KEMP IT LAW

Tech Law at the Apex

White Paper

Legal Aspects of Data: Rights and Duties

Richard Kemp

March 2024

LEGAL ASPECTS OF DATA – RIGHTS AND DUTIES: TABLE OF CONTENTS

TABLE OF FIGURES

LEGAL ASPECTS OF DATA – RIGHTS AND DUTIES

A. AUDIENCE, PURPOSE AND SCOPE

1. **Who should read this white paper**. The primary audience of this white paper is in-house legal counsel lawyering their organisations' data estate and data operations.

2. **Purpose**. The purpose of this white paper is to provide a practical guide to legal rights and duties in relation to data – what they are, how they arise and what they mean.

3. **Scope**. **Section C** overviews data across a number of different verticals (financial services, open banking, insurance, air transport, recorded music, healthcare, education and public sector) before briefly considering a number of data policy aspects. **Section D** looks at different types of data before offering a common 8-layer framework for the legal analysis of data. **Sections E** to **J** then work through the first seven levels of the framework: (1) platform infrastructure, (2) information architecture, (3) IP rights in relation to data (4) contracting for data, (5) regulation of personal data, (6) regulation of non-personal data and (7) information security. In this edition of our data white paper we are separating out level 8 (data governance) into a new standalone white paper to be published shortly. In the meantime, the previous version is available on its own as 'Legal Aspects of Data Governance'.

4. **Our white papers**. This is the third edition of our white paper on the legal aspects of data. It is one in an occasional series on aspects of tech law. Others focus on artificial intelligence ('**AI**'), cloud contracting, cloud security, digital transformation, demystifying tech for lawyers and demystifying tech lawyering. All our white papers are available at www.kempitaw.com. This paper is not legal advice. It is written as at 30 November 2023 and from the standpoint of English law.

B. DATA – THE ELUSIVE ASSET

5. **Data – the elusive asset**. Since the first edition of this white paper in 2014, we've all got used to data tropes – 'big data', 'data as the new oil', 'the data economy' – and the recent emergence of generative AI and its snappy utility have given data even greater prominence, which shows every sign of accelerating in the months and years ahead.

   But even with heightened public awareness, data remains elusive in legal terms. Unlike other assets, data is non-rivalrous – it can be used time and again without lessening value – and comparisons with oil fall at the first hurdle: oil is tangible goods under English law[1] and 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.

   Although legally inert in and of itself however, legal rights and duties of increasing scope and complexity apply to and act on data in different ways and use cases: a useful heuristic is '*there are no rights **in** data, but rights and duties 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 currently developing rapidly. These rights and duties – as intellectual property ('**IP**'), contract and regulation – are the main subjects of this white paper.

[1] See, for example, Benjamin, *Sale of Goods*, 11th Edition (Sweet & Maxwell, 2021), paragraph 1-087, pp. 75 & 76.

6. **Growth in data volumes**. As expression and communication, data is infinite and as a resource, data volumes are doubling every two years. Five areas of technology development particularly are currently driving this growth:

   - hyperscale data centres at the cloud's core;
   - proliferating compute capabilities at the edge;
   - ubiquitous mobile phones and devices;
   - connected sensors and the Internet of Things ('**IOT**'); and
   - most recently, generative AI and the vast foundation models underpinning the remarkable developments in this area.

7. **Growth in data regulation**. Also quite remarkable at the present time is the full-on regulatory response that the development of these technologies is bringing about, most notably in the EU's Digital Decade Policy Programme 2030.[2] Here, many major new sets of rules are proposed, in transition or in force centring on or impacting data. This extended toolbox of new rules represents the most complete and vigorous policy response to the demands of data and digitisation yet seen anywhere. Whilst some of these legislative developments do not centre on digital data, we have overviewed them in this white paper, primarily from the standpoint of UK law (and bearing in mind that EU legislation enacted since Brexit does not form part of UK law).

C. THE BUSINESS CONTEXT: DATA IN KEY VERTICALS

8. **Introduction**. As we stand on the threshold of the transformative application of AI across industry and society, this section illustrates how data is already the lifeblood of a wide variety of key verticals, including financial market data (paragraph 9), open banking (paragraph 10), insurance (paragraph 11), air transport (paragraph 12), recorded music (paragraph 13), healthcare (paragraph 14), education (paragraph 15) and public sector (paragraph 16).

9. **Financial market data**. The financial sector is one of the largest users of IT globally. Trading platforms – computer systems to buy and sell securities, derivatives and other financial instruments – are its beating heart. 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, technology developments (including the growing ability of AI to predict and interpret data), and market volatility have in recent years fuelled increasing demand for financial market data (where global spend reached $37.3bn in 2022).[3]

   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

[2] European Commission, 5 January 2023, Digital Decade Policy Programme 2030 | Shaping Europe's digital future (europa.eu).

[3] TPICAP, 18 April 2023, Global spend on financial market data totals a record $37.3 Billion in 2022, rising 4.7% on demand for research, pricing, reference and portfolio management data – New Burton Taylor Report | TP ICAP.

(i) the reservation to the data provider of all IP (copyright, confidentiality, trade secrets and (in the UK and EU) database right) in the data being supplied; and (ii) 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**'), service improvement 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.

10. **Open banking**. In 2018, two important data-related developments took place in the UK banking industry. First, the UK implemented the second EU Payment Services Directive ('**PSD2**'), enabling (among other things) banks and other payment account providers, their customers and third parties to share data securely with each other.[4] 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. This mandated 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 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 spring 2023, there were 336 regulated providers and 7 million users in the Open Banking ecosystem.

    The banking sector is consequently moving towards an increasingly standardised approach to IT around the structure and design of information architecture ('**IA**') in the shared trading, software, online and other information environments that characterise the banking world. For example, two industry standards bodies, the Open Group Architecture Framework (TOGAF) (which operates an open-standards based enterprise infrastructure architecture framework) and the Banking Industry Architecture Framework (BIAN) (which operates a banking specific standard IA based on service oriented architecture) have worked together to facilitate the development of standardised IA and accelerate the transformation that is under way in the sector.

    Central to any IA and so to the collaboration between BIAN and TOGAF is data modelling, the analysis and design of the data in the information systems concerned. The IA's formal structure and organisation of the database:

    - starts with the flow of information in the "real world" (for example, orders for products placed by a customer on a supplier);
    - takes the information through levels of increasing abstraction; and
    - maps the information to a data model (a representation of that data and its flow categorised as entities, attributes and interrelationships). It does this in a way that all information systems conforming to the IA concerned can recognise and process.

[4] 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.

Although this example is taken from the banking world, the underlying method and analysis of IA and data modelling apply generally across industry sectors and are central to solving the technical challenges of big data management projects.

11. **Insurance**. 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 out if the loss occurs. The combination of large datasets, foundation models, and generative and other types of AI enables insurance risk to be assessed and predicted much more precisely than in the past by reference to specific data about the insured, the risk insured and other indicators. In turn, these factors enable the price of the policy to be calculated more accurately and claims assessed more quickly.

    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;
    - in home insurance, smart domestic sensors help improve responsiveness to the risk of fire, flooding or theft at home; and
    - in health and life insurance, health apps and wearables provide relevant information.

    Comparing this specific data with insights gleaned from AI/ML algorithms trained on vast foundation models enables further accuracy in calibration.

