. Personal data and non-personal data are major data-types mostly governed by the separate regulations such as General Data Protection Regulation (i.e. EU 2016/679) or India's Personal Data Protection Bill for personal data, whereas, EU's Framework for the Free Flow of Non-personal data (i.e. EU 2018/1807) or India's guidelines in report by the Committee of Experts on Non-Personal Data Governance Framework, 2020 for non-personal data. Most of the countries added an extra layer to identifiability by framing a data-type with higher level of sensitivity, known as sensitive personal data. Another data-type, known as mixed data which stands for the both personal and non-personal data linked to each other inextricably (FFD Guidance, 2019)[xxvi]. Another set of data classification characterises data based on the output of data processing (refer Table 3A) e.g. identified data, pseudonymised data, anonymised data, and aggregated data. Identified data equates to personally identifiable data elements. The pseudonymised data-type are considered non-personal in nature, unless no additional information is clubbed to recover an individual's identifiability (GDPR, 2016)[xxvii]. However, anonymised data remains unidentifiable even after the fusion of additional data (FFD Guidance, 2019)[xxviii]. The aggregated data loses individual's indefinability due to aggregation of individual-level data (Hashimzade, N. et al, 2017)[xxix].

3.2 Addressing technical requirements

This section refers classifying data assets based on their technical characteristics, technical users and usability of data, and knowledge management viewpoint. Understanding technical characteristics of the data may often be the first step in developing data classification criteria. This second archetype is distinct from others with its unique features of data classification criteria primarily seeking technical dimensions and measures of the data, or the content and structure of the data. It also focuses on the dimensions and defintions prescribed by Allen and

Cervo (2015)[xxx] namely, completeness (level of data missing or unusable), conformity (degree of data stored in a nonstandard format), consistency (level of conflicting information), accuracy (degree of agreement with an identified source of correct information), uniqueness (level of non-duplicates), integrity (degree of data corruption), validity (level of data matching a reference), and timeliness (degree to which data is current and available for use in the expected time frame).

This archetype includes the technical data classification practices adopted at data warehouses, working directories, or repositories. The data classification categories hereunder can facilitate different levels of summarisation such as metadata (the data defining warehouse data), current detailed data (mostly stored on disk), older detailed data (usually on tertiary storage), lightly summarized data, and highly summarized data (might be physically housed or not) (Han et al., 2012)[xxxi]. On the parallel lines, Krishnan (2003)[xxxii] reports data-driven integration for data warehouses where all the data in an organisation are segmented with respect to format and structure of data-type and associated data processing requirements abiding the business rules embedded in program workflows. Characteristics and examples of processed data categories are captured in the following Table 3B. The data classification categories can also be designed for file monitoring and tracking as seen in the case of Git index, file management, and Git status report identifying file-type as tracked (file in repository and file staged in index), ignored (repository files declared invisible or ignored), and untracked (files excluded from other two) (Loeliger and McCullough, 2012)[xxxiii].

Such technical data classifications can also be carried out for fuelling knowledge management activities such as targeting data mining applications and construction knowledge discovery methods and processes. Stundner and Al-Thuwaini (2001)[xxxiv] report an interesting example of modelling methods like neural networks used to integrate different types of data into reservoir management. The model employs depth-related data (e.g. well logs, drilling data, core data), well properties (e.g. PI, skin factor, location), time series data (e.g. pressures, production history, well tests), and areal distributions/layer data (e.g. OOIP, permeability, etc.). The data classification can also be adopted for facilitating knowledge-led decision making or optimising time and effort for collection, processing, and quality control processing of data. Grundstein and Rosenthal-Sabroux (2003)[xxxv] have shown one such data classification method for the ease of decision making by extended company's employees into three data-types namely, main-stream-data, shared-data, and source-of-knowledge-data.

Following Table 3B captures the various classification criteria and practices considering technical features of data.

3.3 Managing operational responsibilities

This section discusses profiling of the data assets to facilitate the operational efficiency and socio-cultural responsibilities acknowledged to individuals or organisations. This archetype taps into the data classification aiming to fulfil the sociocultural and operational responsibilities bore by the individuals or organisations. Data classification criteria, hereunder, focus on boosting an organisation's operational efficiency. They help to classify the data at organisation for process-oriented tasks or goal-oriented tasks (Kato, et al., 2020) or facilitate design and development, logistics and production, or after sale services (Motohashi, K., 2017). Companies are generally advised (Gregg, M., 2006)[xxxvi] to adopt classification practices for both paper and electronic documents with the labels such as *public* data obtained by anyone inside or outside company; *internal* data not accessible outside the company; the data with *limited distribution* accessible to individuals with necessary clearances and each copy is uniquely identifiable and additional copies are never made; and the *personal* for data related to the employee's details.

Companies mainly use two kinds of data, customer data and supplier data (Motohashi, 2007)[xxxvii]. Customer and product hierarchy management are highly sought across companies and entail customer and product data relationships to represent company's organizational structures. Hierarchical classification of the master data is the critical first step and can provide superior insights helpful to market campaigns, cross-selling, and up-selling. (Allen & Cervo, 2015)[xxxviii] Marr (2016)[xxxix] has also identified a challenge of managing huge volume and increasing growth rate of data hampering the ability to analyse it.

Public authorities have launched government open data initiatives and smart city missions. They are leveraging digital public infrastructure and data obtained therefrom for policymaking. Public sector is employing predictive modelling and other data science tools for prevention rather than for reactionary or remedial purposes. For example, government departments are using data mining to discover tax fraud based on links between companies and known characteristics of offenders (Barbero et al., 2016)[xl].

3.4 Risk mitigation

This section discusses adoption of appropriate safeguards to achieve desired expectations of data security and data access for both user and system information. This archetype deals with the data classification criteria ensuring data security and data access of user information and system information. Data classification is the fundamental first step to effective risk management in any organisation. FISMA (2002) has prescribed three data security objectives namely, confidentiality, integrity, and availability. On top of that, FIPS 199 (2004) defines three levels of potential impact on an individual or organisation after the breach of these data security objectives namely, low, moderate, and high. These levels of impact correspond to limited, serious, and severe/catastrophic adverse effects, respectively, on the organizational operations, organizational assets, or individuals associated. These adverse effect can contribute various levels of damages to organisational assets, financial loss, and harm to individuals.

