KEMP IT LAW

IT Law at the Apex

DATA LAW

White Paper

Legal Aspects of Managing Data

Richard Kemp

October 2019

LEGAL ASPECTS OF MANAGING DATA

TABLE OF CONTENTS

TABLE OF FIGURES

LEGAL ASPECTS OF MANAGING DATA

A. INTRODUCTION

1. **'The world's most valuable resource is no longer oil but data'** ran the title of a leader in The Economist in May 2017, drawing a comparison between Standard Oil in the early 20th century and the world's most valuable listed companies today.[1] But, even as 'data as the new oil' has become a trope of the fourth industrial revolution, it's easy to overdo the parallels: oil is a finite natural resource whilst data is created by people – "the world's most renewable resource" in the words of Microsoft CEO Satya Nadella[2] - and oil used by one person can't be consumed by anyone else, where data can be used time and again without lessening its value.

2. **Data volumes created are doubling every two years or so**. Information and data are also infinite as expression and communication and as a resource, data volumes growing exponentially. In the April 2019 edition of its '*Data Age 2025*' white paper, research company IDC estimated that global data volumes will grow at a compound annual growth rate of 30% to 40% to reach 6x current levels by 2025. [3] Four things are driving this growth: hyperscale data centres at the cloud's core; proliferating compute capabilities at the edge; ubiquitous mobile phones and devices; and (particularly) connected sensors, where the Internet of Things ('**IOT**') "aims to do for information what electricity did for energy".[4] To put this growth in perspective, data volumes created have already doubled since The Economist's May 2017 leader.

3. **The value of personal data is becoming more measurable**. Attributing value to data is another area of difference between oil and data. Whilst the market accurately informs the price and value of oil, the price and value of data are opaque. Data valuation will depend on what is being looked at – financial market data differs from individuals' personal data – but two recent studies of personal data show how value is becoming more measurable. In March 2019 the US Democratic Party strategy group Future Majority found the value generated from Americans' personal data to be US$366 per person in 2018, set to rise to US$612 in 2022.[5] Then in July 2019 accounting firm EY published a study on valuing UK health care data which estimated that the entire dataset of the UK's National Health Service ('**NHS**') was worth £9.6bn per year in benefits, or £175 (US$217) per person across the 55 million record NHS estate.[6] These two studies show that individuals' personal data is currently worth in the low $00s per person per year and that value is rising.

---

[1] The Economist, Leaders, 6 May 2017 - https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data; see also '*Data, data everywhere*', The Economist, Special Report, 27 February 2010: "*data is becoming the new raw material of business: an economic input on a par with capital and labour*" - https://www.economist.com/special-report/2010/02/27/data-data-everywhere

[2] '*Tools and Weapons: the Promise and the Peril of the Digital Age*', Microsoft President Brad Smith and Carol Anne Brown, Hodder & Stoughton, 2019 at p.274.

[3] '*Data Age 2025 – the digitisation of the world from edge to core*', IDC White Paper sponsored by Seagate, April 2019 - https://www.seagate.com/gb/en/our-story/data-age-2025/

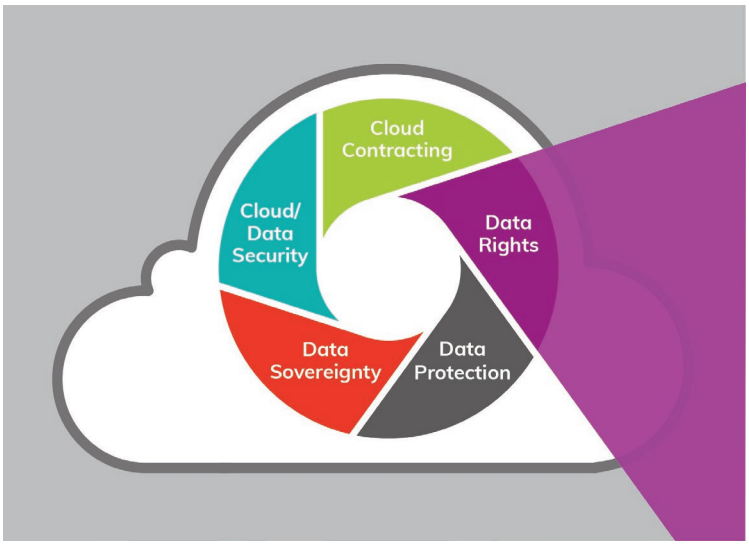
[4] '*Drastic falls in cost are powering another computer revolution*', The Economist, 12 September 2019 - https://www.economist.com/technology-quarterly/2019/09/12/drastic-falls-in-cost-are-powering-another-computer-revolution

[5] '*Who owns Americans' personal information and what is it worth?*', Robert Shapiro and Siddhartha Aneja, Future Majority, 8 March 2019, at p. 21.

[6] '*Realising the value of health care data: a framework for the future*', EY, 19 July 2019

4. **Data is elusive in legal terms**. For all the rapid growth in data volume, value and value measurability, data remains elusive in legal terms. When extracted, oil is goods under English law[7] and as such may be owned, bought, sold or stolen. Data on the other hand isn't a tangible and can't be bought and sold, at least in the same way; and under UK criminal law information has been held not to be intangible property either so it can't be stolen.[8] Although legally inert in and of itself however, different legal rights and duties apply to and act on data in different ways and use cases: a useful heuristic is '*there are no rights **in** data, but rights arise **in relation to** data*'. Reflecting the increasing value of data, the legal aspects of these rights (and the duties that are their converse) are developing rapidly. To quote The Economist again, in the context of IOT, "a world of connected sensors will generate huge amounts of data. It will also generate arguments about who can use and analyse those data".[9] These rights and duties - as intellectual property ('**IP**'), contract and regulation and especially in the context of big data - are the main subject of this white paper.

5. **Purpose and scope of this white paper**. Accordingly, the purpose of this white paper is to provide a practical guide to legal rights in data - what they are, how they arise and how they can be managed. Its primary audience is in-house legal counsel lawyering their organisation's data estate and data operations. **Section B** overviews data across a number of different verticals (financial services, insurance, air transport, recorded music, healthcare and the public sector) and developing data policy. **Section C** looks at different types of data before offering a common 8-layer framework for the legal analysis of data. **Sections D**, **E** and **F** work through each level of the framework: (1) platform infrastructure, (2) information architecture, (3) IP rights, (4) contract rights, (5) data regulation, (6) data protection, (7) information security and (8) data governance.



This white paper is one in an occasional series on aspects of IT law. Others include the legal aspects of artificial intelligence, cloud contracting, cloud security and IOT, and demystifying IT law.[10] This paper is not legal advice. It is written as at 30 September 2019[11] and from the standpoint of English law.

---

[7] See, for example, Benjamin, *Sale of Goods*, 10th Edition (Sweet & Maxwell, 2017), paragraph 1-087, pp. 75 & 76.

[8] *Oxford* v *Moss* ([1979] Crim LR 119) is authority that there is no property in data (in that case, confidential information in an exam question) as it was not 'intangible property' within the meaning of the Theft Act 1968.

[9] '*When humans are connected – what happens when humans are connected to smart machines*', the Economist, 13 September 2019 - https://www.economist.com/technology-quarterly/2019/09/12/hugo-campos-has-waged-a-decade-long-battle-for-access-to-his-heart-implant

[10] '*Legal Aspects of Artificial Intelligence*' (September 2018), '*Legal Aspects of Cloud Computing: Cloud Contracting*' (June 2019), '*Legal Aspects of Cloud Computing: Cloud Security*' (June 2018), '*Legal Aspects of the Internet of Things*' (June 2017), '*Demystifying IT Law*' (June 2018).

[11] We have not addressed here issues relating to the withdrawal of the UK from the EU (Brexit) and will update the paper if necessary when the position is clearer.

B. THE BUSINESS AND POLICY CONTEXTS: DATA IN KEY VERTICALS

6. **Introduction**. This section briefly looks at data in the context of financial markets (paragraph 7), open banking (para 8), insurance (para 9), air transport (para 10), recorded music (para 11), healthcare (para 12) and public (para 13) sectors before looking at the policy perspective and directions of travel (para 14).

7. **Financial market data**. The financial sector is one of the largest users of IT globally. Trading platforms – complex computer systems to buy and sell securities, derivatives and other financial instruments – are its beating heart and data its lifeblood. Based on an ecosystem of exchanges, index providers, data vendors and data users (asset managers on the buy-side and banks and brokers sell-side), these platforms generate market data, indexes, reference data and analytics and together form the world's financial market data/analysis industry. Increasing regulatory requirements, the growing ability of AI to interpret data and rising market volatility are currently fueling increasing demand both for financial market data (where global revenues hit $30bn for the first time in 2018) and exchanges (whose global revenues were $34bn in 2018).[12]

   In legal terms this complex ecosystem is held in place by contract, with market practice based on agreement structures that license, restrict and allocate risk around data use. These contracts have grown up over the years and constitute a stable, cohesive normative framework in markets that have seen little litigation. Exchanges and data vendors will seek to apply their standard terms, which are almost universally based on the reservation to the data provider of all IP (copyright, database right in the EU and confidentiality) in the data being supplied and a limited licence to the customer to use the data for specified purposes. Points of contention in exchange, index and data vendor agreements typically centre on:

   - scope of licence and redistribution rights (internal use only or onward supply, and increasingly data use for AI, machine learning ('**ML**') and data science purposes);
   - treatment of data derived from the data initially supplied (who owns it; what the user may do with it);
   - use of the data after termination of the agreement; and
   - scope of compliance audits and remedies for unpermissioned use and over deployment.

8. **Open banking**. In January 2018, two important data-related developments took place in the UK banking industry. First, the UK implemented the second Payment Services Directive ('**PSD2**'), which aims (among other things) to enable banks and other payment account providers, their customers and third parties to share data securely with each other.[13] Second, in a sort of 'own brand' version of PSD2, the UK went live with its own Open Banking initiative, representing an important endorsement of Open Data principles (see paragraph 14 below). This mandates the nine largest UK banks to allow their personal and small business customers to share their account data securely and directly with third party providers regulated by the Financial Conduct Authority ('**FCA**') and enrolled in the Open Banking initiative. The Open Banking Ecosystem

---

[12] See '*Exchange industry revenues reach record levels in 2018*', David Takaba, Burton-Taylor, Inc., 15 July 2019 - https://burton-taylor.com/exchange-industry-revenues-reach-record-levels-in-2018/; '*LSE Refinitiv deal would create the world's largest exchange group and second largest market data supplier*', Burton-Taylor, Inc, 31 July 2019 - https://burton-taylor.com/london-stock-exchange-refinitiv-deal-would-create-the-worlds-largest-exchange-group-and-second-largest-market-data-supplier-burton-taylor-report/

[13] Directive (EU) 2015/2366 of 25 November 2016 on payment services in the internal market (and amending previous directives) - https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32015L2366&from=EN; PSD2 was implemented in the UK by the Payment Services Regulations 2017 (UK SI 2017/752) - http://www.legislation.gov.uk/uksi/2017/752/contents/made

refers to all the components of Open Banking, including the Application Programming Interface ('**API**') standard and the security, processes and procedures, systems and governance to support participants in the initiative. As of September 2019, 143 FCA regulated providers are enrolled in Open Banking.[14]

9. **The insurance sector**. Insurance is based on the insured transferring the risk of a particular loss to the insurer by paying a premium in return for the insurer's commitment to pay if the loss occurs. The combination of big data and AI/ML enables insurance risk to be assessed and predicted much more precisely than in the past by reference to specific data about the insured and the risk insured. In turn, these factors enable the price of the policy to be calculated more accurately.

   As well as the traditional 'top down' statistical and actuarial techniques of risk calibration and pricing, insurers can now rely on data relating to the insured person and insights delivered by AI. For example in vehicle insurance, location based data from the driver's mobile can show where they were at the time of the accident and other telematics data from on-board IT can show how safely they were driving; smart domestic sensors help improve responsiveness to the risk of fire, flooding or theft at home; and health apps and wearables provide information relevant to health and life insurance. Comparing this specific data with insights gleaned from trained AI/ML algorithms enables further accuracy in calibration.

   These examples – data from location based services, vehicle telematics, home sensors and wearables – is having a material impact on vehicle, home and health insurance pricing and terms. Big data and analytics in insurance also point up two other common themes. First, the tension between the privacy of the insured's personal data and its availability to others – a tension that insurers are wrestling with in the context of genetic pre-disposition to illness and the socialisation of risk. Secondly, as in the banking sector, increasing regulatory scrutiny is accentuating the importance of data analytics. For example, the Solvency II directive[15] regulates the amount of capital that an EU insurance company must hold against the risk of insolvency, and this required capital amount is based on likelihood of aggregated policy pay outs where again the predictive insights of big data and AI/ML are critical.

10. **The air transport industry ('ATI')**. The ATI has grown up with computerisation and standardisation as key components in getting passengers (4.3 billion globally in 2018, up by 75% from 2008) and their baggage to the airport of departure, on to the plane, and to and from the airport of arrival. In doing so, airlines and other ATI companies generate and hold vast amounts of data during all stages of the customer journey – for example, the average transatlantic flight generates 1 terabyte of data. But this data can be siloed in a particular application or airline, so big data techniques have emerged to support the service and efficiency improvements that lie at the heart of ATI growth. A recent study quoted in the Financial Times found that big data analytics was a higher priority for the ATI than any other industry[16] as gathering, analysing and using big data enable airports and airlines to develop insights about customers and their air travel preferences and harness competitive advantage.