    AI and analytics in insurance also point up a number of common themes. First, the tension between the privacy of the insured's personal data and its availability to others – a tension that insurers wrestle with in the context of genetic pre-disposition to illness and the socialisation of risk. Second, bias in AI models may lead to discrimination in outcomes. Third, as in the banking sector, increasing regulatory scrutiny is accentuating the importance of data analytics.

12. **The air transport industry ('ATI')**. The ATI has grown up with computerisation, data and standardisation as key components in getting passengers and their baggage to the airport of departure, on to the plane, and to and from the airport of arrival.

    Airlines manage their inventory (seats) and sell seat tickets to passengers:

    - directly via their own websites;
    - indirectly via 'two step' distribution – through global distribution systems ('**GDSs**') operated by the likes of Amadeus, Sabre and Travelport and then through travel agents; and
    - indirectly via 'one step' distribution – through travel agents, either bypassing the GDS altogether or passing only tangentially through it.

    Airlines pay the GDS a booking fee for their seats booked through that GDS. The GDS also pays commission to travel agents, who frequently engage with a single GDS provider in return for higher commissions. This also means that airlines in practice may need to present their content on all GDSs so as to ensure full coverage.

    As large IT systems, GDSs are central to airline distribution in the way they aggregate data ('content') about airlines' inventory from each airline's passenger service system and present it to travel agents searching for seats for passengers.

13. **Recorded music**.[5] The recorded music industry is a $26bn global business in full digital transformation as streaming has come to dominate music consumption, accounting for over 80% of the industry's revenues. The structure of the industry has grown up around norms based on the individual and collective management and licensing 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 track is the product unit for the sector and PPL, the UK CMO (Collective Management Organisation ('**CMO**')) for the public performance rights of its 150,000 recording and performer members, operates a repertoire database of more than 20 million tracks that is growing by 50,000 new recordings per week. The digital representation of the track is the 'line of rights data', and more than 200 million lines of rights data are provided to PPL each year. Management of data is a large part of PPL's work, driving more accurate distributions, business services to other CMOs and better international collections.

    The record industry is another sector where AI and data techniques are enabling new music content to be generated as well as rapid insights into consumer preferences. These insights have historically been the province of record company A&R (artists and repertoire) teams but data is increasingly influencing musical taste, fashion, trends and therefore the creation of music itself in a way that has not been possible before.

14. **Healthcare**. Healthcare remains the sector where data use will have the greatest impact on people's daily lives. Four drivers lie behind UK healthcare data innovation:

    - 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 AI, open data and interoperability standards.

    Day to day public spending on healthcare in the UK (principally the NHS) is forecast to be around £175bn for the 2023/24 financial year, accounting for roughly 20% of total UK day to day public spending of £930bn.

    Following an independent review,[6] NHS Digital (the national provider of information, data and IT systems for healthcare commissioners, analysts and clinicians) and Health Education England merged with NHS leader NHS England in the first part of 2023. In the words of the review:

    > "Digital technology is transforming every industry including healthcare. Digital and data have been used to redesign services, raising citizen expectations about self-service, personalisation, and convenience, and increasing workforce productivity. The pandemic has accelerated the shift to online and changed patient expectations and clinical willingness to adopt new ways of working. In addition, it facilitated new collaborations both in the centre of the NHS and wider local health and care systems. Together, these changes have enabled previously unimaginable progress in digitally enabled care pathways."

15. **Education**. By the UK Education Act 1996, the Secretary of State is empowered to require every maintained (publicly funded) school to record and keep large amounts of data relating to students, teachers and parents. The data to be kept includes personal details relating to students and staff, and educational records relating

[5] Statistics in this paragraph include those from the IFPI GLOBAL MUSIC REPORT 2023 and PPL Annual Report.

[6] NHS England » Health Education England, NHS Digital and NHS England have merged into a single organisation

to attendance, attainment, school curriculum, exams and educational provision. These record keeping requirements for the UK's 30,000 schools have led to the development of the UK's education technology sector over the last forty or so years, where each school keeps the required data and records in its management information system ('**MIS**'), which connects with complementary software applications that read the data in the MIS to perform additional functions like payment, engagement, accounting and messaging.

16. **The public sector**. As in all developed states, HMG's database about its citizens is the largest in the country, and government departments like Business and Trade, Education, Health and Social Care, Home Office, Revenue and Customs, Science, Innovation and Technology 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 intensifies data sharing and use of AI, 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 overreach and maximising the benefits of technological progress for citizens.

    The UK's National Data Strategy policy paper sets out the broad opportunity for data in the UK public sector:

    > "We are currently in the middle of a fourth industrial revolution. Technological innovation has transformed our lives, changing the way we live, work and play. At the same time, this innovation has brought with it an exponential growth in data: in its generation and use, and in the world's increasing reliance upon it. By embracing data and the benefits its use brings, the UK now faces tangible opportunities to improve our society and grow our economy. If we get this right, data and data-driven technologies like AI could provide a significant boost to the entire economy. Data can improve productivity and provide better-quality jobs. But it can also transform our public services and dramatically improve health outcomes nationally. It can keep us safe and assist the reduction of crime, speed the journey to decarbonisation, and, used well, drive efforts to create a more inclusive, less biased society."[7]

17. **The policy perspective**. It is fair to say that the regulation of data and digital technology is one of the thorniest policy issues impacting the global economy. Faced with a wide range of challenges and opportunities, different countries and trading blocs have adopted varying policy responses.

    The EU has embarked on an energetic, ambitious and comprehensive policy response to the development of digital technologies, most recently in its Digital Decade Policy Programme 2030 published in January 2023, which sets out at paragraph 13 that:

    > "[a] sustainable digital infrastructure for connectivity, microelectronics and the ability to process big data are critical enablers for taking advantage of the benefits of digitalisation, for further technological developments and for digital leadership by the Union."

    We are in a remarkable period that sees many major new sets of rules proposed, in transition or in force (see Figure 1 below). Covering AI, cybersecurity, data, platforms, ePrivacy, healthcare and workforce data and product liability, this extended toolbox of new rules has digital data as its foundation and represents the most complete and vigorous policy response to the demands of technology change yet seen anywhere.

[7] National Data Strategy - GOV.UK (www.gov.uk), 8 December 2020, point 2, the data opportunity

Figure 1 – Table of Proposed, In Transition and In Force EU Digital Technologies Legislation

Area / Measure	Content	Instrument* and status (at 31.10.23)
Artificial Intelligence		
• AI Act	sets out harmonised rules on AI	Regulation, proposal of 21.04.21[8]
Cybersecurity		
• Cyber Resilience Act	addresses connected device software vulnerabilities	Regulation, proposal of 15.09.22[9]
• Critical Entities Resilience (CER) Directive	improve organisational resilience and incident response capacities	Directive EU 2022/2557 in force on 16.01.23, national law transposition date is 17.10.24
• Amendment to Cybersecurity Act	for managed security services	Regulation, proposal of 18.04.23[10]
• NIS 2 Directive	replaces and overhauls and broadens the original directive (EU 2016/1148)	Directive EU 2033/2555 in force on 16.01.23, national law transposition date is 17.10.24
Data		
• Data Act	harmonises rules on data fair access and use	Regulation, proposal of 23.02.22[11]
• Data Governance Act	aims to ensure trust in data sharing, neutrality of data markets and public sector data use	Regulation 2022/868 in force on 24.06.22, most terms apply from 24.09.23[12]
Online platforms		
• Digital Markets Act	aims to foster 'Big Tech' fair competition	Regulation in force on 01.11.22, most terms apply from April 2023[13]
• Digital Services Act	regulates online services and intermediary service providers	Regulation in force on 18.11.22, most terms apply from Feb. 2024[14]
Privacy and data protection		
• ePrivacy Regulation	replaces and overhauls Privacy and eCommunications Directive (2002/58)	Regulation, draft awaits European Parliament reading position[15]

[8] European Commission, *AI Act Proposal* [COM(2021)206 final] (21 April 2021) <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN>.