The archetype also encompasses data access-oriented classification criteria enabling either the enhanced access to data, to grant a role-based access control, or to provide a level-based access control on various data-types. A range of legal instruments in Europe, as cited in Table 3D, enact the tiered provisions of classifying electronic data into categories called, *subscriber data*, *traffic data*, and *content data*. These categories were formed earlier (through the Council of Europe's Cybercrime Convention) to help the law enforcement agencies to access and process electronic communication data which were further extended (through the European Unions' proposed e-evidence regulation)[xli] to electronic data, in general (Warken, C., 2019). Article 2 of the proposed regulation distinguishes electronic data in to four categories. The *subscriber data* includes user's identity, types of services used, duration of such use, but, excludes sensitive data such as authentication data created or provided by the user. The *access data* includes data related to commencement or termination of sessions or service including sign-in or sign-off details, user's IP address, the interface used by user, and user ID. The *transactional data* relates to the provision of a service offered by a service provider that serves to provide context or additional information about such service. It includes the source and destination of message or similar interaction, device's location data, date, time, duration, size, route, format, protocol used, and the compression type. The *content data* includes any stored data in a digital format such as text, voice, videos, images, and sound other than subscriber, access or transactional data. In fact, the proposed classification creates new data category called *access data*, separated from the *transaction data*.

The information that is excluded from subscriber data, such as personal passwords or unlock keys, service bills or usage history, and others, is further categorised under the *residual* category (by the Dutch law) where higher level of order from prosecutor is required to access this information. On the other hand, the US's Stored Communication Act 1986 and the Clarifying Lawful Overseas Use of Data Act 2018 classify electronic data into two categories, first – content data and another – as a combination of the subscriber data and the traffic data (Warken, C. 2019).

Data access focused classification criteria try to ease out the concerns and major obstacles regarding cross-border access to electronic evidence. Data classification can also be based on the level of access granted by IT personnel as per predefined policies and procedures. Caballero (2014)[xlii] has noted that the discretionary access control has become less popular recently which offers end data user or data creator to define the data access levels and mandatory access control is 'more of a militant style' in granting blanket access to a particular level of members in the organisation.

3.5 Meeting internal policies and use requirements

This section discusses data classification based on company's internal policy provisions, voluntary guidelines, audit requirements and standard operating procedures to classify data assets of an organisation. This archetype encompasses the data classification criteria focusing on corporate policies and adopted voluntary criteria in an organisation. Organisations try to monetise their data assets uniquely, analytically, and synthetically. The logical first step in achieving the successful data-driven business model is identifying and sorting the data. In such mapping exercises, companies monitor the nature of data use, internally and externally (Motohashi, K., 2017). This further helps in managing data auditing activities at an organisation.

The data audit framework by Jones and team (2009)[xliii] finds the data classification as a crucial step that sets the scope of data auditing activities at an organisation. They advise organisations to classify their data in three categories, *vital*, *important*, and *minor* data. The vital datasets correspond to the functioning, efficient management, and protection of the organisation. The category includes the datasets that are frequently used by the organisation, being continuously created or added to, or are offered to external clients. The important datasets include the data for which the organisation is responsible, that are less frequently used, or that can be potentially used to offer services in future. Minor data category includes the data that is not explicitly required for the organisation, or the data for which the organisation no longer wants to host the responsibility. The granularity of the audit framework-based classification can be increased by sub-classifying the important data as per the needs to *retain* and *manage it on-site* or the minor data considering the need of *active data management* or *its disposal*.

Data storage preferences can also lead to data classes such as *primary data*, *backup data*, *archival data* (Assunção and Lefèvre, 2012). Universities and academic organisations are adopting guidelines on data classification focusing on information security where data is classified into *restricted data*, *private data*, or *public data* (Carnegie Mellon University, 2008)[xliv]. Restricted data seeks highest level of security controls as it includes an information whose unauthorised disclosure, alteration, or destruction can cause significant level of risk to an institutions and its affiliates. Such risk can be moderate in the case of private data and little or no risk in the case of public data.

Organisations also consider the durability and depreciating values of data assets while framing the data classification criteria to identify *perishable data* whose value declines once it is used, and *durable data* whose value of durable holds up over times (Stahl, et al, 2010)[xlv]. Stahl and team showed the application of these categories in decision making for optimal sampling and pricing of information products. Organisation seek to monetise their proprietary data internally by fuelling new product development, increase sales, effective marketing, cost reduction in manufacturing, improving existing products or manufacturing processes, and strengthening overall business management. To evaluate this, Kazuyuki (2017) surveyed big data use in the Japanese manufacturing firms. He identified that use of different types of data at company-level demonstrates a higher performance impact in Japanese manufacturing sector along with the variation in different usage styles by firm sizes. The study reported the frequently used data categories across three sections of production as – CAD data, CAE simulation, material library data (in design and development); manufacturing process data, logistics location data, purchase data, logistics and delivery data (in logistics and production); equipment operations data, customer complaint data, product defect data, call centre data (in the after-sales services). On the parallel lines, it is seen that the classification of data at mechatronic units (combination of mechanical, electrical, and control related components) of control system engineering or cyber-physical production systems engineering enhance functionalities of automation systems. The involved data is generally classified according (1) to the related engineering discipline, (2) to the plant structure, or (3) to the data structures described. (Mordinyi & Biffl, 2015)[xlvi].

It is interesting to see how data-driven organisations like Google classify their data. Google in its privacy policy (effective February 4, 2021)[xlvii] classifies the *'Information that Google collects'* into two major categories '*Things you create or provide to us*' and '*Information we collect as you use our services*'. The prior category includes information submitted by users while creating or using Google account. The latter category is further sub-classified into three namely, '*Your apps, browsers & devices'*, '*Your activity*', and '*Your location information*', whereas they include information collected by Google while users engage with their services such as information about users' apps, browsers, and devices, and users' activity and location. A portion disclosed after these data categories talks about the ways by which data is collected from third-party sources such as publically available sources, or their trusted partners including marketing and security partners, advertisers, and research services. Craddock et al. (2016) note that Google's privacy policy dated 29 August 2016 lacks the details of categories of personal data collected by them to ensure transparency in the data processing for number of reasons, most of them are still relevant for the Google's recent privacy policy. First, use of the terms 'that includes' or 'an example' makes the data categories less exhaustive. Second, details are more focused on way of information collection rather than the contents of the data categories. Third, data categories primarily focus on the 'provided' and 'observed' datatypes and lack focus on the 'derived' and 'inferred' datatypes. Fourth, data categories enlisted are not exhaustively linked to the purpose of data collection or its uses.