---

[14] See also '*Open Banking – guidelines for Open Data Participants*', Open Banking limited, July 2018 - https://www.openbanking.org.uk/wp-content/uploads/Guidelines-for-Open-Data-Participants.pdf

[15] Directive 2009/138/EC of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II) (OJ L 335, 17.12.09, p.1) - https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32009L0138&from=en

[16] '*How airlines aim to use big data to boost profits: technology provides carriers with treasure trove of information to optimise customer service*', Financial Times, 8 May 2018 - https://www.ft.com/content/f3a931be-47aa-11e8-8ae9-4b5ddcca99b3

Using big data is also improving ATI efficiencies, as illustrated by Resolution 753 passed by ATI trade association and standards body IATA (the International Air Transport Association) which has required its member airlines since June 2018 to track passengers' baggage from start to finish. Over the decade to 2018, a combination of increasing standards, smart technology, automation and new processes has enabled a 50% reduction in baggage items mishandled (from 47 to 25 million) while passenger numbers have risen by 75%.[17]

11. **The recorded music industry**. The recorded music industry is a $19bn global business in full digital transformation as streaming comes to dominate music consumption. The structure of the industry has grown up around norms based on the individual and collective licensing and management of the various and distinct copyrights that arise in a song's composition, lyrics and publication, and in its recording and performance. These copyright norms operate primarily on a national basis with harmonisation and equivalence established internationally through copyright treaties like the Berne Convention and WIPO Treaties.

    The big three record companies (Universal, Sony and Warner) together account for around 70% of the global recorded music market. The music track is effectively the product unit for the sector and PPL, the UK CMO (Collective Management Organisation) for the public performance rights of its 100,000 recording and performer members, operates a repertoire database of 15 million tracks that is currently growing by 37,000 new recordings per week. Management of data is a large part of PPL's work, driving more accurate distributions and better international collections, where the trend is towards standardising of data submission and exchange formats between country CMOs, their members and licensees.

    The record industry is another sector where data techniques are enabling rapid insights into consumer preferences. These insights have historically been the province of record company A&R (Artist & Repertoire) teams but data is increasingly influencing musical taste, fashion, trends and hence the creation of music itself in a way that has not been possible before. In the words of Geoff Taylor, CEO of UK trade body, BPI:

    > "Increasingly, data, in all its forms – spanning metadata to big data – is playing a key role in shaping this process. As streaming comes to dominate music consumption, data is becoming a progressively more important part of the process of producing and marketing music and, arguably, one of the determinants of a song's popularity."[18]

12. **The healthcare sector**. Healthcare remains the sector where data use will have the greatest impact on people's daily lives. Four drivers lie behind data innovation in UK healthcare: intensifying cost pressures leading to demands for better data; increasing availability of national collections of clinical and treatment outcome datasets; growing investment in anonymising, aggregating and analysing data from individual care centres; and government support of open data and interoperability standards. Public spending on healthcare in the UK (principally the NHS) at around £162bn for the 2020 financial year accounts for roughly 20% of total UK public spending of £848bn. NHS Digital, part of the UK's Department of Health and Social Care, is responsible for the standardising, collecting and publishing of data from across UK health and care systems and in its September 2018 paper '*Data, insights and statistics*', it commented:

    > "Artificial intelligence, machine learning, predictive analytics and the internet of things are no longer dreams of tomorrow. They are here and evolving at pace to support increasingly personalised care. The power of data also has huge potential to drive our economy. The NHS has an unrivalled data set covering

---

[17] SITA, '*2019 Baggage IT Insights*' - https://www.sita.aero/resources/type/surveys-reports/baggage-it-insights-2019

[18] '*Magic Numbers: how data and analytics can really help the music industry – a special insight report by music:)ally for the BPI and the Entertainment Retailers Association*', July 2018, at p.2, https://musically.com/wp-content/uploads/2018/07/MagicNumbersBPI_ERA-1.pdf

a single, large population, stretching back two decades and supported by robust governance and the unique identifying NHS number. It is a major national asset and we have a responsibility to continue to build the trust and infrastructure that will allow the UK and the NHS to be global leaders in this space."[19]

13. **The public sector**. Like all developed states, HMG's database about its citizens is the largest in the country, and government departments like BEIS, Education, Health and Social Care, HMRC, Home Office and Work and Pensions have huge and growing databases. As individual government departments increasingly master their own digital data and central government as a whole starts to move towards data sharing, HMG's data estate is now recognised as a valuable national asset. Looked at as an asset, managing the UK's data estate raises complex policy questions as to protection, growth, maintenance and monetisation, along with the reconciliation of competing interests, including protection of privacy and other individual liberties, the security of the State and its citizens, crime and fraud prevention, commercial interests, safeguards against State overreaching and maximising the benefits of technological progress for citizens.

    The '*Public Sector Data Report 2019*'[20] noted four major trends. First, data and analytics are being widely used in the UK to help address the challenges of public spending cuts; second, security and data breaches top UK public sector concerns; third, half of the survey respondents were either confident and well-trained or confident and eager to learn about working with data and analytics; and fourth data and analytics are seen as high value in the UK public sector. For further information about AI and cloud security in the public sector, please see our white papers on the Legal Aspects of AI and Cloud Security[21].

14. **The policy perspective**. In June 2018, the UK government announced that it would develop a national data strategy. The first step was to set up of a new Centre for Data Ethics and Innovation[22] and in June 2019 an open call for evidence was announced based on three areas of focus – people, economy and government.[23] The European Commission has since 2014 driven a number of policy initiatives designed to build the data economy in the context of its Digital Single Market strategy, including the creation of a common European data space, reviewing the directive on the re-use of public sector information and the recommendation on access to and preservation of scientific information and guidance on data sharing with private sector bodies.[24]

    The elusive nature of data in legal terms has tended to confuse rather than clarify policy debates around data, but the following draws out a number of current themes:

    (a) ***Data: ownership or control?*** A seminar held in October 2018 under the auspices of the British Academy, Royal Society and techUK found that:

---

[19] '*Data, insights and statistics*', NHS Digital, September 2018 - https://digital.nhs.uk/data-and-information/data-insights-and-statistics

[20] https://bigdataldn.com/wp-content/uploads/2019/04/The-Public-Sector-Data-Report-2019.pdf

[21] '*Legal Aspects of Artificial Intelligence*', September 2018, pages 33 to 36. '*Legal Aspects of Cloud Computing: Cloud Security*, June 2018, pages 11 to 17.

[22] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/715760/CDEI_consultation__1_.pdf

[23] https://www.gov.uk/government/publications/national-data-strategy-open-call-for-evidence/national-data-strategy-open-call-for-evidence

[24] https://ec.europa.eu/digital-single-market/en/policies/building-european-data-economy

"use of the term "data ownership" raises significant challenges and may be unsuitable because data is not like property and other goods that can be owned or exchange. Instead discussion should explore the rights and controls individuals, groups and organisations have over data, and should encompass a societal as well as individual point of view. Broader debate could help to better describe the data rights and controls that are often associated with the concept of 'data ownership'."[25]

(b) ***Data: asset or utility?*** Should organisations think of the data they use as an asset with value on the balance sheet or a utility like electricity?

(c) ***Data: asset or liability?*** Increasingly as GDPR and other obligations and duties are perceived as giving rise to significant potential liabilities, organisations are looking at data not only as benefit and an asset but also as a risk and potential liability.

(d) ***Data: proprietary or open?*** From the 1980s, the open source software ('**OSS**') movement rejected the traditional 'cathedral' based approach to software development in favour of an open 'bazaar' approach, and today OSS accounts for a large and growing share of software markets around the world. Similar developments are occurring in data where government and the public sector are increasingly open sourcing publicly held datasets and APIs and making open sourcing of data and research a condition of public funding. This is leading to change in market sectors including open banking (see paragraph 8 above) and scientific and academic research publishing. The new Copyright in the Digital Single Market Directive (see paragraph C.27(a) below) is an example of copyright legislation moving to accommodate this new approach.

(e) ***Data: regulation or market forces?*** Finally, there is a growing groundswell of views in the USA and the EU about whether and if so how to address the perceived influence and power of large data-oriented businesses like Google, Apple, Facebook and Amazon, particularly around whether competition rules should be refashioned or developed to provide greater regulation (see paragraph 31 below).

C. TOWARDS A COMMON LEGAL FRAMEWORK FOR DATA

15. **What is data?** A reasonable start point for a discussion about a legal framework for data is to ask: what is the nature of information and data? For the purposes of this white paper, *information* is that which informs and either is expressed or conveyed as the content of a message or arises through common observation; and *data* is digital information. In the language of the standards world[26]:

"**information** (in information processing) is knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, that within a certain context has a particular meaning; [and] **data** is a reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing [which] can be processed by humans or by automatic means".

Unlike oil, information and data as expression and communication are boundaryless and it would be reasonable to suppose that subjecting information to legal rules about ownership would be incompatible with its nature as without limit. Yet data as digital information is only available because of investment in IT,

---

[25] '*Data ownership, rights and controls: reaching a common understanding*', 3 October 2018, https://royalsociety.org/-/media/policy/projects/data-governance/data-ownership-rights-and-controls-October-2018.pdf

[26] See ISO/IEC (the International Organization for Standardization/the international Electrotechnical Commission) standard 2382-1: 1993(en), Information Technology – Vocabulary. See https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:-1:ed-3:v1:en. Information and data are used interchangeably in this paper.

just as music, books and films require investment in creative effort. To give a bit more colour, it may be helpful to overview some of the types of data we're talking about before exploring the legal aspects.

16. **What types of data are we talking about?** This paragraph very briefly describes the types of data that in-house counsel are likely to be involved with in advising on data projects – AI datasets; big data; derived data; linked data; metadata; real-time, delayed and reference data; and structured and unstructured data.

    (a) ***AI datasets***. AI and ML are a set or stream of technologies not a single one. The main streams are natural language processing, expert systems, vision, speech, planning and robotics. The main ML streams are deep, supervised and unsupervised learning. In each computers learn by example or by being set goals and then teaching themselves to recognise patterns or reach the goal without being explicitly programmed to do so. They do this through using different types of datasets – training datasets to train the AI/ML in the objective to be achieved; test datasets to test the training; and the very large operative datasets used in the production environment.

    (b) ***Big data*.** As used in this paper, 'big data' is shorthand for the aggregation, analysis and value of vast exploitable datasets of structured and unstructured data, although the term has been somewhat eclipsed by the advent at scale of the very large datasets processed by AI/ML. Definitions of big data focus on IT's ability to capture, aggregate, and process an ever-greater volume, velocity, and variety of data. It is characterised by:[27]

    - **aggregation**: of vast volumes of digital data (***size***), in many variable formats (text, image, video, sound, etc.) (***shape***), in unstructured vs unstructured (typically, 80% vs 20%) varieties (***structure***) and arriving at a faster velocity (***speed***);
    - **analysis**: of these aggregated datasets on a ***real-time*** rather than ***batch*** basis, by ***AI/ML*** software and algorithms, enabling a shift from ***retrospective*** to ***predictive*** insight; and
    - **value:** facilitating small but constant, fast and ***incremental business change*** enhancing ***competitiveness***, ***efficiency*** *and* ***innovation*** and the value of the data so used.

    (c) ***Derived data***. (Second or subsequent generation) data that is created or derived from (first generation) data is known as derived data (at its simplest, creating a graph for financial index data). Creating derived data may involve use of the first generation data in ways that infringe the rights of the first generation data owner (like copying, extracting from a database or using confidential information) and so require that person's permission or licence. As data's value rises, so (first generation) data owners become more concerned around the creation, use and ownership of derived data. As a rule of thumb (drawn from the financial market data world) derived data creation is frequently permitted where it can't be reversed back to the first generation data or used as a replacement or substitute for it.

    (d) ***Linked data*** is the method by which structured data can be published, looked up via HTTP, queried and linked to other data by computers. Linked data is to the semantic web what documents are to the world wide web. The www enables a document identified by a URL (uniform resource locator) and containing

[27] See the following from 2013/2014:
- '*Big Data: Seizing Opportunities, Preserving Value*', White House Executive office of the President, 1 May 2014 http://www.whitehouse.gov/issues/technology/big-data-review;
- Liran Einav and Jonathan Levin, '*The Data Revolution and Economic Analysis*,' Working Paper, No. 19035, National Bureau of Economic Research, 2013, http://www.nber.org/papers/w19035;
- '*Towards a thriving data-driven economy*', Communication from the European Commission, 2 July 2014 - https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52014DC0442&from=EN

a standard machine-readable HTML (hypertext markup language) link to be accessed via HTTP (hypertext transfer protocol) from other www locations. The semantic web enables data identified by a URI (uniform resource identifier) and structured in a machine-processable format to be accessed via HTTP, queried and linked to other data by computers.

(e) ***Metadata***. Metadata is 'data about data' – data that provides information about other data. An example of metadata is service or account data derived from the use of a service by a customer but that is not what the customer inputted into or returned from the service. Metadata types here include time and date of creation, data source, file size and data quality. Metadata can be the raw material for AI/ML and data science so, as derived data, its creation, use and ownership are increasingly subject to negotiation.

(f) ***Real-time, delayed and reference data***. In the financial market data world, real-time information is data delivered virtually instantaneously with its creation – a stock price delivered to a service user's terminal when the underlying trade is made, for example. The boundary between (chargeable) real-time and (typically non-chargeable) delayed data varies between different data providers – for example for the London Stock Exchange it is 15 minutes.[28] Reference data is distinct from real-time and delayed data and includes issuer, security and venue identifier codes identifier codes, end of day data and other historical or non-real-time information relating.

(g) ***Structured data***. Data is structured when it is formatted and organised in a pre-defined way so that processing and analysis functions can be applied to its elements. Examples include:

- data in a spreadsheet or database;
- data packets transmitted through the Internet - consisting of a header with the sender's and receiver's IP address, protocol used and packet number; message content as data payload; and trailer showing end of packet and error correction;
- real-time data relating to a securities trade;
- Type B messaging in the ATI for secure message exchange – where the message consists of formal statements relating to message origin and destination (airport, airline code etc) and text (relating to time, flight, route and free text).

(h) ***Unstructured data.*** Unstructured data on the other hand is data that is not organised or defined in a way that is set before the message is sent. In the words of White House's Executive Office of the President (EOP) from a May 2014 report it is:

> "large, diverse, complex, longitudinal, and/or distributed datasets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources". [29]

The ambiguities in and irregularities of unstructured data make it more difficult for traditional programs to process than structured data but much of the data captured or generated by IOT sensors is produced in an unstructured way, and it is estimated that 80% of all big data originates in an unstructured form.