[9] European Commission, *Cyber Resilience Act Proposal* [COM(2022)454 final] <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52022PC0454&from=EN>.

[10] Proposal for a regulation amending Regulation (EU) 2019/88 as regards managed security services EUR-Lex - 52023PC0208 - EN - EUR-Lex (europa.eu)

[11] European Commission, *Data Act Proposal* [COM(2022)68 final] <1_EN_ACT_part1_v8.docx (europa.eu)>.

[12] Regulation 2022/868 of 30 May 2022, *Data Governance Act* <L_2022152EN.01000101.xml (europa.eu)>.

[13] Regulation 2022/1925 of 14 September 2022, *Digital Markets Act* <L_2022265EN.01000101.xml (europa.eu)>.

[14] Regulation 2022/2065 of 19 October 2022, *Digital Services Act* <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:2022:277:FULL&from=EN&pk_campaign=todays_OJ&pk_source=EURLEX&pk_medium=TW&pk_keyword=Digital%20service%20act&pk_content=Regulation%20>, pp.3-104.

[15] European Commission, *2017 ePrivacy Regulation Proposal* [COM(2017)10 final] <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52017PC0010&from=EN>.

Area / Measure	Content	Instrument* and status (at 31.10.23)
• Health Data Space Regulation	establishes a common space for individuals to manage, and private and public sector entities to access, health, healthcare and genomic data (the EHDSR is the first of several anticipated "domain-specific" common EU data spaces)	Regulation, proposal of 03.05.22[16], draft awaiting European Parliament committee opinion, expected by 2025
• Platform Workers Directive	Includes limits on monitoring of platform workers' psychological state, private conversations and device use outside of platform work	Directive, proposal of 09.12.21[17]
Product liability		
• Liability Directive for AI	adapts non-contractual civil liability rules to AI	Directive, proposal of 28.09.22[18]
• Liability Directive for Products	product liability rules extended to cover digital products	Directive, proposal of 28.09.22[19]

* Under EU law, Directives require transposition under national law whilst Regulations are directly applicable.

In the UK, institutions are still grappling with the consequences of Brexit for tech – examples include the approach of the Intellectual Property Office (IPO) to exhaustion of IP rights and the interplay between AI and IP.

The lack of continuity in the Conservative administrations since Brexit has hampered development of key domestic policy (like data protection reform) and legislation (like the Digital Markets, Competition and Consumers Bill). How HMG will push ahead with regulating big tech remains a key theme; as will be the balance to be struck between the GDPR's reach and HMG's stated aim of reducing the privacy burden on business.

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

- **ownership or control**? Is it more helpful to think data as property or in terms of the control over it that people have?
- **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?
- **asset or liability?** Increasingly as the General Data Protection Regulation ((EU) 2016/679) (**GDPR)**, security and other obligations and duties are perceived as giving rise to significant potential liabilities, organisations are looking at data not only as a benefit and an asset but also as a risk and potential

[16] European Commission, *European Health Data Space Proposal* [COM(2022)197 final] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0197 .

[17] European Commission, Improving Working Conditions in platform work proposal [COM(2021)762 final] < https://ec.europa.eu/social/BlobServlet?docId=24992&langId=en>

[18] European Commission, *AI Liability Directive Proposal* [COM(2022)496 final] <1_1_197605_prop_dir_ai_en.pdf (europa.eu)>.

[19] European Commission, *Product Liability Directive Proposal* [COM(2022)495 final] <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52022PC0495&from=EN>.

liability.

- **proprietary or open?** Traditional proprietary "cathedral" based approach to data licensing or the open source "bazaar" approach?
- **regulation or market forces?** Finally, there is a growing groundswell of views in the UK, the USA and the EU about whether and if so how to address the perceived influence and power of large data-oriented businesses like Amazon, Alphabet, Apple, Meta and Microsoft, particularly around whether competition rules should be refashioned or developed (see below).

D. TOWARDS A COMMON LEGAL FRAMEWORK FOR DATA

18. **What is data?** The start point for the 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[20]:

> "**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."

Information and data as expression and communication are non-rivalrous (i.e. can be used time and again without lessening value) and 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, 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.

19. **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; foundation models, LLMs (large language models) and generative AI; linked data; metadata; open data; Smart Data; real-time, delayed and reference data; and structured and unstructured data.

    19.1 ***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.

    19.2 ***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 eclipsed recently by developments in 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:

[20] 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.

19.2.1 **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***);

19.2.2 **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

19.2.3 **value**: facilitating small but constant, fast and ***incremental business change*** enhancing ***competitiveness***, ***efficiency*** *and* ***innovation*** and the value of the data so used.

19.3 ***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 misusing confidential information) and so require that person's permission. 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.

19.4 ***Foundation models, large language models ('LLMs') and generative AI*** hit the headlines in 2023. Foundation models depend on vast amounts of raw data, the algorithms to harness and train them, and powerful computing capabilities (in the shape of graphics (GPUs) and tensor (TPUs) processing units).

The raw data is typically taken from the internet – for generative AI ChatGPT-3 it was sourced from snapshots of the whole internet between 2016 and 2019 by Common Crawl, a web crawler. It is then cleaned up and trained using 'self-supervision'. This removes the need for the underlying data to be labelled and enables the system to learn by itself, massively speeding things up.

The foundation model then undergoes a number of processes (see Figure 2 below), all happening many times over in parallel to speed things up even further. As the UK Competition and Markets Authority ('**CMA**') AI Foundation Models: Initial Report[21] of 18 September 2023 explains (at page 6):

> "In preparation for the training process, the vast training data sets are broken down into billions of small tokens. In the case of text data, each token may represent a word or parts of a word. During training, the model learns the probabilistic relationships between each token and every other token in the data set they are provided."