4. Discussion

The five archetypes of data classification criteria discussed in Section 3 are distinct, non-mutually exclusive, overlapping, and complementary. An organisation can adopt multiple classification criteria from different archetypes simultaneously to fulfil its interests. The archetype *risk mitigation* (Section 3.4) involves but is not limited to the legislative provisions related to data security and data access control. However, the archetype *regulatory and statutory compliance* (Section 3.1) only includes legally binding requirements related to privacy or other fundamental rights but categorically excludes security and access dimensions. These archetypes help advance practitioners' conceptual understanding by providing knowledge about the classification criteria-to-be-adopted and shaping practitioner's interaction with the respective criteria. These archetypes can change the way practitioners interact with data while adopting multiple classification perspectives to the same data in different contexts. The epistemological discussion in this paper also aims to advance the practitioners' conventional perspective of data classification that focuses on the nature of data assets (e.g., documents) by hybridising it with cognitive behaviour.

The archetypes proposed in the study seek the attention of practitioners to establish interoperability of data classification as a structure that varies sartorially. The study inspires data classification structure developers to consider the different classificatory labels representing the same concept or parallel concepts for the same label. The proposed archetypes can be a starting point in brainstorming to design the data classification structure, protocol, or organisational policy. The archetypes envisage helping organisations complying regulatory and statutory requirements of the state and pave a platform to extract implicit value by data silos. Practitioners can build new relationships across data silos and new bundling approaches whose value exceeds traditional groupings of information in an organisation. Adopting appropriate classification strategies and bundling approaches can potentially catalyse the way the organisation monetises its data or puts a price tag on its data assets.

There are limitations to these proposed archetypes as they are based on the existing examples of data classification criteria. Though the archetypes are ambitious to assist in exploring new relationships, they might need to be revisited regularly to check the inclusion of new or radical classification criteria. There is a need for contributions from academia and practicing communities to design standardised data classification structures that can facilitate the free flow of data across the data economy.

REFERENCES

1. (ISO/IEC 2382-1:1993) ISO/IEC 2382-1. (1993). Information technology — Vocabulary — Part 1: Fundamental terms. Now revised as ISO/IEC 2382:2015. https://www.iso.org/standard/7229.html
2. [1] (CNSSI 4009 Glossary, 2015) CNSS Secretariat. (2015). Committee on National Security Systems (CNSS) Glossary. https://www.serdp-estcp.org/content/download/47576/453617/file/CNSSI%204009%20Glossary%202015.pdf
3. [1] (IETF RFC 4949, 2007) Internet Security Glossary, Version 2. The IETF Trust. Network Working Group. https://tools.ietf.org/pdf/rfc4949.pdf
4. [1] (Banterle, F., 2020) Banterle, F. (2020) Data Ownership in the Data Economy: A European Dilemma. In: Synodinou TE., Jougleux P., Markou C., Prastitou T. (eds) EU Internet Law in the Digital Era. Springer, Cham. doi:10.1007/978-3-030-25579-4_9 https://link.springer.com/chapter/10.1007/978-3-030-25579-4_9
5. [1] Jacob, E., 2004. Classification and categorization: a difference that makes a difference. Library Trends 52 (3), Winter 2004, 515-540. https://www.ideals.illinois.edu/handle/2142/1686
6. [1] Jacob, E., 2004. Classification and categorization: a difference that makes a difference. Library Trends 52 (3), Winter 2004, 515-540. https://www.ideals.illinois.edu/handle/2142/1686
7. [1] (NIST FIPS 199, 2004) National Institute of Standards and Technology (2004) Standards for Security Categorization of Federal Information and Information Systems. (U.S. Department of Commerce, Washington, D.C.), Federal Information
8. Processing Standards Publication (FIPS) 199. https://doi.org/10.6028/NIST.FIPS.199
9. [1] (OMB A-130, 2016) Office of Management and Budget Memorandum Circular A-130, Managing Information as a Strategic Resource, July 2016. https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/circulars/A130/a13
10. 0revised.pdf
11. [1] Jacob, E., 2004. Classification and categorization: a difference that makes a difference. Library Trends 52 (3), Winter 2004, 515-540. https://www.ideals.illinois.edu/handle/2142/1686
12. [1] Jacob, E., 2004. Classification and categorization: a difference that makes a difference. Library Trends 52 (3), Winter 2004, 515-540. https://www.ideals.illinois.edu/handle/2142/1686
13. [1] (Council of Europe, 1981) Council of Europe. (1981). Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data. (ETS No. 108). Article 2(a). https://rm.coe.int/1680078b37
14. [1] (FFD Regulation, 2018) Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union. Article 3(1). https://eur-lex.europa.eu/eli/reg/2018/1807/oj

15. [1] (FFD Guidance, 2019) Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union. (2019). Section 2.2. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2019:250:FIN
16. [1] Mordinyi, R., & Biffl, S. (2015). Versioning in Cyber-physical Production System Engineering -- Best-Practice and Research Agenda. 2015 IEEE/ACM 1st International Workshop on Software Engineering for Smart Cyber-Physical Systems. doi:10.1109/sescps.2015.16
17. [1] Jacob, E., 2004. Classification and categorization: a difference that makes a difference. Library Trends 52 (3), Winter 2004, 515-540. https://www.ideals.illinois.edu/handle/2142/1686
18. [1] Markman, E. M. (1989). Categorization and naming in children: Problems of induction. Cambridge,
19. MA: MIT Press. https://mitpress.mit.edu/books/categorization-and-naming-children
20. [1] Clark, A. Being there: putting brain, body, and world together again. Cambridge, Mass.: MIT Press, 1997.
21. [1] Engelbart, D.C. A conceptual framework for the augmentation of man’s intellect. In: Howeron, P.E., ed. Vistas in information handling. Washington, DC: Spartan Books, 1963, 1–29.
22. [1] Bowker, G.C. and Star, S.L. Sorting things out: classification and its consequences. Cambridge, Mass.: MIT Press, 1999.
23. [1] Webster, J. & Watson, R.T. (2002) 'Analysing the past to prepare for the future: writing a literature review', MIS Quarterly, Vol. 26, No.2, pp.13-23.
24. [1] Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. Journal of Advanced Nursing, 62(1), 107–115. doi:10.1111/j.1365-2648.2007.04569.x
25. [1] Kyngäs, H., Mikkonen, K., & Kääriäinen, M. (Eds.). (2020). The Application of Content Analysis in Nursing Science Research. doi:10.1007/978-3-030-30199-6
26. [1] Horowitz, B. M. (2019). Policy Issues Regarding Implementations of Cyber Attack: Resilience Solutions for Cyber Physical Systems. Artificial Intelligence for the Internet of Everything, 87–100. doi:10.1016/b978-0-12-817636-8.00005-3 https://www.sciencedirect.com/science/article/pii/B9780128176368000053
27. [1] Xie, I., & Matusiak, K. K. (2016). New developments and challenges. Discover Digital Libraries, 319–339. doi:10.1016/b978-0-12-417112-1.00011-9 https://www.sciencedirect.com/science/article/pii/B9780124171121000119
28. [1] European Commission (2020), European free flow of data initiative within the Digital Single Market —
29. Inception impact assessment. https://ec.europa.eu/smart-regulation/roadmaps/docs/2016_cnect_001_free_flow_data_en.pdf
30. [1] Banterle F. (2020) Data Ownership in the Data Economy: A European Dilemma. In: Synodinou TE., Jougleux P., Markou C., Prastitou T. (eds) EU Internet Law in the Digital Era. Springer,