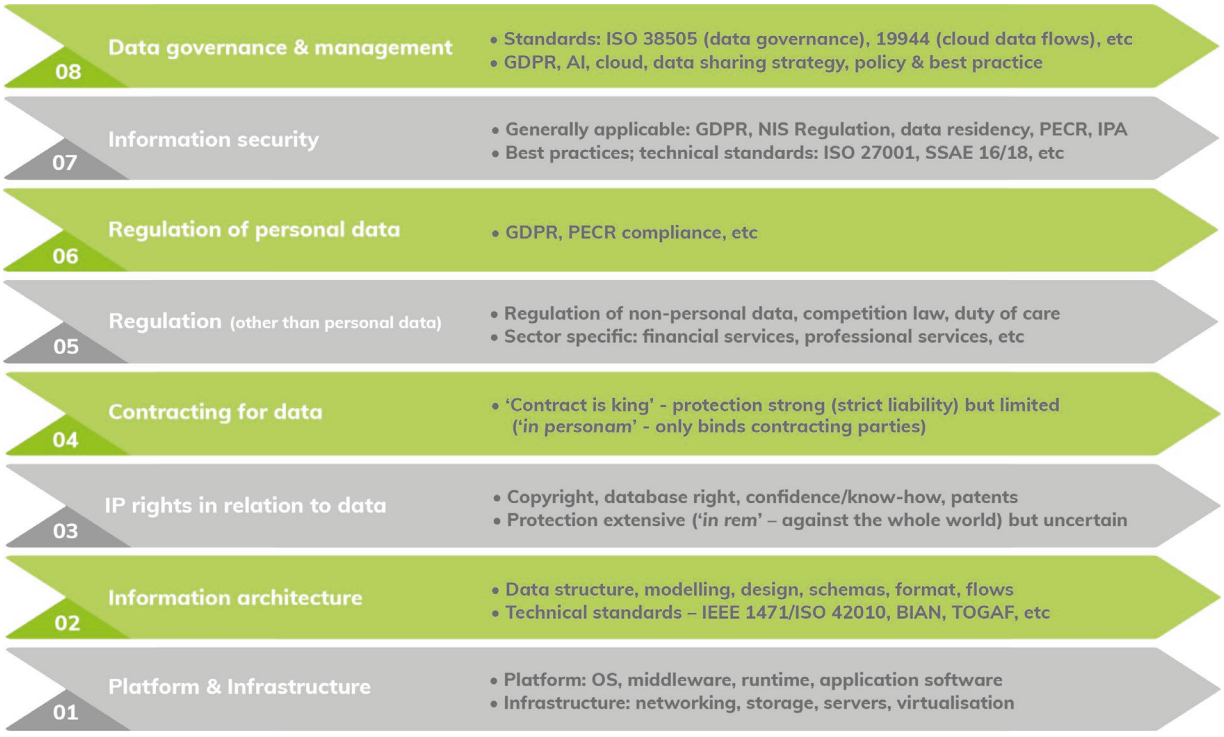
17. **What is data in legal terms?** The equivocal position of data as boundaryless but only available as a result of investment in IT is reflected in the start point for the legal analysis, which is that data is elusive stuff in legal terms. As mentioned in the introduction, this is best explained by saying '*there are no rights **in** data but that*

[28] https://www.londonstockexchange.com/products-and-services/market-data/realtimedata/pricesandpolicies/realtimedataagreement2018.pdf

[29] n 27.

*rights and duties arise **in relation to** data'*. The 1979 UK criminal law case of *Oxford* v *Moss* (see footnote 8 above) is authority that there is no property ***in*** data as it cannot be stolen; and a 2014 UK Court of Appeal ('**CoA**') case confirmed that a lien (a right entitling a person with possession to retain it in certain cases) does not subsist over a database.[30] However, the legal rights and duties that arise ***in relation to*** data are both valuable and potentially onerous and, as an area of law, developing rapidly at the moment.

Figure 1 - Towards a common legal framework for data: the 8-layer stack



These rights and duties arise through IP, contract and regulation. They are important as (positively, in the case of IP and contract) they can increasingly be monetised and (negatively) breach can give rise to extensive damages and other remedies (for IP infringement and breach of contract) and fines and other sanctions (breach of regulatory duty)[31]. Current developments in each of these areas mean that 'data law' is emerging as a new area in its own right around IP, contract and regulation.

18. **Towards a common legal framework for data: the 8 layer stack**. IP, contract and regulation in the context of data can be conceptualised in a legal model as the middle four layers in an eight layer stack, sandwiched between platform infrastructure and information architecture below and information security and data management above (see Figure 1 above, Towards a common legal framework for data: the 8-layer stack).

[30] *Your Response Ltd* v *Datateam Business Media Ltd*, judgment of the Co A on 14 March 2014 [2014] EWCA 281; [2014] WLR(D) 131. See http://www.bailii.org/ew/cases/EWCA/Civ/2014/281.html. A lien in English law is a possessory remedy available only in respect of a 'thing' (or 'chose') in *possession* – as personal tangible property. A database on the other hand is a 'thing' (or 'chose') in *action* – something ultimately capable of enjoyment only through legal action. This case should therefore not be taken as authority for the proposition that that there is no property at all in a database, just that there is no personal tangible property.

[31] For a more detailed review of the technical aspects of data law see Kemp et al, '*Legal Rights in Data*' (27 CLSR [2], pp. 139-151).

D. LEGAL RIGHTS IN DATA: THE 8-LAYER STACK

19. **Level 1: platform infrastructure**. This level consists of the platform's physical infrastructure – data centre (increasingly, the cloud), connectivity, routers, servers, storage and virtualisation - and the software resident on the platform – operating system and middleware. The legal analysis at this level is typically around software copyright issues (rights in computer languages, software 'look and feel', etc.) and the interrelationships between copyright and database right in accessing and extracting the data held in that software.[32]

20. **Level 2: information architecture ('IA')**. The IA is the level between platform infrastructure and the data itself. The IA's ***database schema*** is the formal structure and organisation of the database. It starts with the flow of information in the real world (for example, orders for products placed by a customer on a supplier) and takes it through levels of increasing abstraction, mapping it to a ***data model***. The data model is a representation of that data and its flow categorised as entities, attributes and interrelationships in a way that all information systems conforming to the IA concerned can recognise and process.

    The underlying method and analysis of IA and data modelling apply generally across industry sectors and are central to solving the technical challenges of all projects managing very large datasets. IAs' are consequently becoming increasingly standardised. For example, the International Organisation for Standardisation ('**ISO**') has published (first in 2007 and as updated in 2011) ISO/IEC 42010 on IAs; TOGAF (The Open Group Architecture Framework)[33] operates an open standards based enterprise IA framework; and BIAN (the Banking Industry Architecture Network)[34] operates a banking specific IA standard based on SOA.[35] Advances in the cloud, IOT and AI/ML are leading to further work, including lambda, an IA used to handle very large datasets for real time ('hot' path) processing and batch ('cold' path) processing; and kappa, an alternative to lambda with the same objectives but using a single 'hot' or real-time path for all data flows.

    The IP position of the IA is easily overlooked in practice. Here the documentation describing and specifying the IA will attract traditional literary copyright protection in the normal way; and the database schema (as distinct from the data content of a database) will be protectible by copyright in the EU under Chapter II,

---

[32] See for example *Navitaire Inc* v *Easyjet Airline Company and Bulletproof Technologies, Inc* [2004 EWHC 1725 (Ch) - http://www.bailii.org/ew/cases/EWHC/Ch/2004/1725.html.

[33] See http://www.opengroup.org/subjectareas/enterprise/togaf. TOGAF is also active in other industry sectors.

[34] See https://bian.org/about-bian/. BIAN's financial institution members include many of the large continental European banks and its industry members include many of the large IT suppliers.

[35] Service Oriented Architecture. SOA is a ***software*** development technique ***oriented*** towards associating the business processes or services that the customer requires around the tasks that the developer's software can perform, where the ***architecture*** consists of *application software* that is (i) integrated through a middleware ESB (*Enterprise Service Bus*) messaging framework and (ii) selected, linked and sequenced through *orchestration software*, a metadata menu of available applications. See e.g. http://en.wikipedia.org/wiki/Service-oriented_architecture.

Article 3 of the Database Directive.[36] In the context of a standardised IA the question how the IP in it will be licensed will normally be determined by the IP rights policy applicable to the relevant Standards Setting Organisation ('**SSO**') that manages the standard.

21. **Level 3: IP rights in relation to data - introduction**. The main IP rights in relation to data are copyright (paragraph 22), database right (paragraph 23) and confidentiality (paragraph 24). Patents and rights to inventions can apply to software and business processes that manipulate and process data, but generally not in relation to data itself. Trademarks can apply to data products (like indices) but again, generally not in relation to the actual data.

    IP rights in relation to data are currently of uncertain scope and data IP law will continue to develop in the coming years as data increases in volume, value and value measurability. Historically, IP development has followed the commercialising of innovation, and as the value of data rises, so will the value of the IP rights underpinning it. Case law around database right, database copyright, confidentiality and trade secrets is therefore likely to continue to grow. Whilst of uncertain scope, IP rights are nevertheless extensive as rights '*in rem*' (enforceable against the whole world) with powerful infringement remedies, from temporary and permanent injunctions (court orders requiring termination of the infringement) to damages and account of profits.

22. **Copyright**.

    (a) ***Copyright – general***. Copyright protects the form or expression of information but not the underlying information itself. It applies to software, certain databases, literary, musical, artistic and theatrical works and films, videos and broadcasts. It arises automatically by operation of law in the EU (so does not require to be registered). It is a formal remedy that does what it says on the tin and stops unauthorised copying and the unauthorised carrying out of other acts protected by copyright (best seen as a 'bundle of rights' in this respect).

    (b) ***Ingredients for a successful copyright infringement claim***. A successful claimant for copyright infringement must show:

    - that copyright ***subsists*** in the work – generally, that it is original (where the usual UK standard is low and normally that the work concerned has not been copied from elsewhere) and sufficient to warrant copyright protection (where the English courts have historically taken the pragmatic view that 'what is worth copying is worth protecting');
    - that the claimant can prove ***ownership*** of or can otherwise sue on that copyright;
    - that the work was within copyright ***duration*** (life plus 70 years in the case of software, database copyright and other literary works); and
    - that copyright ***infringement*** has taken place – for example, a qualitatively substantial part of the work had been reproduced without licence or authorisation in circumstances where a copyright permitted act exception did not apply.

[36] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases - http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML

(c) ***Copyright and data***. In the data context, traditional literary copyright subsists in documentation – research publications[37], stock market analysis[38], technical and user documentation and information architecture (see paragraph 20). Software (as instructions to the computer to process data) has benefited from literary work copyright protection in the UK since 1985, and preparatory design material for software since 1993. Moral rights (including the rights to be identified as author and to object to derogatory treatment) apply to literary copyright works but not to software itself.

(d) ***Database copyright***. Database copyright differs from copyright in software and other literary work. This is the result of changes to ss.3 and 3A of the UK Copyright, Designs and Patents Act 1988[39] (‘**CDPA**’) made in 1998 on the introduction into UK law of database right (see para 23). These changes:

- removed the old literary work copyright protection for tables and compilations;
- introduced a new definition of “database”, essentially as a searchable and systematically or methodically arranged collection of independent works, data or other materials; and
- conferred literary work copyright protection on a “database” as so defined, but only where the selection or arrangement of the database’s contents was “the author’s own intellectual creation”, a new and higher originality threshold (borrowed from civil law) than the traditional low UK copyright law threshold of ‘not copied from elsewhere’.

(e) ***Database copyright and the Football Dataco cases***. The new database copyright raised two new questions under UK law: first, what is the relationship between the database and its contents? and second, what was the new “author’s own intellectual creation” originality standard as it applied to content selection or arrangement? These questions were considered between 2010 and 2012 in the context of football fixtures in the *Football Dataco Ltd* v *Brittens Pools Ltd/Yahoo UK Ltd* cases in the UK High Court and CoA and the European Court of Justice (‘**ECJ**’).[40]

Briefly, Football Dataco Ltd (‘**FDL**’) had been appointed by the English and Scottish professional football leagues as their agent to license football fixture lists. FDL brought claims against a number of companies including Brittens Pools and Yahoo! alleging infringement of the leagues’ database copyright and database right. On reference from the CoA, the ECJ held that the policy objective behind the legislation was to stimulate and protect “data storage and processing systems” not to protect the creation of materials capable of being collected in a database. The ECJ held as regards database copyright that only the selection or arrangement of the data *once created* – effectively the structure of the database - and not the creation of the data *in the first place* was to be taken into account when considering originality. This meant that the resources applied by the leagues and FDL were not relevant in assessing whether football fixture lists were eligible for database copyright protection as they were deployed in order to create the data and not to select or arrange them once created.

---

[37] For example *Energy Intelligence Group, Inc.* v *UBS Ltd* (2010)

[38] *Lowry’s Reports, Inc.* v *Legg Mason Inc., et al.* (271 F.Supp.2d 737, Civil No. WDQ-01-3898 (D. Md., July 10, 2003))

[39] http://www.legislation.gov.uk/ukpga/1988/48/contents

[40] Floyd J gave judgment in the UK High Court on 23 April 2010 ([2010] EWHC 841 (ch) - http://www.bailii.org/cgi-bin/markup.cgi?doc=/ew/cases/EWHC/Ch/2010/841.html&query=football+and+dataco&method=boolean). The CoA gave judgement on appeal from Floyd J’s decision on 9 December 2012 ([2010] EWHC 1380 - http://www.bailii.org/ew/cases/EWCA/Civ/2010/1380.html). The ECJ gave judgment on the questions referred to it by the CoA on 1 March 2012 (Case C-604/10 - http://curia.europa.eu/juris/document/document.jsf?text=&docid=119904&pageIndex=0&doclang=en&mode=req&dir=&occ=first&part=1&cid=524892). The CoA finally decided on 20 November 2012.

Turning to the originality threshold, the "author's own intellectual creation" standard in relation to the database structure was met when the author expressed creative ability in an original manner by making free and creative choices – in effect by putting their personal touch on the work. It followed when the case went back to the CoA that the football fixture lists did not benefit from database copyright.