In a bit more detail:[22]

19.3.1 the underlying data is converted into numerical tokens;

19.3.2 the token is then given a definition and embedded into a 'meaning space' near other tokens with similar meanings;

19.3.3 the system's 'attention network' next develops associations between the tokens over billions of training runs;

[21] Full report (publishing.service.gov.uk), CMA, 18 September 2023

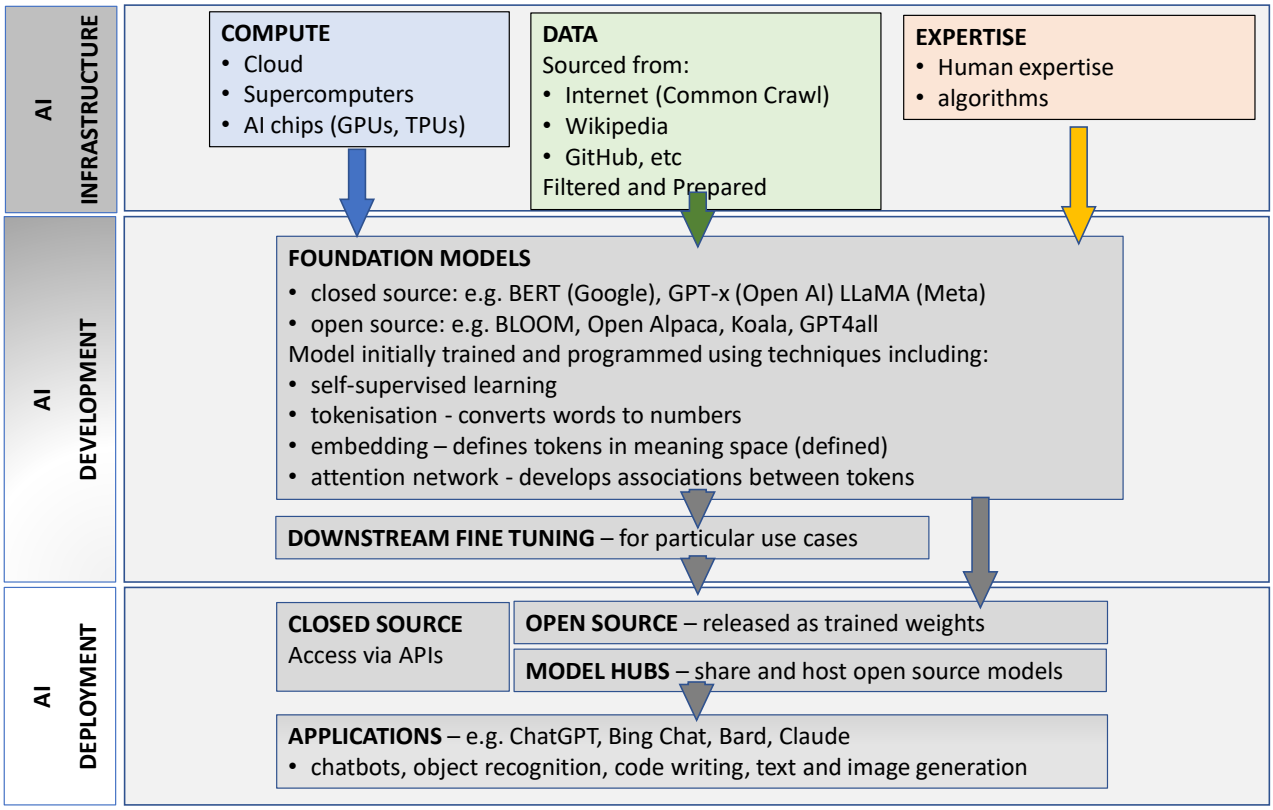
[22] For two good backgrounders see Large, creative AI models will transform lives and labour markets | The Economist, 22 April 2023 and Big Tech is racing to claim its share of the generative AI market | Financial Times, 20 April 2023

19.3.4 the system gradually codes as weights what it sees as numbers and uses these weights to make the token associations closer and more accurate;

19.3.5 the model is fine tuned for particular use cases; and

19.3.6 finally, the application is deployed – for example as chat or generative, code writing or image generation.

Figure 2 – Foundation models[23]



For all the interest in foundation models at the moment, it is worth remembering that they predict, rather than reason: the model does not apply judgement; rather, it is just trained to predict what is most statistically probable. In the words of the Economist:

> "it is much more like an abacus than it is like a mind."[24]

19.4 ***Linked data*** is the method by which structured data can be published, looked up via HTTP (hypertext transfer protocol), 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 a standard machine-readable HTML (hypertext markup language) link to be accessed via HTTP 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.

[23] Source: CMA Initial Review on AI Foundation models, page 8 4 May 2023 (as adapted)

[24] See the Economist article referred to in the previous footnote.

19.5 ***Metadata***. Metadata is 'data about data' – data that provides information about other data. An example 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.

19.6 ***Open data***. See paragraph 34 below.

19.7 ***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. ***Reference data*** is distinct from real-time and delayed data and includes richer content like issuer, security and venue identifier codes, end of day data and other historical or non-real-time information.

19.8 ***Smart Data***. See paragraph 35 below.

19.9 ***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:

19.9.1 data in a spreadsheet or database;

19.9.2 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;

19.9.3 real-time data relating to a securities trade; and

19.9.4 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).

19.10 ***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. It includes large, diverse, complex, longitudinal, and/or distributed datasets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or other digital sources. 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.

20 **What is data in legal terms?** The equivocal position of data as non-rivalrous and 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. This is best explained by saying '*there are no rights **in** data but that rights and duties arise **in relation to** data*'. The 1979 UK criminal law case of *Oxford* v *Moss* 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.[25]

[25] *Your Response Ltd* v *Datateam Business Media Ltd*, judgment of the CA 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 UK law is a possessory remedy available for a 'thing' (or 'chose') in *possession* – as personal tangible property. A database however is a 'thing' (or

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 particularly rapidly at the moment

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).[26] Current developments in each of these areas mean that 'data law' has emerged as a new area in its own right around IP, contract and regulation.

21 **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 3 below.

Figure 3 – Towards a common legal framework for data: the 8-layer stack

'chose') in *action* – capable of enjoyment (or enforcement) only through legal action. The is not authority that that there is no property at all in a database, just that there is no personal tangible property.

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

E. PLATFORM, INFRASTRUCTURE AND INFORMATION ARCHITECTURE

22 **Platform and 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 and interfaces, software 'look and feel', etc.) and the interrelationships between copyright and database right in accessing and extracting the data held in that software.[27]

23 **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:

- **ISO/IEC 42010**: the International Organisation for Standardisation ('**ISO**') has published ISO/IEC 42010:2022 on 'a conceptual model – "meta model" – of the terms and concepts pertaining to Architecture Description';
- **TOGAF** (Open Group Architecture Framework)[28] operates an open standards based enterprise IA framework;
- **BIAN** (Banking Industry Architecture Network)[29] operates a banking specific IA standard based on SOA;[30]
- **Lambda**: 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

---

[27] 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.

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

[29] 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.

[30] 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.

- **kappa** which has the same objectives as lambda but uses 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) may be protectible by copyright in the EU under Chapter II, Article 3 of the Database Directive.[31] 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.

F. INTELLECTUAL PROPERTY RIGHTS IN RELATION TO DATA

24 **IP rights in relation to data – introductory**. The main IP rights in relation to data are copyright (paragraph 25), database right (paragraph 26) and confidentiality and trade secrets (paragraph 27). 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.

25 **Copyright**.

25.1 ***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 certain other acts protected by copyright (best seen as a 'bundle of rights' in this respect).