Cham. doi:10.1007/978-3-030-25579-4_9 https://link.springer.com/chapter/10.1007/978-3-030-25579-4_9

31. [1] Poikola, A., Kuikkaniemi, K. Honko, H. (2014). MyData – A Nordic Model for human-centered personal data management and processing. Ministry of Transport and Communications (Finland). urn.fi/URN:ISBN:978-952-243-455-5 https://julkaisut.valtioneuvosto.fi/handle/10024/78439
32. [1] Warken, C. 2018. Classification of Electronic Data for Criminal Law Purposes. EUCRIM Issue 4/2018 pp 226 – 234 https://eucrim.eu/articles/classification-electronic-data-criminal-law-purposes/ http://sci-hub.se/10.30709/eucrim-2018-023
33. [1] Mróz, K. (2020). Threats to the individual's right to privacy in relation to processing of personal data in order to prevent and combat crime. Ius Novum, 14(1), 82-97. doi:10.26399/iusnovum.v14.1.2020.05/k.mroz https://iusnovum.lazarski.pl/iusnovum/article/view/1080
34. [1] W. Sahqani and L. Turchet, "Co-designing Employees' Data Privacy: a Technology Consultancy Company Use Case," 2021 28th Conference of Open Innovations Association (FRUCT), Moscow, Russia, 2021, pp. 398-406, doi: 10.23919/FRUCT50888.2021.9347593.
35. [1] Poikola, A., Kuikkaniemi, K. Honko, H. (2014). MyData – A Nordic Model for human-centered personal data management and processing. Ministry of Transport and Communications (Finland). urn.fi/URN:ISBN:978-952-243-455-5 https://julkaisut.valtioneuvosto.fi/handle/10024/78439
36. [1] (FFD Guidance, 2019) Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union. (2019). Section 2.2. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2019:250:FIN
37. [1] (GDPR, 2016) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Article 4(5). https://eur-lex.europa.eu/eli/reg/2016/679/oj
38. [1] (FFD Guidance, 2019) Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union. (2019). Section 2.2. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2019:250:FIN
39. [1] Hashimzade, N., Myles, G., Black, J. (2017). A Dictionary of Economics. Oxford University Press. doi:10.1093/acref/9780198759430.001.0001. ISBN 978-0-19-875943-0. https://www.oxfordreference.com/view/10.1093/acref/9780198759430.001.0001/acref-9780198759430
40. [1] Allen, M., & Cervo, D. (2015). Data Quality Management. Multi-Domain Master Data Management, 131–160. doi:10.1016/b978-0-12-800835-5.00009-9 https://www.sciencedirect.com/science/article/pii/B9780128008355000099
41. [1] Han, J., Kamber, M., & Pei, J. (2012). Data Warehousing and Online Analytical Processing. Data Mining, 125–185. doi:10.1016/b978-0-12-381479-1.00004-6

42. [1] Krishnan, K. (2013). Integration of Big Data and Data Warehousing. Data Warehousing in the Age of Big Data, 199–217. doi:10.1016/b978-0-12-405891-0.00010-6 https://linkinghub.elsevier.com/retrieve/pii/B9780124058910000106
43. [1] Loeliger, Jon; McCullough, Matthew. 2012. Version Control With Git: Powerful Tools and Techniques for Collaborative Software Development. https://www.google.co.in/books/edition/Version_Control_with_Git/qIucp61eqAwC
44. [1] Stundner, M., & Al-Thuwaini, J. S. (2001). How Data-Driven Modeling Methods like Neural Networks can help to integrate different Types of Data into Reservoir Management. SPE Middle East Oil Show. doi:10.2118/68163-ms
45. [1] Grundstein, M. and C. Rosenthal-Sabroux. “Three Types of Data for Extended Company's Employees: A Knowledge Management Viewpoint.” (2003). https://www.academia.edu/download/3485489/irma03_lastversioncorrected_.pdf
46. [1] Gregg, M. (2006). The People Layer. Hack the Stack. Syngress. ISBN 9781597491099. 353-400. doi: 10.1016/B978-159749109-9/50013-7. https://www.sciencedirect.com/science/article/pii/B9781597491099500137
47. [1] Motohashi, K. 2017. Survey of Big Data Use and Innovation in Japanese Manufacturing Firms. Policy Discussion Papers 17027, Research Institute of Economy, Trade and Industry (RIETI). https://ideas.repec.org/p/eti/polidp/17027.html
48. [1] Allen, M., & Cervo, D. (2015). Data Quality Management. Multi-Domain Master Data Management, 131–160. doi:10.1016/b978-0-12-800835-5.00009-9
49. [1] Marr, B. (2016). Big data in practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results. Wiley.
50. [1] Martina Barbero, Jo Coutuer, Régy Jackers, Karim Moueddene, Els Renders, Wim Stevens, Yves Toninato, Sebastiaan van der Peijl, Dimitry Versteele. 2016. European Commission DG INFORMATICS (DG DIGIT). 2016 https://datos.gob.es/sites/default/files/blog/file/dg_digit_study_big_data_analytics_for_policy_making.pdf
51. [1] FIPS 199. (2004). Standards for Security Categorization of Federal Information and Information Systems. pp. 6. doi:10.6028/NIST.FIPS.199 https://csrc.nist.gov/publications/detail/fips/199/final
52. [1] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018PC0225
53. [1] Warken, C., van Zwieten, L., & Svantesson, D. (2019). Re-thinking the categorisation of data in the context of law enforcement cross-border access to evidence. International Review of Law, Computers & Technology, 1–21. doi:10.1080/13600869.2019.1600871
54. [1] European Commission Services. 2017. Improving cross-border access to electronic evidence: Findings from the expert process and suggested way forward. https://ec.europa.eu/home-affairs/sites/default/files/docs/pages/20170522_non-paper_electronic_evidence_en.pdf