23. **Database right**.

(a) ***Database right – general***. Database right, a separate IP right from copyright, was also introduced into English law in 1998, when the UK implemented the EU Database Directive[41] through the Copyright and Rights in Databases Regulations 1997.[42]

(b) ***Database right – 'made in the EU'***. Database right applies only to databases made in the EU[43]. This has led to qualitatively different legal treatment between databases developed in the EU and in the USA, as in the USA a database made by "sweat of the brow" alone and without "some minimal degree of creativity" will not meet the originality requirement under US law to benefit from copyright protection.[44]

(c) ***Ingredients for a successful database right infringement claim***. Database right arises in a database (which bears the same meaning as under the CDPA – see paragraph 22 above) in whose "obtaining, verifying or presentation" ('**OVP-ing**') the maker has made a "substantial investment". The first owner of database right is generally the maker of the database as the person who takes the initiative in and assumes the risk of OVP-ing its contents. The right lasts for 15 years from initial creation, effectively refreshed wherever "any substantial change" is made. It is infringed by "extraction and/or re-utilization" of a substantial part of the database contents either on a one-off basis or repeatedly and systematically of insubstantial parts.

(d) ***Database right: the Fixtures Marketing and BHB cases***. The first significant cases to consider database right were a series of football fixtures marketing and horse racing cases decided by the ECJ in 2004 of which the *BHB* case[45] is the most important. Here the ECJ considered what was meant in the EU Database Directive by investment in 'obtaining' the contents of a database so as to determine what databases were protectible by database right. The Court espoused the principle that the investment in *creating the materials* that made up the contents of a database was to be disregarded and only the investment in *collecting them in the database* counted:

> "the expression 'investment in … the obtaining … of the contents' of a database in … [the EU Database Directive] must be understood to refer to the resources used to seek out existing independent

---

[41] n 37.

[42] SI 1997/3032 - http://www.legislation.gov.uk/uksi/1997/3032/contents/made

[43] By Article 11(1), database right "shall apply to database whose makers or rightholders are nationals of a Member State or who have their habitual residence in the territory of the Community".

[44] The leading case in the USA is *Feist Publications Inc.* v *Rural Telephone Service Co. Inc,* [499 U.S. 340; 18 USPQ 2d 1275 (1991)] which considered the copyrightabilty of a compilation of names and addresses in a telephone directory. The US Supreme Court held that in order to meet the originality requirement under US copyright law 'sweat of the brow' was not enough to show originality and there needed to be 'some minimal degree of creativity'. The directories in question did not meet this low standard and so were not protected.

[45] Case C-203/02, *The British Horseracing Board Ltd and Others* v *The William Hill Organization Ltd*; Case C-338/02 ECJ Grand Chamber judgment of 9 November 2004. See also Kemp et al, '*Database right after BHB v William Hill: enact in haste and repent at leisure*' (22 CLSR [6], pp 493-498).

materials and collect them in the database. It does not cover the resources used for the creation of materials which make up the contents of a database."

Equally, investment in 'verifying' had to come after the creation of the underlying database materials in order to count for database right purposes. These cases were considered to narrow the scope of database right, especially for real time databases in the financial market data industry for example where the creation of underlying trade data, their collection into a database and their verification may be seen as effectively instantaneous.

(e) ***Database right and the Football Dataco cases***. That the ECJ's principle is not without difficulty was shown in the CoA judgment of 6 February 2013[46] in another case involving FDL, this time where the counterparties were Sportradar GmbH and Stan James plc. Here, the subject of the dispute was FDL's '*Football Live*' service which published live and online factual match information (like goals, scorers, substitutions and red and yellow cards). Defendant Sportradar published a competitive service called '*Sport Live Data*' which it licensed to bookmaker Stan James plc. In compiling '*Sport Live Data*', Sportradar scraped and copied other online sources including '*Football Live*'. Sportradar, following *BHB*, claimed that database right did not arise in the '*Football Live*' database because the investment went into creating the data – recording the facts of the match – not collecting existing materials. Giving the CoA's judgment, Sir Robin Jacob rejected this argument and held that FDL's resources went into collecting the data generated from the football matches, not creating that data, and upheld the Floyd J's first instance judgment[47] that '*Football Live*' was protected by database right. The CoA judgment in *Football Dataco* v *Sportradar* marks a move away from the 'minimalist' stance of the ECJ in *BHB* eight years earlier towards a more nuanced view of the difference between creation and collection of data.

(f) ***Infringement of database right***. The elements of infringement of database right – 'extraction and/or re-utilization' of a substantial part on a one-off basis, or repeatedly and systematically of insubstantial parts – have also been subject to a judicial ebb and flow over the last fifteen years. On the 'minimalist' side, *BHB* is authority that, in the case of a one-off extraction, infringement only occurs if the extraction is substantial, both quantitatively (amount extracted in relation to total database volume) and qualitatively (scale of investment in OVP-ing the part extracted); and that for repeated and systematic extraction to be infringing, the cumulative effect must be that a substantial part of the initial database has been reconstituted.

On the other hand, indirect as well as direct acts can constitute infringing extraction and re-utilisation; exhaustion of rights (the EU term for the first sale doctrine in the USA) does not apply to re-utilisation (*BHB*); and re-utilisation covers any distribution of any part of the database, and can take place in any EU country where the alleged infringer intends to target members of the public (*Sportradar* in the ECJ).[48]

(g) ***Euronext* v *TOM & BinckBank***.[49] In this important July 2015 judgment of the Hague District Court in the Netherlands, the facts were that Euronext, as successor to the Amsterdam Stock Exchange, operated

---

[46] *Football Dataco et al* v *Sportradar GmbH, Stan James plc et al* ([2013] EWCA Civ 27) http://www.bailii.org/ew/cases/EWCA/Civ/2013/27.html

[47] Judgment of 8 May 2012 [2012] EWHC 1185 (Ch) http://www.bailii.org/ew/cases/EWHC/Ch/2012/1185.html

[48] Judgment of ECJ (Third Chamber) of 18 October 2012 - http://curia.europa.eu/juris/document/document.jsf?text=&docid=128651&pageIndex=0&doclang=en&mode=req&dir=&occ=first&part=1&cid=4768956

[49] Case/Registration No. C/09/442420/HA ZQ 13-152. The full text of the judgment (in Dutch) is viewable at http://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2015:8312.

the AEX index of Dutch companies whose shares were traded on its exchange, and a series of options based on the AEX index. TOM, an options trading platform, developed, issued and offered a different options contract by "almost completely copying Euronext's AEX index and options database". The Dutch court gave short shrift to the ***BHB*** line of cases saying (at para 4.36) that the investment in collating a football fixtures list in those cases "required no particular effort" and did not compare with Euronext's investment in its AEX index option series, which included around 50,000 components annually, the accuracy of each of which was critical. Accordingly, in the first judgment that financial market data is protectible by database right, the court found that TOM had infringed the Euronext's database right in its AEX index options database (para 4.38).[50]

(h) ***Other recent cases***. Two cases in the medical field from the UK in August 2017 (an ECG system database)[51] and April 2018 (a UK physiotherapy clinics database)[52] and one from Spain in March 2018 (a pharmaceutical healthmap database)[53] have each held that the database in question was protected by database right, confirming the pendulum swing away from the minimalist position in ***BHB***.

(i) ***EU April 2018 review***. Database right in particular has proved to be a thorny issue for the EU in IP policy terms and it has carried out reviews in 2005 and, in the context of the EU's Digital Single Market initiative, in 2018. The Commission staff evaluation report of 25 April 2018[54] concluded that the right should be retained, although highlighting widely articulated issues around:

- creating/obtaining data, especially in the era of IOT sensor and machine generated databases and the very large datasets processed by AI/ML software;
- interpretation of who is the 'maker' of the database;
- the necessary 'substantial investment' for the right to arise in the first place; and
- the threshold for 'substantial part' of the database infringement purposes.

24. **Confidentiality and trade secrets**.

(a) ***Data and confidentiality***. Copyright and database right each protect the expression and form of information rather than its substance. On the other hand, equitable rules protecting confidentiality of information (where 'equity will intervene to enforce a confidence') may provide a better form of IP protection as they can protect from disclosure the substance of data that is not generally publicly known. Further, a long line of UK[55] cases shows that protection can extend to aggregation of information even

---

[50] In other parts of the case, the defendants were found to have engaged in misleading advertising; BinckBank was found to have breached the derived data terms of the Euronext Market Data Agreement; and TOM was found not to have infringed Euronext's trademarks.

[51] *Technomed Ltd and Another* v *Bluecrest Health Screening Ltd and Another* [2017] EWHC 2142 (Ch) - http://www.bailii.org/ew/cases/EWHC/Ch/2017/2142.html

[52] *Health and Case Management Ltd* v Physiotherapy Network Ltd [2018] EWHC 869 (QB) - https://www.bailii.org/ew/cases/EWHC/QB/2018/869.html

[53] Decision of the Spanish Supreme Court in IMS Health, SA v Infonis, SL, March 2018

[54] Staff working document and executive summary on the evaluation of the Directive 96/9/EC on the legal protection of databases, European Commission, 25 April 2018 - https://ec.europa.eu/digital-single-market/en/news/staff-working-document-and-executive-summary-evaluation-directive-969ec-legal-protection

[55] *Albert (Prince)* v *Strange*, ([1849] 1 M&G 25); *Exchange Telegraph Co. Ltd* v *Gregory & Co.*, ([1896] 1 QB 147); *Exchange Telegraph Co. Ltd v* Central News Ltd ([1897] 2 Ch 48); *Weatherby & Sons* v *International Horse Agency and Exchange Ltd*, ([1910] 2 Ch 297). The last three are the 'wireline' cases.

where parts of it are in the public domain and so not otherwise confidential. In these 'wireline' cases the information concerned was essentially in the public domain but the courts held that the structure of the information in its aggregated form was not and so was protectible as confidential. Protection may also extend to trace through to later generations data derived from the initial confidential data.

(b) ***The EU Trade Secrets Directive***. The EU Trade Secrets Directive[56] brings EU law more closely into line with Article 39 of the WTO TRIPS Agreement[57] (which gives IPR protection to trade secrets as undisclosed information) and the US Uniform Trade Secrets Act[58]. Article 2(1)(a) defines a trade secret as information that (i) is not "as a body or in the precise configuration and assembly of its components generally known or readily accessible", (ii) has commercial value because it is secret and (iii) has been subject to reasonable steps to keep it secret. The directive came into effect in the UK on 9 June 2018 through the Trade Secrets (Enforcement, etc.) Regulations 2018 ('**TSR**').[59] The TSR's Explanatory Note stated that many of the directive's provisions had already "been implemented in the UK by the principles of common law and equity relating to breach of confidence in confidential information, and by statute and court rules". These are the directive's rules on lawful and unlawful acquisition, use and disclosure of trade secrets (Articles 3, 4 and 5) and remedies, process and sanctions (Articles 6, 7 and 16). So the TSR addressed "those areas where gaps occur or where the implementation of the … Directive in the UK, across its jurisdictions, may be made more transparent and coherent". The main substantive change is setting a limitation period of six years for the UK, except in Scotland where it is five (directive Article 8, TSR 4 to 9). Regulations 10 to 19 address various aspects of legal proceedings including legal procedures, remedies and powers of the court and the factors that it is to take into account.

25. **IP rights in relation to data – practical points**. Market participants aiming to maximise their data IP rights should consider the following steps:

- assert (by contract and by website, documentation and other relevant notices) copyright, database right, confidentiality and trade secrets for data existing in, generated by, derived from and transmitted using the systems and services;
- ensure across all website and other notices and contracts that all relevant data is stated to be confidential and trade secret in order to minimise leakage;
- consider the copyright position as a whole, taking into account literary copyright in information architecture and documents associated with the data;
- assert in written methodologies and specifications that the way in which the contents of the database(s) and dataset(s) concerned are selected and arranged is the product of the author's own intellectual creation in order to maximise the likelihood of database copyright availability;
- ensure relevant documentation shows substantial 'OVP-ing' investment in collecting the data in the database as well as creating it so as to maximise the likelihood of database right availability;

---

[56] Directive 2016/943 of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure (OJ L157/2016) - https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016L0943&from=EN

[57] World Trade Organisation Agreement on Trade-Related Aspects of Intellectual Property Rights http://www.wto.org/english/tratop_e/trips_e/t_agm3d_e.htm#7

[58] http://www.uniformlaws.org/Act.aspx?title=Trade+Secrets+Act

[59] SI 2018/597 - http://www.legislation.gov.uk/uksi/2018/597/made

- document the steps taken to keep data secret to maximise the availability of trade secret protection;
- take effective assignments of present and future copyright and database right (and as necessary, trade secrets and confidential information) in all relevant contracts;
- consider whether text and data mining is to be barred and if so prepare a 'written and appropriate' reservation of rights (see paragraph 27(a) below); and
- review the contractual definitions of:
  - confidential information so as to assess what data is included and ensure it covers trade secrets; and
  - IP rights so as to assess whether confidential information and trade secrets are included;

  and ensure consistency of treatment between data as confidential information and data as IP rights.

Contracting for data 04

• 'Contract is king' - protection strong (strict liability) but limited ('in personam' - only binds contracting parties)

26. **Level 4: contracting for data - introduction**. It is the strength of contract law that underpins durable ecosystems like financial market data referred to at paragraph 6 and it is fair to say that 'contract remains king' in the world of data. Contract rights in relation to data are technically entirely separate from IP rights. Their value was confirmed in a UK High Court case in 2006 where the judge said:

> "I agree with [the data supplier] that it is entitled, in principle, to impose a charge for use of its … data by, and for the benefit of, [users], whether or not [it] has IP rights in respect of the data, and, in particular, database rights under the Databases Directive and the Databases Regulations or copyright, and irrespective of the extent of any such rights. [The data supplier] has, in the data, a valuable commodity, for which it is entitled to charge. There is no authority to the contrary, including the [*BHB*] case".[60]

Conversely to IP law, contract confers rights and imposes duties that the law recognises as strong and certain. But whilst data contracts are strong in this way, they operate '*in personam*' – unlike IP rights which operate '*in rem*', they are only enforceable against a party to the agreement and not against the whole world. Data agreements may also impose contractual duties relating to IP rights and the data and materials those IP rights apply to. This means that 'contract' IP (rights and duties imposed by agreement) needs to be analysed under contract law whilst IP rights 'proper' (IP rights and duties arising under and imposed by law) should be analysed separately under the applicable IP rules. This can lead to interpretation challenges, especially where the licence or permission granted under the agreement is effectively the defence to what would otherwise be IP right infringement.