25.2 ***Ingredients for a successful copyright infringement claim***. A successful claimant for copyright infringement must show:

25.2.1 that copyright ***subsists*** in the work – generally, that it is both *original* and *sufficient* to warrant copyright protection. On *originality*, the historical UK standard was low - normally that the work concerned had 'not been copied from elsewhere' and was the result of more than trivial 'skill and labour'. However, a number of pre-Brexit EU cases have suggested that the correct

[31] 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

test is 'author's own intellectual creation', a higher standard than the 'skill and labour' test implying an element of creativity. The (higher) 'author's own intellectual creation' standard now appears to be correct in the light of the November 2023 UK CoA judgment of Arnold LJ in *THJ Systems* v *Sheridan*.[32] On *sufficiency*, the English courts have historically taken the pragmatic view that 'what is worth copying is worth protecting';

25.2.2 that the claimant can prove ***ownership*** of or can otherwise sue on that copyright;

25.2.3 that the work was within copyright ***duration*** (life plus 70 years in the case of software, database copyright and other literary works); and

25.2.4 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.

25.3 ***Copyright and data***. In the data context, traditional literary copyright subsists in documentation – research publications[33], stock market analysis[34], technical and user documentation and information architecture (see paragraph 23). 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.

25.4 ***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[35] ('**CDPA**') made in 1998 on the introduction into UK law of database right (see para 26). These changes:

25.4.1 removed the old literary work copyright protection for tables and compilations;

25.4.2 introduced a new definition of 'database', essentially as a searchable and systematically or methodically arranged collection of independent works, data or other materials; and

25.4.3 conferred literary work copyright protection on a 'database' as so defined 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'.

25.5 ***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**').[36]

---

[32] [2023] EWCA Civ 1354

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

[34] *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))

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

[36] 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 -

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.

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.

26 **Database right**.

26.1 ***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 through the Copyright and Rights in Databases Regulations 1997.[37]

26.2 ***Database right – 'made in the UK'***. Broadly, until Brexit, database right applied under UK law to databases made in the EEA (including the UK), and database right existing in the UK or EEA before January 2021 continues to apply in the UK. However since the expiry of the Brexit transition period on 31 December 2020, database right under UK law applies only to databases made since then if they were made in the UK, and UK citizens, residents, and businesses are not eligible to receive or hold database right in the EEA for databases created after December 2020.

EU and UK databases receive different legal treatment compared with those created in in the USA, where a database made by 'sweat of the brow' alone and without 'some minimal degree of creativity' will not meet the copyright originality requirement under US law.[38]

26.3 ***Ingredients for a successful database right infringement claim***. Database right arises in a database (which bears the same meaning as under the CDPA) 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

---

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.

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

[38] 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.

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.

26.4 ***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[39] 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 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.

26.5 ***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[40] 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 Floyd J's first instance judgment[41] 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.

26.6 ***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 judicial ebb and flow. On the 'minimalist' side, *BHB* is authority that,

---

[39] 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).

[40] *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

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

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.

26.7 ***Euronext* v *TOM & BinckBank***.[42] 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 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 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.[43]

26.8 ***EU database right review***. Database right was reviewed at EU level 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[44] concluded that the right should be retained, albeit with further questions around (i) who 'makes' a database; (ii) what 'substantial investment' is required; (iii) what is a 'substantial part' for infringement purposes; and (iv) what is meant by 'obtaining' data, especially in the context of databases generated by IOT sensors and machines, and the very large datasets processed by AI/ML software. The last point is proposed to be settled in the draft EU Data Act by a targeted exclusion that database right will no longer apply to databases containing data obtained from or generated by the use of a connected device. It is unlikely that this change will be reflected in the UK if it is enacted in the EU.

27 **Confidentiality and trade secrets**.

27.1 ***Data and confidentiality***. Copyright and database right each protect the expression and form of information rather than its substance. However, 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

[42] 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.

[43] 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.

[44] 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

known. Further, a long line of UK[45] cases shows that protection can extend to aggregation of information even 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.

27.2 ***The EU Trade Secrets Directive***. The EU Trade Secrets Directive[46] brought EU law more closely into line with Article 39 of the WTO TRIPS Agreement[47] (which gives IPR protection to trade secrets as undisclosed information) and the US Uniform Trade Secrets Act.[48]

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**').[49]

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.

27.3 ***Trade secrets and data***. In a legal environment where attaching IP rights to data is challenging, trade secrets have emerged as a likely candidate right, especially in today's more digitally connected, AI- and cloud- enabled world. This is because the area of trade secrecy is relatively structured and harmonised and interoperates benignly with national laws of confidence. However, there remain significant challenges, with a number of key questions to be addressed including:

- how do you show secrecy and prevent erosion in a world of digitally accessible data?
- what constitutes reasonable steps to keep data secret? is the standard similar to the GDPR and NIS Directive standard to take 'appropriate technical and organisational measures'?

---

[45] *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.

[46] 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

[47] 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

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

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

- how do you evidence ownership?
- how do you identify, document and keep records of the trade secret when the algorithm or dataset is dynamic and changing?

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

28.1 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;

28.2 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;

28.3 consider the copyright position as a whole, taking into account literary copyright in information architecture and documents associated with the data;

28.4 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;

28.5 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;

28.6 document the steps taken to keep data secret to maximise the availability of trade secret protection;

28.7 take effective assignments of present and future copyright and database right (and as necessary, trade secrets and confidential information) in all relevant contracts;

28.8 consider whether text and data mining is to be barred and if so prepare a 'written and appropriate' reservation of rights (see paragraph 30.1 below); and

28.9 review the contractual definitions of:

28.9.1 confidential information so as to assess what data is included and ensure it covers trade secrets; and

28.9.2 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.

G. CONTRACTING FOR DATA

29 **Contracting for data – introductory**. It is the strength of contract law that underpins durable ecosystems like financial market data referred to at paragraph 9, 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".[50]

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 by law) should be analysed separately under the applicable IP rules. This can lead to interpretation challenges where the licence or permission granted under the agreement is effectively the defence to what would otherwise be IP infringement; or where the governing law of the agreement is different from the law of the country where the alleged infringement took place.

30. **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'.

30.1 ***Text and data mining ('TDM')***. The EU 'Copyright in the Digital Single Market' Directive[51] 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 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.

The UK did not enact the Directive before Brexit and there has been much discussion about the scope of the TDM exception and at time of writing (end 2023) the question remains open.

30.2 ***Derived data***. As mentioned at paragraph 19.3 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').[52] Where derived data arises through

---

[50] 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.

[51] 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

[52] 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).

TDM, the position should again be considered carefully.

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. As mentioned at paragraph 9, 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 but 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 self-supervises and its precise operation may be difficult to discover).

30.3 ***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.

30.4 ***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.

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

31.1 ***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.

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

31.3 ***Scope of licence***. Consider:

31.3.1 exclusive/sole/non-exclusive;

31.3.2 internal use/onward dissemination/sub-licensing;

31.3.3 geographical/product restrictions;

31.3.4 permitted purposes for use of the data:

31.3.5 check whether all planned and future use cases are expressly permitted;

31.3.6 what is the mechanism for re-purposing/adding new use cases?

31.3.7 particularly with social media data, check that the provider's terms permit anticipated uses;

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

31.5 ***Risk allocation***:

31.5.1 reliance on data being provided – 'as is', reasonable skill and care or other performance standard?

31.5.2 supplier and customer indemnity and liability positions;

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

31.7 ***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).