55. [1] Tiwari, B., & Kumar, A. (2015). Role-based access control through on-demand classification of electronic health record. International Journal of Electronic Healthcare, 8(1), 9. doi:10.1504/ijeh.2015.071637
56. [1] Caballero, A. (2014). Information Security Essentials for IT Managers. Managing Information Security, 1–45. doi:10.1016/b978-0-12-416688-2.00001-5
57. [1] Jones, S., Ross, S., Ruusalepp, R. (2009). Data Audit Framework (DAF) Methodology. Version 1.8. HATII, University of Glasgow. https://www.data-audit.eu/DAF_Methodology.pdf https://www.data-audit.eu/docs/DAF_iPRES_paper.pdf
58. [1] Carnegie Mellon University. (2008). Guidelines for Data Classification. Carnegie Mellon University, Information Security Office, Computing Services. https://www.cmu.edu/iso/governance/guidelines/data-classification.html
59. [1] Florian Stahl, Daniel Halbheer, Oded Koenigsberg, and Donald R. Lehmann. (2010). Sampling information goods: How much should be free? https://www0.gsb.columbia.edu/mygsb/faculty/research/pubfiles/3598/sampling_information_goods.pdf
60. [1] Mordinyi, R., & Biffl, S. (2015). Versioning in Cyber-physical Production System Engineering -- Best-Practice and Research Agenda. 2015 IEEE/ACM 1st International Workshop on Software Engineering for Smart Cyber-Physical Systems. doi:10.1109/sescps.2015.16
61. [1] Google. (2021). Google's Privacy Policy. https://policies.google.com/privacy
62. [1] OECD Working Party on Security and Privacy in the Digital Economy, Summary of the OECD Privacy Expert Roundtable 'Protecting Privacy in
63. a Data-driven Economy: Taking Stock of Current Thinking' http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=dsti/
64. iccp/reg(2014)3&doclanguage=en
65. [1] Borek, A., Parlikad, A. K., Webb, J., & Woodall, P. (2014). Software Tools. Total Information Risk Management, 237–269. doi:10.1016/b978-0-12-405547-6.00012-2 https://www.sciencedirect.com/science/article/pii/B9780124055476000122
66. [1] Wun-Young, L., & Hirao, J. (2009). Chapter 7 - Overview of Auditing, SAP Security Configuration and Deployment. Syngress, 325-352. ISBN 9781597492843. doi: 10.1016/B978-1-59749-284-3.00007-7. https://www.sciencedirect.com/science/article/pii/B9781597492843000077
67. [1] Borek, A., Parlikad, A. K., Webb, J., & Woodall, P. (2014). Software Tools. Total Information Risk Management, 237–269. doi:10.1016/b978-0-12-405547-6.00012-2 https://www.sciencedirect.com/science/article/pii/B9780124055476000122
68. [1] Krishnan, K. (2013). Integration of Big Data and Data Warehousing. Data Warehousing in the Age of Big Data, 199–217. doi:10.1016/b978-0-12-405891-0.00010-6 https://www.sciencedirect.com/science/article/pii/B9780124058910000106

69. [1] Sint, R., Shaffert, S., Stroka, S., and Ferstl, R. 2009. Combining unstructured, fully structured and semi-structured information in semantic wikis. 4th Semantic Wiki Workshop (SemWiki 2009) at the 6th European Semantic Web Conference (ESWC 2009), Hersonissos, Greece, June 1st, 2009. Proceedings. http://ceur-ws.org/Vol-464/paper-14.pdf
70. [1] The DCC Curation Lifecycle Model https://www.dcc.ac.uk/guidance/curation-lifecycle-model
71. [1] Han, J., Kamber, M., & Pei, J. (2012). Introduction. Data Mining, 1–38. doi:10.1016/b978-0-12-381479-1.00001-0 https://www.sciencedirect.com/science/article/pii/B9780123814791000010
72. [1] Sint, R., Shaffert, S., Stroka, S., and Ferstl, R. 2009. Combining unstructured, fully structured and semi-structured information in semantic wikis. 4th Semantic Wiki Workshop (SemWiki 2009) at the 6th European Semantic Web Conference (ESWC 2009), Hersonissos, Greece, June 1st, 2009. Proceedings. http://ceur-ws.org/Vol-464/paper-14.pdf
73. [1] Han, J., Kamber, M., & Pei, J. (2012). Introduction. Data Mining, 1–38. doi:10.1016/b978-0-12-381479-1.00001-0 https://www.sciencedirect.com/science/article/pii/B9780123814791000010
74. [1] Sint, R., Shaffert, S., Stroka, S., and Ferstl, R. 2009. Combining unstructured, fully structured and semi-structured information in semantic wikis. 4th Semantic Wiki Workshop (SemWiki 2009) at the 6th European Semantic Web Conference (ESWC 2009), Hersonissos, Greece, June 1st, 2009. Proceedings. http://ceur-ws.org/Vol-464/paper-14.pdf
75. [1] Sint, R., Shaffert, S., Stroka, S., and Ferstl, R. 2009. Combining unstructured, fully structured and semi-structured information in semantic wikis. 4th Semantic Wiki Workshop (SemWiki 2009) at the 6th European Semantic Web Conference (ESWC 2009), Hersonissos, Greece, June 1st, 2009. Proceedings. http://ceur-ws.org/Vol-464/paper-14.pdf
76. [1] Han, J., Kamber, M., & Pei, J. (2012). Introduction. Data Mining, 1–38. doi:10.1016/b978-0-12-381479-1.00001-0 https://www.sciencedirect.com/science/article/pii/B9780123814791000010
77. [1] OECD (2019), Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies, OECD Publishing, Paris, doi:10.1787/276aaca8-en. https://www.oecd.org/sti/enhancing-access-to-and-sharing-of-data-276aaca8-en.htm
78. [1] COM/2017/09 (2017), Communication From The Commission To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions "Building A European Data Economy", COM(2017) 9 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2017:9:FIN
79. [1] OECD (2019), Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies, OECD Publishing, Paris, doi:10.1787/276aaca8-en. https://www.oecd.org/sti/enhancing-access-to-and-sharing-of-data-276aaca8-en.htm
80. [1] Florian Stahl, Daniel Halbheer, Oded Koenigsberg, and Donald R. Lehmann. (2010). Sampling information goods: How much should be free? https://www0.gsb.columbia.edu/mygsb/faculty/research/pubfiles/3598/sampling_information_goods.pdf