27. **Contracting for data – developing market practice**. This paragraph explores developing contracting market practice around text and data mining, derived data, combined data and metadata and use for 'data science'.

    (a) ***Text and data mining ('TDM')***. The EU 'Copyright in the Digital Single Market' Directive[61] was passed on 17 April 2019 and is due to be implemented by 7 June 2021. It includes terms intended to facilitate TDM by building in 'permitted act' exceptions to copyright and database right. TDM is defined by Art. 2(2) as "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations". Art. 3(1) provides for

---

[60] Etherton J, paragraph 285, *Attheraces Ltd & Another* v *The British Horse Racing Board* [2005] EWHC 3015 (Ch) - http://www.bailii.org/ew/cases/EWHC/Ch/2005/3015.html.

[61] Directive 2019/790 of 17 April 2019 on copyright and related rights in the Digital Single Market, OJ L130/92 - https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L0790&from=EN

a specific 'permitted act' exception where research organisations and cultural heritage institutions carry out TDM for the purposes of scientific research on material protected by copyright and database right that they have lawful access to. This exception may not be excluded by contract (Art. 7(1)). Art. 4(1) provides a more general exception for TDM on lawfully accessible material but by Art. 4(3) this exception may be disapplied where the rightholder has expressly reserved the right "in an appropriate manner such as machine-readable means in the case of content made publicly available online." As a practical matter, care should therefore be taken in drafting any express, 'appropriate' reservation.

(b) ***Derived data***. As mentioned at paragraph 16(c) above, organisations in many different verticals are paying closer attention to the question of derived data – later generation data created from the input data. The background IP law position can be complex so the best outcome is clear express contractual provisions. In copyright law terms, derivative work has a specific meaning under US law (17 USC § 101) but not UK copyright law (where the term used is 'adaptation').[62] Where derived data arises through TDM, the rules in paragraph 27 will apply when implemented. Contractual questions centre on whether the user is permitted to create derived data and if so on what terms and who owns and may use it. The pattern in financial market data is to permit derived data creation and use and for it to be owned by the person creating it so long as the derived can't be reversed back to the input data or used as a replacement or substitute for it. This approach is starting to be followed in other verticals. It is more challenging to apply in the AI/ML context where input data once ingested may not be able to be 'unlearned' by the AI/ML algorithm (especially where the algorithm learns by example or trains itself and its precise operation may be difficult to discover).

(c) ***Combined data***. Combined (or commingled) data is like derived data, but with the user taking input data from more than one source, combining it and creating something different- an analogy is piping yellow and red water into a swimming pool which turns orange. The best solution is for the contract to cover expressly what is to happen, although this may present practical difficulties. In the absence of an express entitlement, the issue for the initial provider wishing to secure an interest in the downstream combined data is that copyright and database right, as formal remedies, may not help, although where the inputs are confidential, confidentiality/trade secrets may help.

(d) ***Metadata and use for 'data science'***. Cloud and data service providers increasingly wish to be able to generate and use metadata. The provider may wish to do this using only data that been anonymised, for purposes of 'data science' (a term increasingly used in practice to mean ethical use of appropriately anonymised data for the purposes of AI/ML, business intelligence/analytics and service improvement) or even without restriction. Again, express contractual wording should be considered to assess what is permitted and prohibited.

28. **Contracting for data – practical points**.

(a) ***Importance of express terms***. A key preliminary practical point for data contracting is to ensure that the agreement expressly addresses all the rights to be granted by the provider and needed by the user and all the restrictions needed by the provider and to be accepted by the user.

(b) ***Data as a licence and as a service***. Although data provision may be expressed in the contract as a licence (that is, permission to do what IP law could otherwise stop), if what is actually being provided is access

[62] See CDPA ss.16(1)(e) and 17. Adaptation is defined at s. 17(3)(a) and for a literary work means translation, in turn defined for software at s.17(4).

to data or supply of data as a feed, then the terms applicable to that service supply should also be expressly addressed.

(c) ***Scope of licence***. Consider:

- exclusive/sole/non-exclusive;
- internal use/onward dissemination/sub-licensing;
- geographical/product restrictions;
- permitted purposes for use of the data:
  - check whether all planned and future use cases are expressly permitted;
  - what is the mechanism for re-purposing/adding new use cases?
  - particularly with social media data, check that the provider's terms permit anticipated uses;

(d) ***Compliance warranties***. (mutual?) warranties of compliance with laws and regulation – data protection, information security, sector specific regulation, audit/investigation;

(e) ***Risk allocation***:

- reliance on data being provided – 'as is', reasonable skill and care or other performance standard?
- supplier and customer indemnity and liability positions;

(f) ***Duration*** of licence (co-terminous with agreement?), duration, suspension and termination of supply;

(g) ***Post-term use of data***. What happens on contract termination where the contract is silent? Can the user /licensee continue to use the data supplied up to termination in the same way as before? Or must it expunge or purge all the data from its systems? Relevant areas of background contract law include rules about contract construction, implication of terms and the applicability of section 3(2)(b) of the Unfair Contract Terms Act 1977 (contract performance different from that which was reasonably expected).

29. **Level 5: non-personal data regulation - introduction**. The third legal area of increasing importance for data is regulation. We have broken this down fairly arbitrarily into data protection (as layer 6 – see paragraph 33) and non-data protection (layer 5), where Regulation 2018/1807, competition law and sector specific regulation are becoming increasingly important. General consumer regulation may also apply to big data but is not considered further here.

30. **Regulation 2018/1807 and non-personal data.** Regulation 2018/1807 came into effect at the end of May 2019 and aims to put in place a framework designed to ensure free movement within the EU of electronic data other than personal data. Without affecting national competent authorities' powers to obtain and access data for their official duties, the Regulation seeks to bar unjustified national data localisation requirements and to enable data porting for non-consumer users' data.[63]

[63] Regulation 2018/1807 of 14 November 2018 on a framework for the free flow of non-personal data within the European Union - https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1807&from=EN

31. **Competition law**. In the EU and the UK competition law operates in three principal areas:

- merger control - at EU level, Regulation 139/2004[64] and in the UK, the Enterprise Act 2002[65] as amended by the Enterprise and Regulatory Reform Act 2013[66];
- rules outlawing abusive behaviour by companies holding a dominant position in a relevant market (Article 102 TFEU and Chapter II UK Competition Act 1998 ("**CA**")); and
- rules outlawing agreements which appreciably restrict competition and affect intra-EU or intra-UK trade (Article 101 TFEU and Chapter I CA).

The competitive impact of mergers is decided in the EU by the Commission and in the UK by the Competition and Market Authority ("**CMA**"). In the case of abuse of dominant position and unlawful anti-competitive agreements, an innocent party who can prove loss will have remedies in the UK including recovery of damages for the tort of breach or statutory duty. It can also complain to the regulator.

Historically, the financial market data industry has been the crucible where data competition law has developed, essentially showing that markets for various types of financial data and the business patterns in them are capable of analysis on traditional competition law lines.[67]

A number of factors have made the competition law analysis of data markets more complex over the last few years, including:

- rapid technological change (fuelled by, and fuelling, big data and AI/ML) and making data a critical competitive input into many downstream products and markets, as well as primary markets;
- the nature of:
  - data as non-rivalrous (data can be used time and again without lessening its value);
  - many services as payment-free (free of direct payment by the user); and
  - consumer behaviour as multihoming (subscribing to or using many competing services together);
- competitive advantage for incumbents perceived as conferred by:
  - significant returns to scale (production costs are non-proportional to customer numbers); and
  - network externalities (higher customer numbers increase service convenience);
- the growth of Google, Apple, Facebook and Amazon as influential data-oriented businesses; and

---

[64] Regulation 139/2005 of 20 January 2004 on the control of concentrations between undertakings (the EC Merger Regulation - https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32004R0139&from=en

[65] http://www.legislation.gov.uk/ukpga/2002/40/contents

[66] http://www.legislation.gov.uk/ukpga/2013/24/contents/enacted

[67] See for example:
- merger control: Reuters/Telerate, Case COMP/M.3695, Art 6(1)(b) Decision of 23 May 2005 - https://ec.europa.eu/competition/mergers/cases/decisions/m3692_20050523_20212_en.pdf; Thomson Corporation/Reuters Group, Case COMP/M.4726, Art 8(2) Decision of 19 February 2008 - https://ec.europa.eu/competition/mergers/cases/decisions/m4726_20080219_20600_en.pdf;
- Article 102: Standard & Poor's – ISINs, Case COMP/39592, Decision of 15 November 2011, –- http://europa.eu/rapid/press-release_IP-11-1354_en.htm?locale=en; Case COMP/39654 – Reuters Instrument Codes, Decision of 20 December 2012 - http://ec.europa.eu/competition/elojade/isef/case_details.cfm?proc_code=1_39654; and
- Article 101: Case AT.39745 – CDS Information Market, Decision of 20 July 2016 - https://ec.europa.eu/competition/antitrust/cases/dec_docs/39745/39745_14238_7.pdf.

- developing policy trends for competition law to 'do more' in regulating data markets.[68]

The EU has taken a robust line where, for example, Google has been found to hold a dominant position in the market for general internet search and fined on three occasions by the EU for abusive conduct.[69]

These cases, together with EU merger cases Microsoft/LinkedIn[70] and Facebook/WhatsApp[71] are leading to the development of new frameworks for competition law analysis of data markets.[72] For example, in '*Is big data a big deal? A competition law approach to big data*', the authors (at pages 201 and 202) propose a 4-step approach:

- do the parties own or control the relevant data?
- is the relevant data commercially available as a product or input for downstream competitors?
- is the relevant data proprietary to the owner's/controller's products and a competitively critical input?
- do reasonably available substitutes for the relevant data exist or is it unique?

Data markets are also under antitrust scrutiny in the USA, with the US Department of Justice reported in June 2019 to be launching investigations into Google and Apple and the Federal Trade Commission reportedly looking into Facebook and Amazon.[73]

32. **Sector specific regulation**. Data regulation is also deepening in many vertical industry sectors. This is not a new thing – the rules on the confidentiality of client information and privilege have been cornerstones of the

---

[68] See for example the following reports from regulators:
- Australia: '*Digital Platforms Enquiry*', July 2019, Australian Competition and Consumer Commission - https://www.accc.gov.au/publications/digital-platforms-inquiry-final-report;
- EU: '*Competition Policy for the Digital Era*', May 2019, DG Competition, European Commission - https://ec.europa.eu/competition/publications/reports/kd0419345enn.pdf;
- UK: '*Unlocking Digital Competition*', March 2019, HM Treasury - https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/785547/unlocking_digital_competition_furman_review_web.pdf; and
- France and Germany: '*Competition Law and Data*', May 2016, Autorité de la Concurrence and Bundeskartellamt - https://www.bundeskartellamt.de/SharedDocs/Publikation/DE/Berichte/Big%20Data%20Papier.pdf?__blob=publicationFile&v=2.

[69] Case AT.39740 - Google Search (Shopping), Decision of 27 June 2017 - https://ec.europa.eu/competition/antitrust/cases/dec_docs/39740/39740_14996_3.pdf, Case AT.40099 – Google Android, Decision of 18 July 2018 - https://ec.europa.eu/competition/antitrust/cases/dec_docs/40099/40099_9993_3.pdf and Case AT.40411, Google AdSense, Commission Press Release of 20 March 2019 - https://europa.eu/rapid/press-release_IP-19-1770_en.htm. See also in the UK, *Streetmap.EU Limited* v *Google* [2016] EWHC 253 (ch), judgment of Roth J of 12 February 2016 - https://www.bailii.org/ew/cases/EWHC/Ch/2016/253.html

[70] Case COMP/M.8124 - Microsoft/LinkedIn, Art 6(1)(b) and 6(2) Decision, 6 December 2016 - https://ec.europa.eu/competition/mergers/cases/decisions/m8124_1349_5.pdf

[71] Case COMP/M.7127 – Facebook/WhatsApp, Art. 6(1)(b) Decision, 3 October 2014 - https://ec.europa.eu/competition/mergers/cases/decisions/m7217_20141003_20310_3962132_EN.pdf

[72] See for example 'Competition Law and Data', Bundeskartellamt and Autorité de la concurrence and, 10 May 2016 - http://www.autoritedelaconcurrence.fr/doc/reportcompetitionlawanddatafinal.pdf, 'Is big data a big deal? A competition law approach to big data', Greg Sivinski, Alex Okuliar and Lars Kjolbye, European Competition Journal, 13:2-3, 199-227 - https://www.tandfonline.com/doi/pdf/10.1080/17441056.2017.1362866?needAccess=true and 'Competition policy for the digital era', European Commission, 5 April 2019 - https://ec.europa.eu/competition/publications/reports/kd0419345enn.pdf and

[73] '*Antitrust and "Big Tech"*', US Congressional Research Service, 11 September 2019 - https://fas.org/sgp/crs/misc/R45910.pdf

legal profession for generations, for example. The explosive growth and digitisation of data are however changing the picture fundamentally in many sectors. These include:

- financial services: where MiFID II has taken MiFID's regulatory template for equities price transparency and extended it to bonds, OTC derivatives and structured finance products, requiring pre- and post-contract price data to be disclosed and reported to the market for trades in all covered securities; and the Benchmark regulation has introduced a regime designed to ensure the accuracy and integrity of indexes and other benchmarks[74];
- insurance: Solvency II;
- the ATI: specific rules on PNR – passenger name record – data about an airline customer's itinerary; and
- healthcare: rules about aggregating anonymised clinical outcome patient data.