H. THE REGULATION OF NON-PERSONAL DATA

32. **Non-personal data regulation – introduction**. The third legal area of increasing importance for data in the UK is regulation. This section covers non-personal data and the next section addresses protection of personal data. For non-personal data, the first point to note is that the UK is not currently attempting anywhere near as ambitious a legislative programme as the EU in this area (see Figure 1, paragraph 17 above). Second, despite this, organisations carrying on business in the UK and the EU have to comply with the EU's burgeoning rule book so that, as a practical matter, they may well decide in their UK operations to follow EU law in the UK as well. Third, the post-Brexit vicissitudes of successive conservative administrations have weighed heavily on policy continuity, so that the law making process for a number of key statutes announced has, and continues to be, protracted. Fourth, an incoming labour administration may have different plans.

    We have broken down the area of non-personal data fairly arbitrarily and this section briefly overviews Regulation 2018/1807 (paragraph 33), Open Data (paragraph 34), Smart Data (paragraph 35), general competition law (paragraph 36), the Digital Markets, Competition and Consumer Bill (paragraph 37) and sector specific regulation (paragraph 38). General consumer regulation may also apply to big data but is not considered further here. Finally, this section focuses on regulation of data rather than the related but more general and broader areas of the digital economy, digital technologies and the cloud.

33. **Regulation 2018/1807 and non-personal data.** Regulation 2018/1807, known as the Free Flow Data Regulation, came into effect in the UK 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.[53]

[53] 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

At the end of the transition period, the UK government repealed the Free Flow Data Regulation from UK law. Instead, both the EU and UK have made commitments, in the UK-EU Brexit trade and co-operation agreement, not to implement data localisation requirements, for example by requiring or prohibiting data storage or processing within either territory.

34. **Open Data**. From the 1980s onwards, the open-source software (OSS) movement rejected the traditional proprietary, "cathedral" based approach to software development in favour of an open, "bazaar" approach. Today OSS accounts for a large and growing share of software in use around the world. Broadly speaking, the open data movement aims to do for data what open source has done for software: to make data freely usable without the sort of ownership, licensing and re-use terms that are key to the approach in the proprietary world. 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 and scientific and academic research publishing.

    The UK government in particular has been a powerful influence in driving public sector data towards free and open access, sharing, use and exploitation. Here, the UK RPSI Regulations 2015[54] moved the default position away from limited availability towards requiring UK public sector bodies to allow re-use of their information except in defined circumstances. In practice, government exhortation to use the broad, permissive Open Government Licence[55] (**OGL**) where possible has emerged as a powerful force in favour of open data.

35. **Smart Data and the Data Protection and Digital Information ("DPDI") Bill**. Although the DPDI Bill's changes to the UK's data protection regime overshadow what the Bill proposes for Smart Data, the new data-sharing framework set out in Part 3 will have a significant effect on consumers, incumbents and new players in affected sectors if/when enacted. A number of Smart Data 'schemes' exist in the UK, but only one – Open Banking (see paragraph 10 above) – is operational. In essence, the Bill's Smart Data proposals provide a legislative framework for schemes to be rolled out in sectors of the UK economy and beyond Open Banking.

    Part 3 of the Bill is built around two key concepts: "customer data" and "business data". "Customer data" is data relating to the customer of a trader – for example, data about a given customer's usage patterns or the price paid. "Business data" is wider information about the goods, services or digital content provided by a trader – for example, data about the trader's general pricing, key contract terms or customer feedback data. The aim is that business data gives context to customer data and helps the customer make better decisions.

    The government intends that the main beneficiaries of the Smart Data framework will be consumers and businesses (particularly SMEs). On the other side of the ledger, new rules will have cost implications for the businesses required to comply with new data sharing arrangements. As they stand, the proposals in the Bill are not sector-specific but the Government appears to have certain areas in its sights: non-bank financial services, utilities and telecommunications are all called out in the impact assessment for Part 3 of the Bill.[56]

    The natural point of comparison for the smart data rules is with the EU Data Governance Act (in force in the EU since September 2023) and the draft Data Act and European Health Data Space rules. In comparison with the EU developments, the Smart Data proposals appear modest: sector-specific Smart Data schemes in the UK versus the radical 'horizontal' proposals for data sharing and interoperability in the Data Act. They do, however, seek to build on the proven formula of Open Banking.

---

[54] The Re-Use of Public Sector Information Regulations 2015 (SI 2015 1415)

[55] Open Government Licence (nationalarchives.gov.uk)

[56] See here, paragraph 10 on p. 10 <https://tinyurl.com/2s4krv7j>.

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

- merger control – at EU level, Regulation 139/2004[57] and in the UK, the Enterprise Act 2002[58] as amended by the Enterprise and Regulatory Reform Act 2013[59];
- rules outlawing abusive behaviour by organisations 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 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 of 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.[60]

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) making data a critical competitive input into many downstream products and markets, as well as primary markets;
- the rise of the cloud;
- 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:

---

[57] 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

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

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

[60] 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.

- significant returns to scale (production costs are non-proportional to customer numbers); and
- network externalities (higher customer numbers increase service convenience);

- the growth of Alphabet, Amazon, Google, Meta and Microsoft as influential data-oriented businesses; and
- developing policy trends for competition law to 'do more' in regulating data markets.

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 several occasions by the EU for abusive conduct.[61]

These cases, together with EU merger cases Case M.8124, *Microsoft/LinkedIn*, 6 December 2016 and Case M.7127, *Facebook/WhatsApp*, 3 October 2014 have helped develop several approaches to the competition law analysis of data markets in recent years. For example, in '*Is big data a big deal? A competition law approach to big data*'[62] the authors have proposed 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? and
- do reasonably available substitutes for the relevant data exist or is it unique?

Although the UK is now a third country and no longer part of the EU internal market, under the principle of extraterritorial application of EU competition law, the EU competition rules (Articles 101 and 102 of the TFEU) will continue to apply post-transition period to agreements or conduct of UK companies that have an effect within the EU. UK companies active within the EU, therefore, still need to comply with EU competition law, as well as applicable domestic law. However since the end of 2020, EU competition law is no longer enforced in the UK by the CMA.

Data markets are also under antitrust scrutiny in the USA, with the US Department of Justice in January 2023 launched an important case against Google alleging monopolising multiple digital advertising technology products in violation of US antitrust law.[63]

37. **The Digital Markets, Competition and Consumer ("DMCC") Bill**. On 25 April 2023, the DMCC Bill was introduced into the UK House of Commons. The Bill gives the CMA power to designate undertakings as having "strategic market status" relating to a digital activity and to impose conduct requirements on those undertakings so designated. The CMA would also have power, following investigation and through pro-competition interventions, to intervene to promote competition where it considers that conduct of a designated undertaking has an adverse effect on competition. The Bill also introduces a duty for designated undertakings to report certain mergers and to produce compliance reports and would give the CMA wider investigatory and enforcement powers.

---

[61] See for example: *Case AT.39740, Google Search (Shopping)*, Decision of 27 June 2017; *Case AT.40099, Google Android*, Decision of 18 July 2018; *Case AT.40411, Google AdSense*; and in the UK, see *Streetmap.EU Ltd* v *Google Inc* [2016] EWHC 253 (Ch).