81. [1] Florian Stahl, Daniel Halbheer, Oded Koenigsberg, and Donald R. Lehmann. (2010). Sampling information goods: How much should be free? https://www0.gsb.columbia.edu/mygsb/faculty/research/pubfiles/3598/sampling_information_goods.pdf

82. [1] Loeliger, Jon; McCullough, Matthew. 2012. Version Control With Git: Powerful Tools and Techniques for Collaborative Software Development. https://www.google.co.in/books/edition/Version_Control_with_Git/qIucp61eqAwC

83. [1] Loshin, D. (2011). Data Requirements Analysis. The Practitioner's Guide to Data Quality Improvement, 147–165. doi:10.1016/b978-0-12-373717-5.00009-9 https://www.sciencedirect.com/science/article/pii/B9780123737175000099

84. [1] Blackman, C., Forge, S. (2017). Data Flows — Future Scenarios: In-Depth Analysis for the ITRE Committee, 09-10, Table 1. http://www.europarl.europa.eu/RegData/etudes/IDAN/2017/607362/IPOL_IDA(2017)607362_EN.pdf

85. [1] Rumbold, J. M. M., & Pierscionek, B. K. (2018). What Are Data? A Categorization of the Data Sensitivity Spectrum. Big Data Research, 12, 49–59. doi:10.1016/j.bdr.2017.11.001 https://www.sciencedirect.com/science/article/abs/pii/S2214579617302010

86. [1] The Council of Europe's Convention on Cybercrime of 2001 (ETS No. 185) https://rm.coe.int/1680081561

87. [1] Article 126na, Dutch Code of Criminal Procedure https://www.legislationline.org/download/id/6416/file/Netherlands_CPC_am2012_en.pdf

88. [1] 18 United States Code, Chapter 121 Stored Wire and Electronic Communications and Transactional Records Access, §2703(c)(2) https://www.govinfo.gov/content/pkg/USCODE-2010-title18/html/USCODE-2010-title18-partI-chap121.htm

89. [1] Article 2(7) (Definitions), Proposal for a Regulation of the European Parliament and of the Council on European Production and Preservation Orders for electronic evidence in criminal matters, COM/2018/225 final - 2018/0108 (COD) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:225:FIN

90. [1] Article 18(3) (Production order), The Council of Europe's Convention on Cybercrime of 2001 (ETS No. 185) https://rm.coe.int/1680081561

91. [1] 18 United States Code, Chapter 121 Stored Wire and Electronic Communications and Transactional Records Access, §2702-2703 https://www.govinfo.gov/content/pkg/USCODE-2010-title18/html/USCODE-2010-title18-partI-chap121.htm

92. [1] Article 2(7) (Definitions), Proposal for a Regulation of the European Parliament and of the Council on European Production and Preservation Orders for electronic evidence in criminal matters, COM/2018/225 final - 2018/0108 (COD) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:225:FIN

93. [1] Article 15 and Article 19(3), The Council of Europe’s Convention on Cybercrime of 2001 (ETS No. 185) https://rm.coe.int/1680081561
94. [1] Article 1(d) (Definitions), The Council of Europe’s Convention on Cybercrime of 2001 (ETS No. 185) https://rm.coe.int/1680081561
95. [1] Article 1(d) (Definitions), The Council of Europe’s Convention on Cybercrime of 2001 (ETS No. 185) https://rm.coe.int/1680081561
96. [1] Article 2(10) (Definitions), Proposal for a Regulation of the European Parliament and of the Council on European Production and Preservation Orders for electronic evidence in criminal matters, COM/2018/225 final - 2018/0108 (COD) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:225:FIN
97. [1] Article 18(3), The Council of Europe’s Convention on Cybercrime of 2001 (ETS No. 185) https://rm.coe.int/1680081561
98. [1] Paragraphs 209 and 229, Explanatory Report to the Council of Europe’s Convention on Cybercrime of 2001 (ETS No. 185) https://rm.coe.int/16800cce5b
99. [1] Article 2(10) (Definitions), Proposal for a Regulation of the European Parliament and of the Council on European Production and Preservation Orders for electronic evidence in criminal matters, COM/2018/225 final - 2018/0108 (COD) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:225:FIN
100. [1] Article 2(8) (Definitions), Proposal for a Regulation of the European Parliament and of the Council on European Production and Preservation Orders for electronic evidence in criminal matters, COM/2018/225 final - 2018/0108 (COD) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:225:FIN
101. [1] Article 2(9) (Definitions), Proposal for a Regulation of the European Parliament and of the Council on European Production and Preservation Orders for electronic evidence in criminal matters, COM/2018/225 final - 2018/0108 (COD) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:225:FIN
102. [1] Criminal Procedure Code of the Kingdom of Netherlands (as of 2012) or Dutch Code of Criminal Procedure, Book 1, Title IVa, Section 7 – articles 126n through 126ng https://www.legislationline.org/download/id/6416/file/Netherlands_CPC_am2012_en.pdf
103. [1] Article 2 (Definitions), Proposal for a Regulation of the European Parliament and of the Council on European Production and Preservation Orders for electronic evidence in criminal matters, COM/2018/225 final - 2018/0108 (COD) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:225:FIN
104. [1] Article 4(3)(c), Proposal for a Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic

Communications) COM/2017/010 final - 2017/03 (COD) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52017PC0010

105. [1] Creative Commons. (2013). About The Licenses 4.0. https://creativecommons.org/licenses/ https://creativecommons.org/2013/11/25/ccs-next-generation-licenses-welcome-version-4-0/

106. [1] FIPS 199. (2004). Standards for Security Categorization of Federal Information and Information Systems. pp. 6. doi:10.6028/NIST.FIPS.199 https://csrc.nist.gov/publications/detail/fips/199/final

107. [1] FIPS 199. (2004). Standards for Security Categorization of Federal Information and Information Systems. pp. 6. doi:10.6028/NIST.FIPS.199 https://csrc.nist.gov/publications/detail/fips/199/final

108. [1] FIPS 199. (2004). Standards for Security Categorization of Federal Information and Information Systems. pp. 6. doi:10.6028/NIST.FIPS.199 https://csrc.nist.gov/publications/detail/fips/199/final

109. [1] Florian Stahl, Daniel Halbheer, Oded Koenigsberg, and Donald R. Lehmann. (2010). Sampling information goods: How much should be free? https://www0.gsb.columbia.edu/mygsb/faculty/research/pubfiles/3598/sampling_information_goods.pdf