The common theme here is sector specific rules applicable to digital data that regulators in the sectors concerned consider significant for carrying out their regulatory functions. These requirements are becoming more intrusive as regulatory authorities obtain wider supervisory powers to obtain information, investigate business practices and conduct and audit organisations under their charge.

33. **Level 6: data protection**.[75]

(a) ***Introduction***. The GDPR, which came into effect on 25 May 2018, applies to personal data used in big data and AI/ML. In her foreword to the ICO's March 2017 paper, '*Big data, artificial intelligence, machine learning and data protection*', the Commissioner said (at page 3):

> "it's clear that the use of big data has implications for privacy, data protection and the associated rights of individuals ... Under the GDPR, stricter rules … apply to the collection and use of personal data. In addition to being transparent, organisations … need to be more accountable for what they do with personal data. This is no different for big data, AI and machine learning." [76]

In addition to the basic principles of GDPR compliance at Arts. 5 and 6 (lawfulness through consent, contract performance, legitimate interests, etc.; fairness and transparency; purpose limitation; data minimization, accuracy; storage limitation; and integrity and confidentiality), big data, AI and ML raise a number of further issues. These include the AI provider's role as data processor or data controller (see (b) below), anonymization (c) and other AI data protection compliance tools (d), research and pseudonymization (e), and profiling/automated decision-making (f). These are now briefly considered.

(b) ***data processor or controller?*** By Art. 4(7) a person who determines "the purposes and means" of processing personal data is a data controller and under the GDPR the data controller bears primary responsibility for the personal data concerned. By Art. 4(8), a data processor just processes personal data on behalf of the controller. Although the data processor does not have direct duties to data subjects

---

[74] Directive 2014/65/EU of 15 May 2014 on markets in financial instruments and amending Directives 2002/92/EC and 2011/61/EU (OJ L 173, 12.6.14, p. 349) ("**MiFID II**"); Regulation (EU) 600/2014 of 15 May 2014 on markets in financial instruments and amending Regulation (EU) 648/2012 (OJ L 173, 12.6.14, p. 84) ("**MiFIR**"); Regulation 2016/1011 of 8 June 2016 (the Benchmarks Regulation).

[75] This para is repeated from paragraph 24 of our Sept 2018 white paper on *Legal Aspects of Artificial Intelligence*'.

[76] https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

for that data, it is required under Arts. 28 to 32 to accept prescriptive terms in its contract with the controller and to take certain other measures. Essentially, an AI provider as a controller has direct duties to the data subjects but as a processor just has direct duties to the controller. Correctly characterising the AI provider as processor or controller is therefore critical to GDPR compliant structuring of the relationship and to allocating risk and responsibility.

However, the borderline between controller and processor can be fuzzy in practice.[77] Where it lies in the AI context was considered for the first time in the UK in the ICO's July 2017 decision on an agreement between the Royal Free Hospital and Google DeepMind.[78] Under the agreement DeepMind used the UK's standard, publicly available acute kidney injury ('**AKI**') algorithm to process personal data of 1.6m patients in order to test the clinical safety of Streams, an AKI application that the hospital was developing. The ICO ruled that the hospital had failed to comply with data protection law and as part of the remediation required by the ICO, the hospital commissioned law firm Linklaters to audit the system. The hospital published the audit report in May 2018[79], which found (at paragraph 20.7) that the agreement had properly characterised DeepMind as a data processor not a controller and observed (at paragraph 20.6) that Streams:

> "does not use complex artificial intelligence or machine learning to determine when a patient is at risk of AKI (which could suggest sufficient discretion over the means of processing to be a data controller). Instead, it uses a simple algorithm mandated by the NHS."

In suggesting that use of "complex" AI or machine learning to determine an outcome could involve "sufficient discretion over the means of processing" to be a controller, the case raises more questions: is algorithm complexity a relevant criterion in assessing who determines the means of processing? If so, where does the border lie? The controller must determine the "purposes and means" of processing, so if the customer determines the purposes (to find out who is at risk of illness, for example) but the AI provider (and not the customer) determines the means of processing (because the AI algorithm is "complex"), is the provider a controller in that case?

(c) ***Anonymization as a compliance tool***. Whilst processing personal data to anonymise it is within the GDPR (because that processing starts with personal data), properly anonymized data is outside the GDPR as it is no longer personal:

---

[77] See further '*ICO GDPR guidance: Contracts and liabilities between controllers and processors' – draft for consultation*', September 2017 - https://ico.org.uk/media/about-the-ico/consultations/2014789/draft-gdpr-contracts-guidance-v1-for-consultation-september-2017.pdf and the EU Article 29 Working Party's, '*Opinion 1/2010 on the concepts of " controller" and "processor"*', 16 February 2010- http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2010/wp169_en.pdf. In the June 2018 *Fashion ID* case, the CJEU held that a website operator featuring the Facebook 'Like' button (a social plugin that causes the transmission to Facebook of website users' personal data) can qualify as a controller jointly with Facebook - https://curia.europa.eu/jcms/upload/docs/application/pdf/2019-07/cp190099en.pdf

[78] '*Royal Free – Google DeepMind trial failed to comply with data protection law*', ICO, 3 July 2017 https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2017/07/royal-free-google-deepmind-trial-failed-to-comply-with-data-protection-law/ See also '*Google DeepMind and healthcare in an age of algorithm*', Powles and Hodson, published version from January 2017 - https://www.repository.cam.ac.uk/bitstream/handle/1810/263693/Powles.pdf?sequence=1&isAllowed=y

[79] '*Audit of the acute kidney injury detection system known as Streams*', Linklaters, 17 May 2018 – http://s3-eu-west-1.amazonaws.com/files.royalfree.nhs.uk/Reporting/Streams_Report.pdf

> "The principles of data protection should … 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" (GDPR Recital 26).

Referring to its *Code of Practice on Anonymisation*[80], the ICO lists anonymisation at one of its six key recommendations for AI:

> "Organisations should carefully consider whether the big data analytics to be undertaken actually requires the processing of personal data. Often, this will not be the case; in such circumstances organisations should use appropriate techniques to anonymise the personal data in the data set(s) before analysis." (paragraph 218.1)

The ICO states that the risk of re-identification is the key criterion here:

> "Organisations using anonymised data need to be able to show they have robustly assessed the risk of re-identification and have adopted solutions proportionate to the risk. This may involve a range of technical measures, such as data masking, pseudonymisation and aggregation, as well as legal and organisational safeguards". (paragraph 135)

(d) ***big data projects: other compliance tools.*** The ICO makes five other recommendations for AI in its '*Big data, artificial intelligence, machine learning and data protection*' paper:

- **privacy notices**: "organisations should be transparent about their processing of personal data … in order to provide meaningful privacy notices" (paragraph 218.2);[81]
- **data protection impact assessments**: "organisations should embed a privacy impact assessment framework into their big data processing activities to help identify privacy risks and assess the necessity and proportionality of a given project" (paragraph 218.3);[82]
- **privacy by design**: "organisations should adopt a privacy by design approach in the development and application of their big data analytics … including implementing technical and organisational measures to address matters including data security, data minimization and data segregation" (paragraph 218.4);[83]
- **ethical principles**: "organisations should develop ethical principles to help reinforce key data protection principles" (paragraph 218.5); and
- **auditable machine learning algorithms**: organisations should implement innovative techniques to develop auditable ML algorithms [including] internal and external audits … to explain the rationale behind algorithmic decisions and check for bias, discrimination and errors" (paragraph 218.6).

(e) ***big data projects: pseudonymization as a further compliance tool in research***. AI/ML and very large datasets will increasingly be used for data and other science research. Personal data processed for scientific research is covered by the GDPR (Recital 159) and Art. 89(1) provides that:

---

[80] '*Anonymisation: managing data protection risk code of practice*', ICO, November 2012 - https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf

[81] See further '*Your privacy notice checklist*', ICO - https://ico.org.uk/media/for-organisations/documents/1625126/privacy-notice-checklist.pdf

[82] See draft '*Data Protection Impact Assessments*', ICO, 22 March 2018 - https://ico.org.uk/media/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/data-protection-impact-assessments-dpias-1-0.pdf

[83] See further '*Data protection by design and default*', ICO - https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-by-design-and-default/

> "Processing for … scientific … purposes … shall be subject to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner."

Art. 4(5) defines pseudonymization (caught by the GDPR by Recital 26) 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".

GDPR Recital 28 provides that:

> "The application of pseudonymisation to personal data can reduce the risks to the data subjects concerned and help controllers and processors to meet their data protection obligations."

Pseudonymisation can therefore help as a GDPR compliance tool for scientific research, which has certain benefits such as:

- processing for science research is considered to be compatible with lawful processing operations (Recital 50 and Art. 6(1)(b));
- the storage limitation principle is somewhat relaxed (Art. 6(1)(e)); and
- the obligations to provide information to data subjects (Recital 52 and Art. 14(5)(b)) and in relation to special categories of data (Recital 65 and Art. 9(2)(j)) are also somewhat wound down.

(f) ***big data projects: profiling and automated decision making***.[84] AI's ability to uncover hidden links in data about individuals and to predict individuals' preferences can bring it within the GDPR's regime for profiling and automated decision making, defined by Art. 4(4) 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."

Art. 22(1) extends data subjects' rights to "decisions based solely on automated processing":

> "the data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her".

The right is qualified not absolute and by Art. 22(2) does not apply if the decision:

---

[84] See also '*Guidelines on automated individual decision-making and Profiling for the purposes of the GDPR*', Article 29 Working Party, 3 October 2017 - http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053; '*Machine learning with Personal Data – Profiling, Decisions and the EU GDPR',* Kamarinou, Millard and Singh, Queen Mary University of London, School of Law Legal Studies Research Paper 247/2016, November 2016 - *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2865811##.* In February 2018, the UK House of Commons Science and Technology Committee launched an inquiry into Algorithms in decision-making - *https://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/news-parliament-2015/algorithms-in-decision-making-inquiry-launch-16-17/*

"(a) is necessary for entering into or performance of, a contract between data subject and data controller;

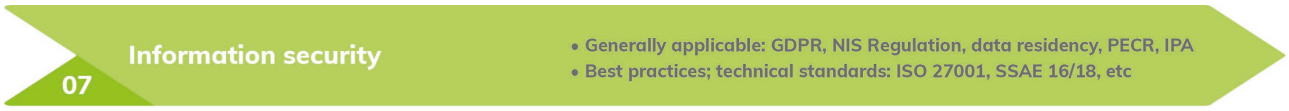
(b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

(c) is based on the data subject's explicit consent."

But by Art 22(3):

"In the cases referred to in points (a) and (c) of [Art.22(2)], the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision."

Art. 22(4) sets out further restrictions relating to processing of special categories of data referred to at Art. 9(1), including racial/ethnic origin, religious beliefs, genetic, biometric or health data or data about an individual's sex life or sexual orientation. The Art. 22 right sits on top of the other rights of data subjects and duties of controllers. Establishing the lawful and fair basis of processing and compliance with the other principles therefore remains important, as does adequately noticing the AI activities concerned. The requirement for decisions to be based 'solely' on automated processing and the safeguarding required by Art. 22(3) are leading AI users to consider interposing human evaluation between the machine and the data subject. This tension between the GDPR and the costs of human intervention is likely to lead to claims about the quality and genuineness of the human decision making.



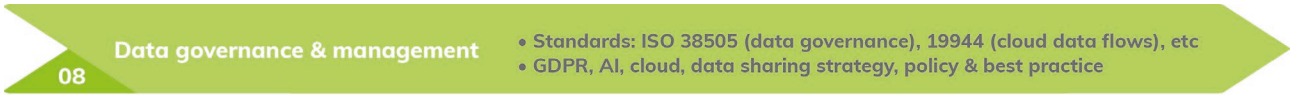
34. **Level 7: Information security**.[85] Towards the top of the data common legal framework sits information security at level 7. The standardisation of data management and security within the organisation has developed significantly over the last few years, and, as with data protection, this is another area where work can potentially be reused when approaching the management of big data. Common standards apply in the payment card industry (**PCI**) whose Security Standards Council (**SSC**) publishes and operates a range of Data Security Standards (**DSS**). More generically, the International Standards Organisation (**ISO**) has published the 27000 series of Information Security Management Systems (**ISMS**) standards and in the USA various audit bodies have published standards on how service companies should report on their information security and other compliance controls (for example SSAE 18 and ISAE 3402).

35. **The legal framework for data: a complex picture**. The legal framework for data presents a complex picture:
    - first, IP (and within IP rights, each of copyright, database right and confidentiality), contract and regulation are discrete sets of norms each with their own technical (and sometimes mutually inconsistent) rules;
    - second, IP rights, contract law and regulation act concurrently on each element of the data stack. A particular dataset – say PNR (passenger name record) data from the ATI – will also potentially be subject to IPR as database right or copyright (in the IT system of an airline); contractual rights and duties (between the airline and a travel agent, say); and data protection regulation (if passenger personal data).

[85] For further information, see our June 2018 white paper on '*Legal Aspects of Cloud Computing: Cloud Security*'.

- third, legal rights and duties arise in a multi-layered way. Data going through several database systems between creation and end use may (in the EU, but not in the USA) be subject to a thin sliver of different database right owned by different actors at each stage as incremental investment is made. A bank subject to regulatory information security and audit duties may seek contractually to impose those requirements on its IT vendors in order to ensure that it is not beholden to its regulator without being able to enforce compliance from suppliers.
- fourth, the computer processes by which data is created – for example financial market data – take place at great speed, so that the evidential burden in formal dispute resolution in showing what happened when is time consuming and costly.
- fifth, IPR rule sets are national rights conferred by national law and enforceable (primarily and initially) in national courts and so operate differently in different countries. Differences vary from the minor (for example, the USA has a generic 'fair dealing' exception to copyright infringement, whereas the UK has a long list of specific 'permitted act' exceptions) to the major (database right is 'made in Europe' and does not apply to databases made in the USA; some countries operate a copyright registration requirement, whilst in others copyright arises by operation of law with no possibility of registration). In the area of regulation, directives in EU law are binding as to the objective to be achieved but leave implementation to each Member State, leading to significant differences in national approach (whereas regulations are directly applicable without the need for national implementation).