[62] Greg Sivinski, Alex Okuliar and Lars Kjolbye (2017), European Competition Journal 13:2-3)

[63] '*Justice Department Sues Google for Monopolizing Digital Advertising Technologies*', 24 January 2023, Office of Public Affairs, United States Department of Justice

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

38.1 ***Financial services***. The UK MiFID II regime derives from the MiFID II Directive ((EU) 2014/65) and has applied in the UK since the end of the UK-EU Brexit transition period with amendments to ensure the regime remains operationally effective. MiFID II has taken MiFID's regulatory template for equities price transparency and extended it to bonds, OTC derivatives and structured finance products. It requires pre- and post- contract price data to be disclosed and reported to the market for trades in all covered securities. This has led to hefty growth in the market data world.

The UK MiFID II regime applies not just to IT platforms and data but across the whole spectrum of financial instrument trading. In particular, trading operations and procedures that have developed incrementally since the onset of computerised trading in the 1970s have been rewritten to comply with the more prescriptive requirements of these rules. The Benchmark regulation has introduced a regime designed to ensure the accuracy and integrity of indexes and other benchmarks[64].

38.2 ***Insurance***. As in the banking sector, increasing regulatory scrutiny has accentuated the importance of data analytics. For example, the UK Solvency II regime[65] regulates the amount of capital that insurance companies must hold against the risk of insolvency, and this required capital amount is based on likelihood of aggregated policy pay-outs where again AI's predictive insights are critical.

38.3 ***the Air Transport Industry***. For example, specific rules on PNR – passenger name record – data about an airline customer's itinerary; and

38.4 ***Healthcare***. Specific 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.

I. THE REGULATION OF PERSONAL DATA

39. **Data protection regulation**.

39.1 **Introduction**. As an organisation's data will undoubtably include personal data, consideration of data protection legislation will form an important and significant part of an organisation's structured

[64] 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).

[65] Based on EU 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.

approach to the legally compliant management and governance of its data operations. In this context, we focus primary on data management as it relates to AI, machine learning and big data. To the extent that personal data is processed as part of these activities, data protection legislation will need to be considered and one or more territorial data protection regimes may be engaged.

The UK data protection regime comprises the UK GDPR (that is, the retained EU law version of the General Data Protection Regulation ((EU) 2016/679) (UK GDPR)), along with the Data Protection Act 2018 (DPA 2018) and the Privacy and Electronic Communications (EC Directive) Regulations 2003 (SI 2003/2426) (PECR). By Article 3(1), the UK GDPR applies to the processing of personal data, in the context of the activities of an establishment of a controller or a processor in the UK, regardless of whether or not the processing takes place in the UK. Its scope is wider than the UK, as it also applies to the processing of personal data by a controller or processor established outside the UK where the processing activities are related to either (i) the offering of goods and services, irrespective of whether payment is required, to data subjects who are in the UK; or (ii) the monitoring of the behaviour of data subjects in the UK.

The UK GDPR also applies to the processing of personal data by a controller not established in the UK but in a place where domestic law applies by virtue of public international law (*Article 3(3) and recital 2, UK GDPR*). In addition, by Article 3, the EU GDPR continues to have extra-territorial effect and controllers or processors may need to comply the EU GDPR (in addition to the UK GDPR) even if they are not established in the EU, but where the processing activities are related to either (i) the offering of goods and services, irrespective of whether payment is required, to data subjects in the EU; or (ii) the monitoring of the behaviour of data subjects in the EU. Organisations may, therefore, find themselves subject to parallel data protection regulatory regimes under the UK GDPR and the EU GDPR.

39.2 The ICO, in its most recent Guidance on AI and data protection of 15 March 2023 (the "**March 2023 guidance**") said:

> "We see new uses of artificial intelligence (AI) everyday, from healthcare to recruitment, to commerce and beyond. We understand the benefits that AI can bring to organisations and individuals, but there are risks too. We have set out some of these risks, such as AI-driven discrimination in ICO25, our strategic plan for the next two years. Enabling good practice in AI has been one of our regulatory priorities for some time, and we developed this guidance on AI and data protection to help organisations comply with their data protection obligations. The guidance:
>
> - gives us a clear methodology to audit AI applications and ensure they process personal data fairly, lawfully and transparently;
> - ensures that the necessary measures are in place to assess and manage risks to rights and freedoms that arise from AI; and
> - supports the work of our investigation and assurance teams when assessing the compliance of organisations using AI." (*Page 5*)

Key requirements of the UK GDPR in relation to AI and the datasets it uses are summarised below,

39.3 **The March 2023 guidance**. The March 2023 guidance (which replaces the earlier 2014 and 2017 versions) looks at AI and data through the lens of the basic principles of UK GDPR compliance and contains a helpful AI and data protection toolkit.

39.4 **March 2023 guidance – basic principles**. The basic UK GDPR principles are dealt with chapter by

chapter (new, old or updated) in the March 2023 guidance around:

- the accountability and governance implications of AI (*updated* with a *new* section 'on things to consider as part of your data protection impact assessment' ("**DPIA**");
- transparency in AI (*new*);
- lawfulness in AI (*updated* with *new* content added on inferences, affinity groups and special category data);
- accuracy in AI (*new* chapter with *old* and *new* content);
- fairness, bias, discrimination and assessing the impact of Article 22 GDPR (automated decision making) (*new* chapter with *new* and *old* content)
- data minimisation (*updated*); and
- ensuring individual rights in AI systems (*updated*).

39.5 **March 2023 guidance – toolkit**. The toolkit is in in the form of a customisable Excel spreadsheet framework which for each of four stages of the AI life cycle:

(i) business requirements and design ("**BRD**");

(ii) data acquisition and preparation;

(iii) training and testing; and

(iv) deployment and monitoring;

sets out:

(a) the UK GDPR reference and a statement of data protection risks (e.g. for BRD, accountability purpose limitation, fairness, transparency, security, data minimisation, storage limitation, individual rights and meaningful human review);

(b) two blank columns for completion headed risk assessment summary and inherent risk rating;

(c) control steps to be taken for each risk area, and the objective of that control (e.g. conducting a data protection impact assessment (DPIA) and making it clear who in the organisation is accountable for AI risk management);

(d) practical steps for to reduce risk (e.g. DPIA, operational procedures and processes and data flow mapping);

(e) specific information about the steps to be taken, who owns the control steps, current status and completion date; and

(f) finally, a statement of residual risk once the action and steps above have been taken.

J. INFORMATION SECURITY

40. **Information security**.[66] Towards the top of the data common legal framework sits information security at level 7.

40.1 ***The NIS Directive***. The EU-originating Cybersecurity Directive ((EU) 2016/1148) (also known as the Network and Information Security Directive or NIS Directive) imposes minimum cybersecurity and incident reporting obligations on certain operators of essential services ("**OES**") in key industry sectors including energy, transport, health, drinking water supply and distribution, and digital infrastructure, and relevant digital service providers ("**RDSPs**") (such as online marketplaces, search engines and cloud services) who meet certain qualifying criteria.

40.2 ***The UK NIS Regulations***. The UK government implemented the Cybersecurity Directive into UK law through the Network and Information Systems Regulations 2018 (SI 2018/506)[67] (the "**NIS Regulations**") which came into force on 10 May 2018, with significant fines (of up to £17 million) for organisations that fail to meet these requirements.

Although the NIS Regulations focus on the security of IT systems, and the continuity of essential services, rather than data, they do, indirectly, contribute to effective data regulation in the relevant key sectors they cover, as the security of data cannot often be separated from the IT systems which hold them.