110. [1] CNSSI 4009. (2015). Committee on National Security Systems (CNSS) Glossary. https://www.serdp-estcp.org/content/download/47576/453617/file/CNSSI%204009%20Glossary%202015.pdf

111. [1] Florian Stahl, Daniel Halbheer, Oded Koenigsberg, and Donald R. Lehmann. (2010). Sampling information goods: How much should be free? https://www0.gsb.columbia.edu/mygsb/faculty/research/pubfiles/3598/sampling_information_goods.pdf

112. [1] Florian Stahl, Daniel Halbheer, Oded Koenigsberg, and Donald R. Lehmann. (2010). Sampling information goods: How much should be free? https://www0.gsb.columbia.edu/mygsb/faculty/research/pubfiles/3598/sampling_information_goods.pdf

113. [1] (DAF Methodology, 2009) Jones, S., Ross, S., Ruusalepp, R. (2009). Data Audit Framework Methodology. Version 1.8. HATII, University of Glasgow. https://www.data-audit.eu/DAF_Methodology.pdf https://www.data-audit.eu/docs/DAF_iPRES_paper.pdf

114. [1] Google. (2021). Google's Privacy Policy. https://policies.google.com/privacy

115. [1] Creative Commons. (2019). Creative Commons Privacy Policy. https://creativecommons.org/privacy/#2_Information_We_Collect

116. [1] Carnegie Mellon University. (2008). Guidelines for Data Classification. Carnegie Mellon University, Information Security Office, Computing Services. https://www.cmu.edu/iso/governance/guidelines/data-classification.html

117. [1] NSF. (2005). Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. National Science Board. Appendix D. Digital Data Collections by Categories. https://www.nsf.gov/geo/geo-data-policies/nsb-0540-1.pdf https://www.nsf.gov/pubs/2005/nsb0540/nsb0540_11.pdf

118. http://www.arabidopsis.org/

119. http://www.plasmodb.org/bdbs.shtml

120. http://www.maizegdb.org/

121. http://canopy.evergreen.edu/home.asp

122. http://ligo.org/

123. http://simbad.u-strasbg.fr/Simbad

124. http://physics.nist.gov/PhysRefData/contents.html

(ISO/IEC 2382-1:1993) ISO/IEC 2382-1. (1993). Information technology — Vocabulary — Part 1: Fundamental terms. Now revised as ISO/IEC 2382:2015. https://www.iso.org/standard/7229.html

[ii] (CNSSI 4009 Glossary, 2015) CNSS Secretariat. (2015). Committee on National Security Systems (CNSS) Glossary. https://www.serdp-estcp.org/content/download/47576/453617/file/CNSSI%204009%20Glossary%202015.pdf

[iii] (IETF RFC 4949, 2007) Internet Security Glossary, Version 2. The IETF Trust. Network Working Group. https://tools.ietf.org/pdf/rfc4949.pdf

[iv] (Banterle, F., 2020) Banterle, F. (2020) Data Ownership in the Data Economy: A European Dilemma. In: Synodinou TE., Jougleux P., Markou C., Prastitou T. (eds) EU Internet Law in the Digital Era. Springer, Cham. doi:10.1007/978-3-030-25579-4_9 https://link.springer.com/chapter/10.1007/978-3-030-25579-4_9

[v] Jacob, E., 2004. Classification and categorization: a difference that makes a difference. Library Trends 52 (3), Winter 2004, 515-540. https://www.ideals.illinois.edu/handle/2142/1686

[vi] Jacob, E., 2004. Classification and categorization: a difference that makes a difference. Library Trends 52 (3), Winter 2004, 515-540. https://www.ideals.illinois.edu/handle/2142/1686

[vii] (NIST FIPS 199, 2004) National Institute of Standards and Technology (2004) Standards for Security Categorization of Federal Information and Information Systems. (U.S. Department of Commerce, Washington, D.C.), Federal Information
Processing Standards Publication (FIPS) 199. https://doi.org/10.6028/NIST.FIPS.199

[viii] (OMB A-130, 2016) Office of Management and Budget Memorandum Circular A-130, Managing Information as a Strategic Resource, July 2016.
https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/circulars/A130/a13
0revised.pdf

[ix] Jacob, E., 2004. Classification and categorization: a difference that makes a difference. Library Trends 52 (3), Winter 2004, 515-540. https://www.ideals.illinois.edu/handle/2142/1686

[x] Mordinyi, R., & Biffl, S. (2015). Versioning in Cyber-physical Production System Engineering -- Best-Practice and Research Agenda. 2015 IEEE/ACM 1st International Workshop on Software Engineering for Smart Cyber-Physical Systems. doi:10.1109/sescps.2015.16

[xi] Jacob, E., 2004. Classification and categorization: a difference that makes a difference. Library Trends 52 (3), Winter 2004, 515-540. https://www.ideals.illinois.edu/handle/2142/1686

[xii] (Council of Europe, 1981) Council of Europe. (1981). Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data. (ETS No. 108). Article 2(a). https://rm.coe.int/1680078b37

[xiii] (FFD Regulation, 2018) Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union. Article 3(1). https://eur-lex.europa.eu/eli/reg/2018/1807/oj

[xiv] Clark, A. Being there: putting brain, body, and world together again. Cambridge, Mass.: MIT Press, 1997.

[xv] Engelbart, D.C. A conceptual framework for the augmentation of man’s intellect. In: Howeron, P.E., ed. Vistas in information handling. Washington, DC: Spartan Books, 1963, 1–29.

[xvi] Bowker, G.C. and Star, S.L. Sorting things out: classification and its consequences. Cambridge, Mass.: MIT Press, 1999.