These differences in technical rules, the concurrent application of different rules to the same data, the 'multi-layered'-ness of rights in the lifecycle of the data flow, speed of processes and differences between national laws each contribute to the legal complexity of the data rights picture and the legal challenge of big data and AI/ML projects.



E. MANAGEMENT AND GOVERNANCE OF THE ORGANISATION'S DATA OPERATIONS

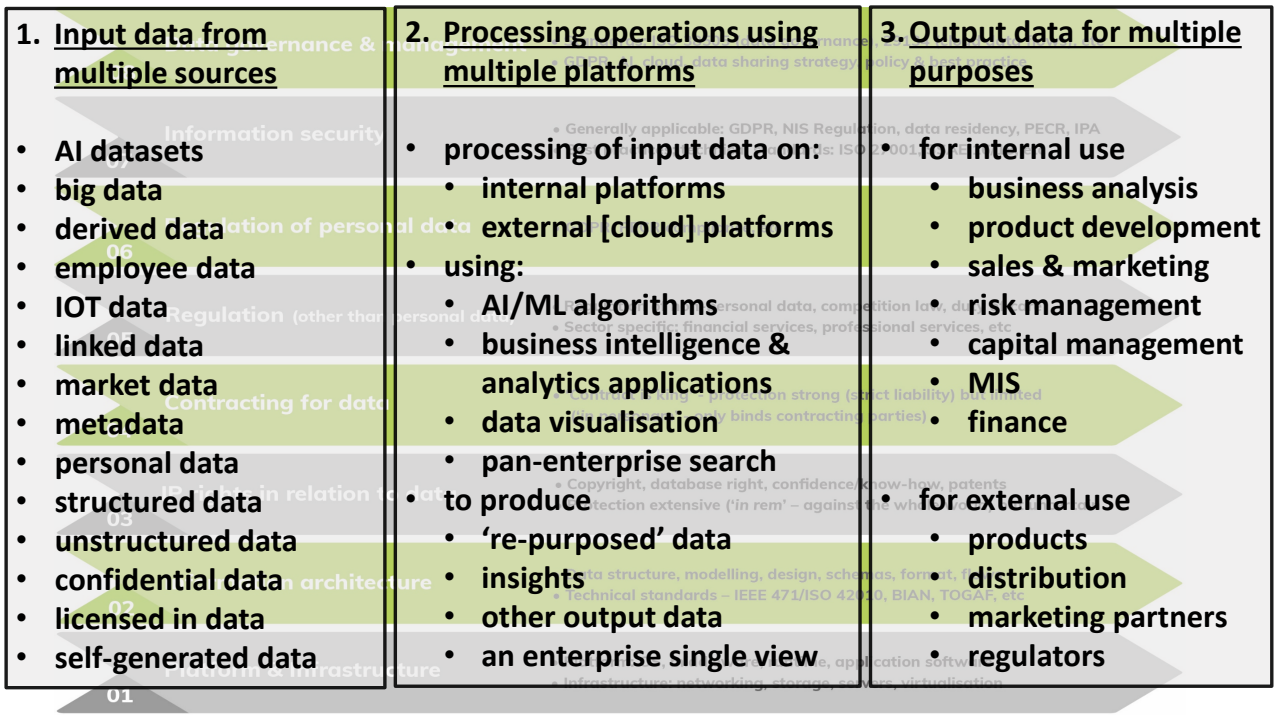
36. **Level 8: data management and governance: introduction**. The 8 layer stack provides this paper's first 'view' - a common legal framework for data. This Section overlays on to that view, first, the organisation's data operations – the input into (paragraph 37), processing within, (paragraph 38) and output from (paragraph 39) the organisation's data 'engine' (see Figure 2 below, 'The data engine – input, processing and output operations'); and secondly a structured approach to managing data projects (See Figure 3 below) based on risk assessment (paragraph 42 below), strategy and policy statements (paragraphs 43 to 46) and processes and procedures (paragraph 49).

37. **Data input operations**. Data comes into the organisation's data engine from an increasingly wide variety of sources. The data can be structured – exchange market data, structured messages or a bought in (licensed) marketing database, for example; it can be confidential or publicly available; it can be personal data or non-personal data; and it can be one or more of these things at the same time. Increasingly, however, it consists of unstructured data like location and other data from mobile and data from home sensors, wearables and other IOT devices and sensors. It is this capturing of 'ever-greater volumes, velocity and variety of data' that, if harnessed effectively, provides the organisation with its big data opportunity.

38. **Data processing operations**. Although data volumes and the power to process them are growing exponentially, there nevertheless remains a gap between the amount of data that organizations can accumulate, and their abilities to leverage that data in a way that is useful. The gap is narrowing as AI/ML

and predictive big data forecasting and modelling techniques catch up with traditional retrospective reporting software. The 'secret sauce' of AI and ML algorithms is assisting organisations in unlocking the 'unspoken secret of big data' – small effects with large aggregated consequences.

Figure 2: The data engine – input, processing and output operations



39. **Data output operations**. The results of the processing then need to go to the places internally within the organisation (the departments and functions where it is of value) and externally (marketing and distribution partners and, increasingly, regulators) where it will be used. Use depends on the industry sector of the company concerned. In insurance for example, vehicle telematics and location based services can inform the insurer of a driver's general skill and care and where he or she was when the accident occurred. This data can be used by underwriters to assess risk and premium costs, by claims assessors to evaluate fault, by the finance department to allocate capital based on risk and hence pay-out profile, by the compliance team for reporting to the regulator, and by product development for new product offerings and for marketing purposes. It is here that the licensing and data protection and other regulatory implications of using data for a different purpose than that for which it was originally obtained become particularly important.

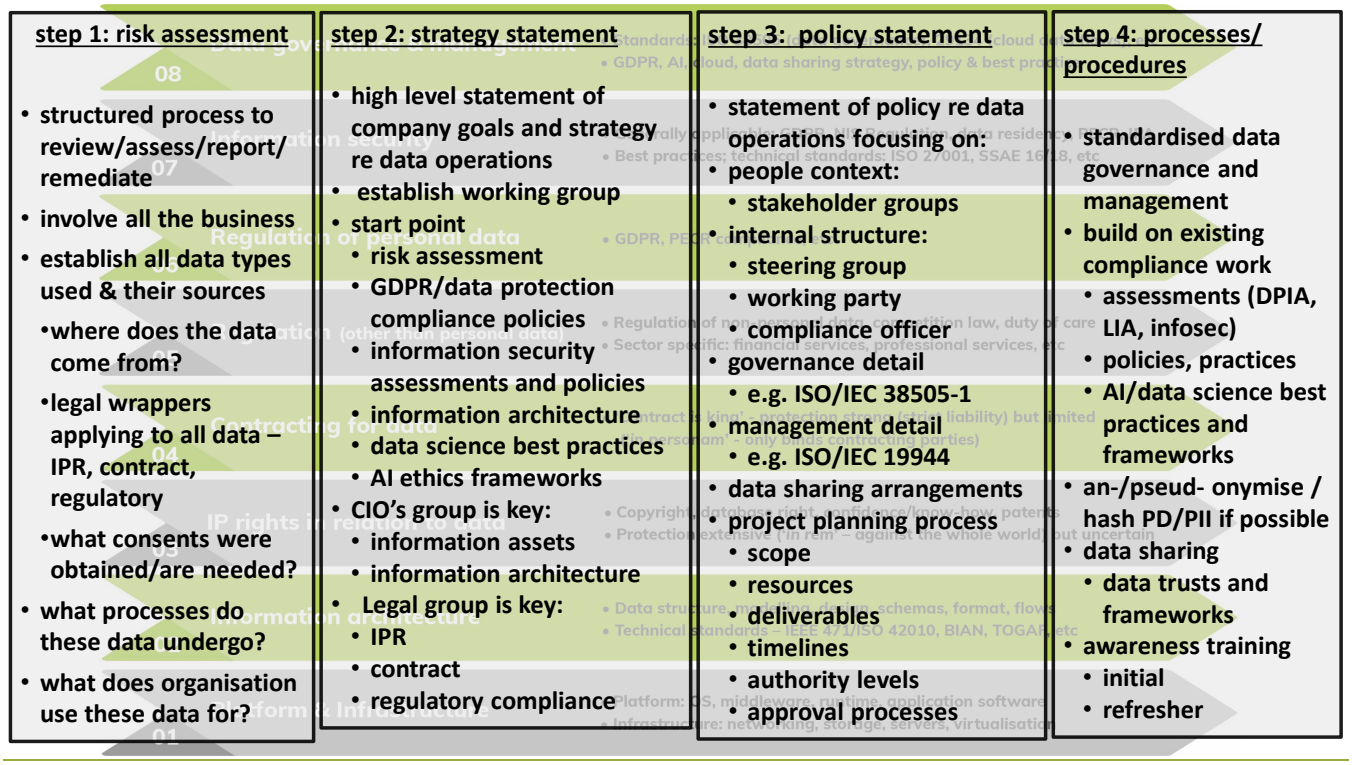
40. **The 'pan-enterprise' view**. This picture conceptualising the data engine is of course over simplified: data input is starting to be but is rarely yet fully coordinated on an enterprise-wide basis: processing operations are likely to be carried out at the desktop as well as at the (on-premise or in-cloud) data centre; and departments may have their own systems and IT requirements. The ways in which an organisation can procure and use big data are also increasing: they may procure data as a service ('**DaaS**') and AI as a Service ('**AIaaS**') from the cloud, rather than make the investment itself, or they may carry out some of these activities in house and some externally.

41. **A structured approach to managing data projects**. The third view of big data – balancing effective and legally compliant use of the organisation's data assets – is superimposed on the first two, the common data legal framework and the big data 'engine'. Here, the objective is a structured approach to managing data projects

with the aim of achieving legally compliant data use across the organisation in a technically enhanced and practical way that allows the business to gain maximum advantage from its data assets.

Data governance does not arise in a vacuum. Large organisations will typically already have in place governance for all or part of their data activities, including data protection and privacy (for example, records of processing activities, impact assessments and data protection by design and by default as required by the GDPR), information security, sector specific data regulatory compliance, information architecture management and AI and data science best practices and ethics frameworks. However, the rise of big data, AI and ML are fuelling a 'democratisation' of the benefits of data utilisation, with operational departments outside the CIO's group looking to use the new capabilities and features. A 'top down' approach to data governance may result in a lack of responsiveness and flexibility, whilst a 'bottom up' approach driven by operational usage may be fragmented and insufficiently address legal, regulatory and business risk in a way consistent with good governance. Practical, incremental management can be built into a structured approach to data governance projects based around four steps – risk assessment, strategy statement, policy statement, and process and procedures - whose key content is shown in Figure 3 below.

Figure 3: Towards a structured approach for managing big data projects



42. **Step 1: risk assessment**. The first step or work stream in a data management and governance project is the risk assessment as to how the organisation is currently using its data, carried out along the normal lines of ***review*** > ***assess*** > ***report*** > ***remediate***. The review will focus particularly on where data is sourced from, the terms under which it is supplied and how it is being used. The next stage will assess whether use is consistent with contractual and licence terms, etc. and whether all consents necessary for the use cases in question have been obtained (including where the data is personal data). The review and assessment will be part of a report to senior management. The review will normally also include recommendations by way of remediation plan to put right any areas of non-compliance that may have been identified in the assessment and also that are forward looking to the strategy and policy aspects of data governance.

43. **Step 2: strategy statement**. The strategy statement is the high level articulation of the organisation's rationale, goals and governance for data, prepared by an inclusive working group or task force consisting of senior management, the legal team, the CIO's team and all other stakeholders. Identification and inclusion of all stakeholders, and articulating the prime objective of each in relation to data and how that objective will be achieved, will be critical to successful data governance and management. The strategy statement for big data will need to align with high level corporate objectives and with other strategy statements in the areas of (i) data protection and privacy, (ii) information security, (iii) information architecture and data methodologies, (iv) data science best practices, (v) AI/ML ethics frameworks and (v) intellectual property management. Organisations will therefore be able to build on work already done in these areas to avoid reinventing the wheel. The role of the CIO's group (looking after IT procurement strategy, the organisation's information architecture and assets and data modelling) and the General Counsel's group (looking after the organisation's IP assets, contracting and regulatory compliance) in formulating the organisation's strategy will be key.

44. **Step 3: policy statement**. Building on and implementing the strategy statement, the policy statement is the next level down and focuses on the people context, internal structure, governance and management detail, approach to data sharing and development of re-usable project planning processes. The working group or task force will be responsible for the third work stream or step of preparing of the data policy statement. As part of its focus on the 'people context' of data governance, the policy statement will generally settle the detail of the institutional framework – for example, steering group, working party or task force, whether there will be a data compliance officer (who may also be the current Data Protection compliance office for example).

    The policy statement will also mandate a project planning process for individual data projects, including setting out scope and dependencies, resources needed, deliverables, timelines (to ensure that projects are to be completed on budget, on time and to standard), authority levels, and approval processes. The working group/task force and policy statement are where the legal considerations around compliant data use across the organisation and the technical considerations around the organisation's information architecture come together. Central to this work is the organisation's overall approach to data management, governance and categorisation.

45. **A standards-based approach to data governance: ISO/IEC 38505-1**. Organisations increasingly look at their data estates recognising that data has value as a business asset but also carries risk and potential liability (for data breach, for example). In order to maximise value and minimise risk, organisations are looking to:

    - establish common processes that apply to their data assets across the data lifecycle;
    - appropriately protect data assets and address any misuse; and
    - enable efficiency gains to be harnessed from a structured, managed, consistent, standardised, repeatable approach that can be applied to all the organisation's data-centred activities, operations and services.