40.3 ***EU developments***. The EU is in the process of implementing an inter-connected suite of EU network and information security directives and regulations, including:

- the proposed Cyber Resilience Act: containing obligations for connected device software vulnerabilities)[68];
- an amendment to the Cyber Security Act for managed service providers;[69]
- the NIS 2 Directive: upgrading rules for network and information systems – and replacing the NIS 1 Directive, the Directive originally implemented in the UK by NIS Regulations[70]; and
- the Critical Entities Resilience Directive: aiming to reduce vulnerabilities of critical entities[71].

40.4 ***UK developments***. Following a consultation in 2022, the UK government announced in December 2022 that intended to update the NIS Regulations. The changes consulted on include:

---

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

[67] The Network and Information Systems Regulations 2018 (legislation.gov.uk)

[68] Regulation, proposal of 15 September 2022

[69] Regulation of 17 April 2019, proposal to amend of 18 April 2023

[70] Directive of 14 December 2022

[71] Directive of 14 December 2022

- extending the regulations' scope by bringing managed service providers (**MSPs**) within the net in order to keep digital supply chains secure;
- upgrading the reporting of cyber incidents to regulators;
- establishing a cost recovery system for enforcement;
- enabling the government to amend the NIS Regulations; and
- enabling the Information Commissioner to take a more risk-based approach to regulating digital services.

As at the date of writing, the direction of travel for these updates remains unclear.

40.5 **Technical standards**. 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).

K. THE LEGAL FRAMEWORK FOR DATA: A COMPLEX PICTURE

41. **The legal framework for data: a complex picture**. The legal framework for data presents a complex picture:

41.1 ***Each type of right is subject to its own rules***: IP (and within IP rights generally, each of copyright, database right, confidentiality and trade secrets), contract and regulation are discrete sets of norms each with their own technical (and sometimes mutually inconsistent) rules;

41.2 ***Rights are primarily national and operate differently in different countries***: not only are rights and duties subject to their own rules, but the rule sets concerned are national, conferred by national law and enforceable (primarily and initially) in national courts and so operate differently in different countries. Differences abound, for example:

41.2.1 the USA has a generic 'fair dealing' exception to copyright infringement, whereas the UK/EU has a long list of specific 'permitted act' exceptions);

41.2.2 database right is 'made in Europe/UK ' and does not apply to databases made in the USA;

41.2.3 some countries operate a copyright registration facility (e.g. the US) or requirements, whilst in others copyright arises by operation of law with no possibility of registration;

41.2.4 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 transposition); and

41.2.5 as a result of Brexit, UK and EU law are likely increasingly to diverge, causing further complexity at the technical level concerned (say IP rights or regulation).

41.3 ***Rights and duties act concurrently on each element of the data stack***: 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 – may also be subject to IP as database right, copyright or

trade secret (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);

41.4 ***Rights and duties are multi-layered***: 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/UK, 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;

41.5 ***Data may be created at great speed, increasing the evidence burden***: the computer processes by which data is created – e.g. 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.

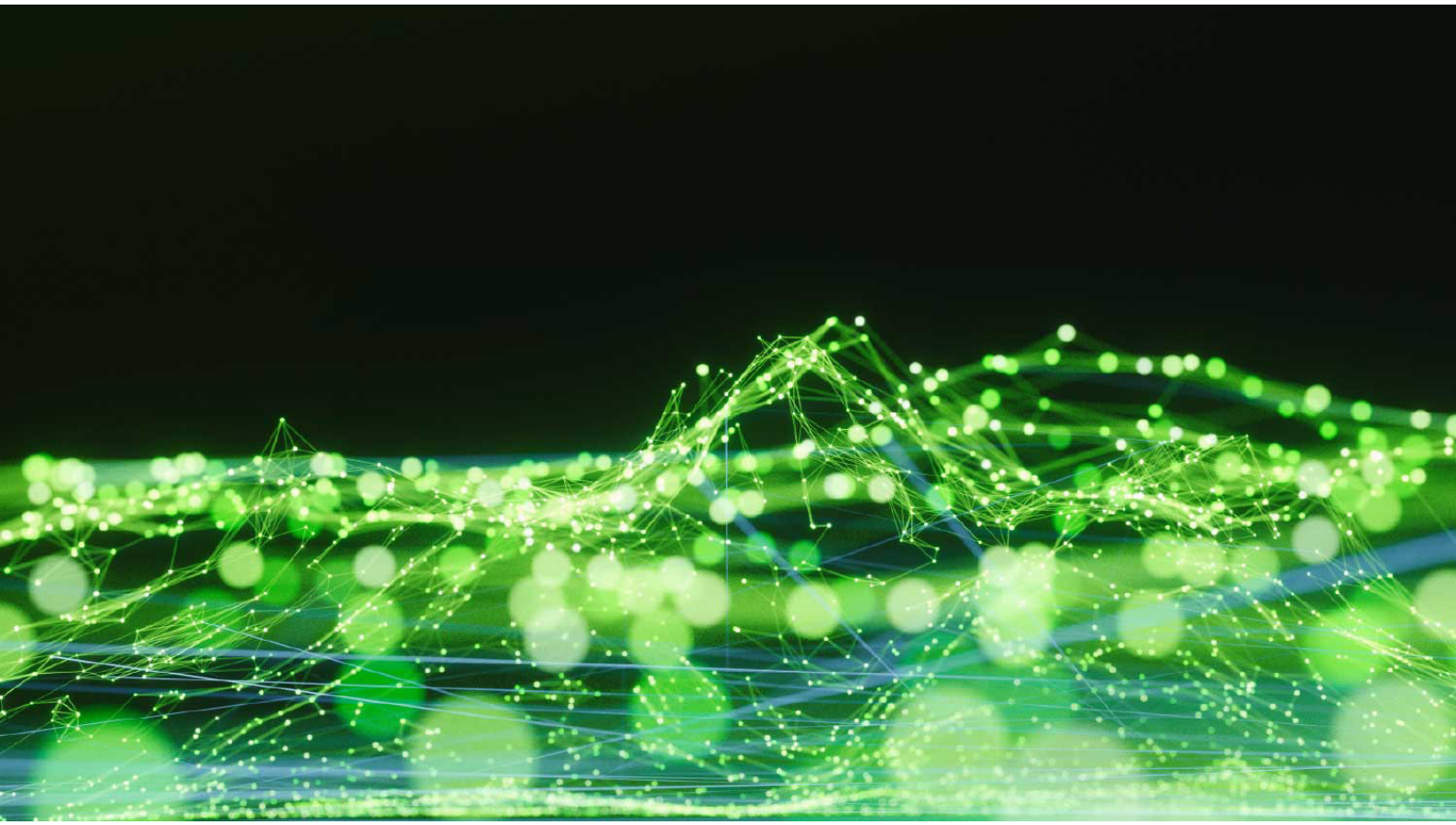
These differences between types of right in one country; the differences between similar rights in different countries; the way in which different rights act concurrently on the stack; the 'multi-layered'-ness of rights in the data lifecycle and the speed of the processes that create the data being assessed each contribute to the legal complexity of the data rights picture and the legal challenges of data projects.

L. CONCLUSION

42. **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.

KEMP IT LAW

Tech Law at the Apex



Richard Kemp
Partner

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