[xvii] Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. Journal of Advanced Nursing, 62(1), 107–115. doi:10.1111/j.1365-2648.2007.04569.x

[xviii] Kyngäs, H., Mikkonen, K., & Kääriäinen, M. (Eds.). (2020). The Application of Content Analysis in Nursing Science Research. doi:10.1007/978-3-030-30199-6

[xix] European Commission (2020), European free flow of data initiative within the Digital Single Market — Inception impact assessment. https://ec.europa.eu/smart-regulation/roadmaps/docs/2016_cnect_001_free_flow_data_en.pdf

[xx] Banterle F. (2020) Data Ownership in the Data Economy: A European Dilemma. In: Synodinou TE., Jougleux P., Markou C., Prastitou T. (eds) EU Internet Law in the Digital Era. Springer, Cham. doi:10.1007/978-3-030-25579-4_9 https://link.springer.com/chapter/10.1007/978-3-030-25579-4_9

[xxi] Poikola, A., Kuikkaniemi, K. Honko, H. (2014). MyData – A Nordic Model for human-centered personal data management and processing. Ministry of Transport and Communications (Finland). urn.fi/URN:ISBN:978-952-243-455-5 https://julkaisut.valtioneuvosto.fi/handle/10024/78439

[xxii] Warken, C. 2018. Classification of Electronic Data for Criminal Law Purposes. EUCRIM Issue 4/2018 pp 226 – 234 https://eucrim.eu/articles/classification-electronic-data-criminal-law-purposes/ http://sci-hub.se/10.30709/eucrim-2018-023

[xxiii] Mróz, K. (2020). Threats to the individual's right to privacy in relation to processing of personal data in order to prevent and combat crime. Ius Novum, 14(1), 82-97. doi:10.26399/iusnovum.v14.1.2020.05/k.mroz https://iusnovum.lazarski.pl/iusnovum/article/view/1080

[xxiv] W. Sahqani and L. Turchet, "Co-designing Employees' Data Privacy: a Technology Consultancy Company Use Case," 2021 28th Conference of Open Innovations Association (FRUCT), Moscow, Russia, 2021, pp. 398-406, doi: 10.23919/FRUCT50888.2021.9347593.

[xxv] Poikola, A., Kuikkaniemi, K. Honko, H. (2014). MyData – A Nordic Model for human-centered personal data management and processing. Ministry of Transport and Communications (Finland). urn.fi/URN:ISBN:978-952-243-455-5 https://julkaisut.valtioneuvosto.fi/handle/10024/78439

[xxvi] (FFD Guidance, 2019) Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union. (2019). Section 2.2. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2019:250:FIN

[xxvii] (GDPR, 2016) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Article 4(5). https://eur-lex.europa.eu/eli/reg/2016/679/oj

[xxviii] (FFD Guidance, 2019) Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union. (2019). Section 2.2. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2019:250:FIN

[xxix] Hashimzade, N., Myles, G., Black, J. (2017). A Dictionary of Economics. Oxford University Press. doi:10.1093/acref/9780198759430.001.0001. ISBN 978-0-19-875943-0. https://www.oxfordreference.com/view/10.1093/acref/9780198759430.001.0001/acref-9780198759430

[xxx] Allen, M., & Cervo, D. (2015). Data Quality Management. Multi-Domain Master Data Management, 131–160. doi:10.1016/b978-0-12-800835-5.00009-9 https://www.sciencedirect.com/science/article/pii/B9780128008355000099

[xxxi] Han, J., Kamber, M., & Pei, J. (2012). Data Warehousing and Online Analytical Processing. Data Mining, 125–185. doi:10.1016/b978-0-12-381479-1.00004-6

[xxxii] Krishnan, K. (2013). Integration of Big Data and Data Warehousing. Data Warehousing in the Age of Big Data, 199–217. doi:10.1016/b978-0-12-405891-0.00010-6 https://linkinghub.elsevier.com/retrieve/pii/B9780124058910000106

[xxxiii] Loeliger, Jon; McCullough, Matthew. 2012. Version Control With Git: Powerful Tools and Techniques for Collaborative Software Development. https://www.google.co.in/books/edition/Version_Control_with_Git/qIucp61eqAwC

[xxxiv] Stundner, M., & Al-Thuwaini, J. S. (2001). How Data-Driven Modeling Methods like Neural Networks can help to integrate different Types of Data into Reservoir Management. SPE Middle East Oil Show. doi:10.2118/68163-ms

[xxxv] Grundstein, M. and C. Rosenthal-Sabroux. "Three Types of Data for Extended Company's Employees: A Knowledge Management Viewpoint." (2003). https://www.academia.edu/download/3485489/irma03_lastversioncorrected_.pdf

[xxxvi] Gregg, M. (2006). The People Layer. Hack the Stack. Syngress. ISBN 9781597491099. 353-400. doi: 10.1016/B978-159749109-9/50013-7. https://www.sciencedirect.com/science/article/pii/B9781597491099500137

[xxxvii] Motohashi, K. 2017. Survey of Big Data Use and Innovation in Japanese Manufacturing Firms. Policy Discussion Papers 17027, Research Institute of Economy, Trade and Industry (RIETI). https://ideas.repec.org/p/eti/polidp/17027.html

[xxxviii] Allen, M., & Cervo, D. (2015). Data Quality Management. Multi-Domain Master Data Management, 131–160. doi:10.1016/b978-0-12-800835-5.00009-9

[xxxix] Marr, B. (2016). Big data in practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results. Wiley.

[xl] Martina Barbero, Jo Coutuer, Régy Jackers, Karim Moueddene, Els Renders, Wim Stevens, Yves Toninato, Sebastiaan van der Peijl, Dimitry Versteele. 2016. European Commission DG INFORMATICS (DG DIGIT). 2016 https://datos.gob.es/sites/default/files/blog/file/dg_digit_study_big_data_analytics_for_policy_making.pdf

[xli] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018PC0225

[xlii] Caballero, A. (2014). Information Security Essentials for IT Managers. Managing Information Security, 1–45. doi:10.1016/b978-0-12-416688-2.00001-5

[xliii] Jones, S., Ross, S., Ruusalepp, R. (2009). Data Audit Framework (DAF) Methodology. Version 1.8. HATII, University of Glasgow. https://www.data-audit.eu/DAF_Methodology.pdf https://www.data-audit.eu/docs/DAF_iPRES_paper.pdf

[xliv] Carnegie Mellon University. (2008). Guidelines for Data Classification. Carnegie Mellon University, Information Security Office, Computing Services. https://www.cmu.edu/iso/governance/guidelines/data-classification.html

[xlv] Florian Stahl, Daniel Halbheer, Oded Koenigsberg, and Donald R. Lehmann. (2010). Sampling information goods: How much should be free? https://www0.gsb.columbia.edu/mygsb/faculty/research/pubfiles/3598/sampling_information_goods.pdf

[xlvi] Mordinyi, R., & Biffl, S. (2015). Versioning in Cyber-physical Production System Engineering -- Best-Practice and Research Agenda. 2015 IEEE/ACM 1st International Workshop on Software Engineering for Smart Cyber-Physical Systems. doi:10.1109/sescps.2015.16

[xlvii] Google. (2021). Google's Privacy Policy. https://policies.google.com/privacy