    Technical standards are therefore becoming more widely used as a way of assuring internal stakeholders and external contracting parties that the organisation's data is being managed and governed appropriately. The ISO first published ISO/IEC 38500 on governance of IT for the organisation in 2008, and the current version is from 2015.[86] This was supplemented in 2017 by ISO/IEC 38505-1[87] specifically on the governance

[86] https://www.iso.org/standard/62816.html

[87] https://www.iso.org/standard/56639.html

of data and in 2018 by ISO/IEC 38505-2, [88] which provides guidance on 38505-1. ISO/IEC 38505-1 sets out a framework for data accountability mapping and governance as shown in Fig. 4.

Figure 4: ISO/IEC 38505-1: Data activities in the lifecycle [A] are value, risk and constraint assessed [B] within a comprehensive framework [C] that constantly evaluates, directs and monitors [D]

DATA ACCOUNTABILITY MAPPING AND DATA GOVERNANCE ELEMENTS (ISO/IEC 38505-1)			
A. For each of the following data activities:	**B. assess value, risk & constraints**	**C. within a data governance framework addressing:**	**D. that is constantly evaluate, directed and monitored**
1. Data **Pass Through**	**Value** **Risk** **Constraints**	1. **Responsibility** for data sharing operations	**Evaluate** ICT management systems SWOT analysis **Direct** ICT use ICT management systems **Monitor** ICT management systems performance/conformance
2. Data **Acquisition**		2. **Strategy** rests on its data capabilities	
3. Data **Storage**		3. **Acquisition** assess benefits, costs & risks	
4. Data **Use**		4. **Performance** deploy analytics for continuous improvement	
5. **Derivation** of new data		5. **Conformance** with applicable norms & regulation	
6. Data **Release** (disclosure)		6. **Human Behaviour** taking note of people in the process	
7. Data **Deletion**			

The standard takes:

- an example data management lifecycle (for example with the stages ***pass through*** > ***acquire*** > ***store*** > ***use*** > ***derive*** > ***release*** > ***delete***) [column A in Figure 4];
- at each stage assesses that data's ***value*** and ***risk*** to the organisation and the ***constraints*** (duties or norms) applying to its use [column B];
- within a comprehensive framework based on the 6 principles that the governing body [column C]:
  1. is accountable for the organisation's use of data and will ensure that responsibilities are understood and accepted internally across the organisation and for the data lifecycle (***responsibility***);
  2. is accountable for a data strategy that aligns with the organisation's overall strategy (***strategy***);
  3. is accountable for all data acquisitions and will ensure that they are appropriate (***data acquisition***);
  4. will identify all relevant performance metrics and ensure appropriate action (***performance***);
  5. will ensure the organisation complies with external obligations and internal policies (***conformance***);
  6. will ensure that all data use takes due note of people (***human behaviour***);
- and where the framework constantly:
  - ***evaluates*** current and future use of data;
  - ***directs*** the preparation and implementation of strategies and policies to ensure that data use meets business objectives; and
  - ***monitors*** conformance to policies and performance against strategies [column D].

[88] https://www.iso.org/standard/70911.html

46. **A standards-based approach to data management and categorisation: ISO/IEC 19944**. ISO 38505-1 makes the point that data governance should not be confused with the field of data management, which has "many well-defined methods for the processing of data as well as mechanisms for ensuring the confidentiality, integrity and availability of that data" (ISO 38505-1, p. 7). ISO/IEC 19944[89] looks at the nitty gritty and provides in the context of data management relating to the cloud and personal data a standardised, structured and repeatable approach based on identifying relevant use cases, management practices and common taxonomy. ISO/IEC 19944 takes terms defined in ISO/IEC 29100's privacy framework (like PII[90] Principal, PII Controller, PII Processor and Privacy Stakeholder) and applies them to 'data use statements'. A data use statement is a description or declaration of a particular use case structured according to the rules of ISO/IEC 19944 so that it sets out (as shown in Figure 5) common elements of the use concerned through responses to a set of standard questions:
    1. what is the data (according to the ***data taxonomy*** of ISO/IEC 29100)?
    2. what is the ***data category*** (identified as PII, pseudonymized or anonymized?)
    3. where does the data come from (***source scope***)?
    4. what services are using the data (***use scope***)?
    5. how is the data to be used (***action***)?
    6. where will the data end up (***result scope***)?

Figure 5: overall structure of an ISO/IEC 19944 data use statement (source: ISO/IEC 19944)

The data use statements for the data set, service or product concerned are then identified in this standardised way and applied to data flows where PII may be involved or impacted. For each statement, the standard prescribes (i) a terse statement based on the ISO/IEC 19944 structure intended to be machine- and human- readable; (ii) an indication of intended outcomes; and (iii) a free narrative form.

47. **Data trusts and data trust frameworks: enabling compliant data sharing**.

    (a) ***Introduction***. Data trusts and DTFs are gaining traction as an innovative way to facilitate trusted and regulatorily compliant data sharing. The idea came to public attention in the October 2017 Hall/Pesenti Growing the UK AI industry report which described data trusts as:

    > "a set of relationships underpinned by a repeatable framework, compliant with parties' obligations, to share data in a fair, safe and equitable way". [91]

    Passing the baton to the Open Data Institute[92] ('**ODI**'), the Report's top recommendation was that:

---

[89] ISO/IEC 19944:2017 '*Information Technology – Cloud Computing – Cloud Services and Device: Data flow, data categories and data use*' - https://www.iso.org/standard/66674.html

[90] PII is 'personally identifiable information', the ISO's superset for data that included personal data under the GDPR.

[91] '*Growing the Artificial Intelligence Industry in the UK*' by Prof. Dame Wendy Hall and Jérôme Pesenti, 15 Oct 2017 - https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf

[92] The ODI was founded in 201 by Sir Tim Berners Less and Sir Nigel Shadbolt to advocate open data - https://theodi.org/.

> "Government and industry should deliver a programme to develop Data Trusts – proven and trusted frameworks and agreements – to ensure exchanges are secure and mutually beneficial".

Although incubated in AI, data trusts have broader potential across the whole field of data science and, more generally, in enabling organisations to manage their GDPR personal data sharing duties, non-personal data sharing, data ethics and data governance. The ODI found in research in April 2019:

> "that there is huge demand from private, public and third sector organisations in countries around the world to explore data trusts. Whilst organisations have different ideas about what data trusts could do, they are nevertheless enthusiastic and eager to find ways of sharing data whilst retaining trust, and still deriving benefits for themselves and others."[93]

In July 2019, the ICO endorsed this view in their draft data sharing code of practice consultation:

> "There is a great deal of interest, both in the UK and internationally, in the concept of 'data trusts'. … In essence they are a new model to enable access to data by new technologies (such as artificial intelligence), while protecting other interests and retaining trust, and following a "privacy by design" approach. They have potential for use in data sharing".[94]

(b) ***Towards a definition of data trust***. The ODI in its research on what is meant by 'data trust' found[95] the term interpreted variously as a 'repeatable framework of terms and mechanisms', 'mutual organisation', 'legal structure', 'store of data' and 'public oversight of data access', before deciding[96] in favour of 'a legal structure that provides independent stewardship of data'. In addition to aligning to the ODI's principles for good data infrastructure, the ODI set out six characteristics that a data trust should have:

- a clear purpose;
- a legal structure 'including trustors, trustees with fiduciary duties and beneficiaries';
- rights and duties over stewarded data;
- a defined decision making process;
- a description of how benefits are shared; and
- sustainable funding.

In the commercial world – likely to be where many data trusts will operate - there are two initial issues with the ODI's suggested definition. First, advocating a legal structure implies a separate legal entity, which in turn imposes formalities[97] that may not be necessary in all cases, particularly where a similar result may be more simply obtained through an ecosystem of clearly defined contract terms that each participant expressly accepts. Second, the imposition of fiduciary duties in a commercial context may give rise to unexpected obligations. This is because fiduciary duties are onerous[98] and challenging to

---

[93] '*Huge appetite for data trusts, according to new ODI research*', 15 April 2019 - https://theodi.org/article/huge-appetite-for-data-trusts-according-to-new-odi-research/

[94] '*Data sharing code of practice – draft code for consultation*', ICO, 16 July 2019, page 85 (consultation period to 9 September 2019) - https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-consultation-on-the-draft-data-sharing-code-of-practice/

[95] '*What is a data trust?*', Jack Hardinges, ODI, 15 July 2018 - https://theodi.org/article/what-is-a-data-trust/

[96] *'Defining a data trust*', ODI, 19 October 2018 - https://theodi.org/article/defining-a-data-trust/

[97] These may arise internally within the entity (between the entity and its trustees or directors, for example) and between the entity and third parties (around capacity, contracting, rights, duties and liabilities, etc.)

[98] See for example Millett LJ in the leading case of *Bristol & West Building Society v Mothew (t/a Stapley & Co)* [1998] Ch 1 at 18: "The distinguishing obligation of a fiduciary is the obligation of loyalty. The principal is entitled to the single-minded loyalty of his fiduciary. This core liability has several facets. A fiduciary must act in good faith; he must not make a profit out of his trust; he must not place himself in a position where his duty and his interest may conflict; he may not

calibrate precisely and because the remedies for breach of fiduciary duty may be more extensive than for breach of contract.[99] This is not to downplay data stewards' responsibilities in any way, but to say that clearly expressed contractual rights and duties may will achieve the desired result. They can also be more easily negotiated and risk insured.

The answer may be to accept that 'data trust, the framework' (what the Hall/Pesenti report described a set of contractual relationships underpinned by a repeatable, legally compliant framework) can live alongside 'data trust, the entity' (proposed by the ODI).

(c) ***What does a data trust framework ('DTF') look like?*** The emerging view is to see the DTF as a legal framework together with a set of common operating rules, technical specifications and interfaces (APIs) agreed by and applying for the DTF's specific purposes and between all the participants of the IT ecosystem concerned. Together, the legal and operating rules, specifications and interfaces enable and manage all 'lifecycle' activities for the data concerned (acquisition, flow, storage, use, sharing, consumption and deletion) within the ecosystem.

The DTF is underpinned by a standardised approach to data categorization, data management and data governance, which can be along the lines of ISO/IEC 38505-1 and ISO/IEC 19944 overviewed at paragraphs 45 and 46 above. Combining an approach based on technical standards with design sprints and usability workshops enables the DTF to be constructed quickly, that DTF to be modified for other use cases efficiently, and for different DTFs to work together.

48. **Examples of data trusts and DTFs**. Although DTFs look set to proliferate in the months ahead, currently they are relatively few and far between. Examples include:

(a) ***Silicon Valley Regional Data Trust***.[100] The SVRDT aggregates and uses:

"data from different educational organisations in California and seeks to enable the use of data currently siloed in different organisations for purposes including policy, research and case management."[101]

(b) ***Trūata***. Trūata, which counts MasterCard and IBM as foundational partners:

"enables its clients to derive the maximum value from their data assets while complying with the highest data protection standards. Offering its clients a service to independently anonymise data, enabling them to conduct privacy-enhanced analytics to drive business growth, uphold customer trust and protect brand reputation."[102]

(c) ***SITA BagTrust***. SITA is the leading IT and communications provider to the ATI and, in the context of IATA Resolution 753 described at paragraph 10 above, offers a range of services to airports and airlines that 'track baggage like a parcel'. These include BagTrust:

"a feature for all SITA baggage customers, [that] lets you effectively manage your GDPR policy [and] decide which partners have access to which pieces of data …"[103]

---

act for his own benefit or the benefit of a third person without the informed consent of his principal…." http://www.bailii.org/ew/cases/EWCA/Civ/1996/533.html

[99] Equitable remedies for breach of fiduciary duty include rescission (setting aside), account of profits and other equitable compensation and proprietary remedies (constructive trusts, tracing and recovering tainted proceeds).

[100] https://www.svrdt.org/

[101] Quoted in the ODI document at endnote 9 above.

[102] https://www.truata.com/2019/08/13/take-the-person-out-of-personal-data-uk-consumers-demand/

[103] https://www.sita.aero/solutions-and-services/products/bagjourney

Here, SITA has used its domain expertise to create published rules and, in line with them, entrusted their airline customers with setting preferences around which airport stakeholder is able to see what data. 'Published rules' are preferred to specific contract terms as building greater transparency and trust.

(d) ***UK Government data trust programme***. In January 2019, The UK government announced[104] it was investing £700,000 in a "world-first 'data trust' programme to be piloted in the UK", with three initiatives including **WILDLABS Tech Hub** to tackle illegal wildlife poaching[105] and **WRAP** to address food waste.[106]

49. **Step 4: processes and procedures**. The policy statement will drill down to the level of the fourth step or work stream, the detailed processes and procedures to be used in the organisation's data management. They will likely align to GDPR impact assessments (DPIAs, legitimate interests, compatibility and information security assessments), work on anonymisation, pseudonymisation and hashing, AI principles and ethical frameworks. They are increasingly likely to be built on technical standards such as ISO/IEC 38505-1, 29100 and 19944 and involve data trusts and data trust frameworks. The processes and procedures will also tie into the organisation's HR policies and provide for awareness training.

F. CONCLUSION

50. **Conclusion**. As AI, ML and big data become ubiquitous, gaining unique competitive insight from data has become an indispensable strategic goal of organisations large and small. A sound legal framework for understanding the rights and duties that arise in relation to data in order to manage risk, and the development of a structured approach to the legally compliant management and governance of data operations across the organisation is becoming essential for success in the data-enabled world.

Richard Kemp, Kemp IT Law, London,
October 2019
richard.kemp@kempitlaw.com

[104] See '*Government launches data trust programme - Digital secretary Jeremy Wright sets out plans for scheme using data sharing to tackle global challenges such as illegal wildlife trade and food waste*', Lisa Evenstad, Computer Weekly, 31 January 2019 - https://www.computerweekly.com/news/252456785/Government-launches-data-trust-programme and '*Digital revolution to use the power of data to combat illegal wildlife trade and reduce food waste*', gov.uk, 31 January 2019 - https://www.gov.uk/government/news/digital-revolution-to-use-the-power-of-data-to-combat-illegal-wildlife-trade-and-reduce-food-waste

[105] https://www.wildlabs.net/

[106] http://www.wrap.org.uk/

KEMP IT LAW

IT Law at the Apex

Richard Kemp
Partner

T: 020 3011 1670
M: 07932 695 615
richard.kemp@kempitlaw.com