Legal Aspects of Managing Big Data

Richard Kemp

September 2014

KEMP IT LAW

kempitlaw.com

21 Napier Avenue London SW6 3PS Tel 020 3011 1667 Web www.kempitlaw.com

TABLE OF CONTENTS

TABLE OF FIGURES

LEGAL ASPECTS OF MANAGING BIG DATA

A. INTRODUCTION

1. **'Big Data is everywhere'**. 'If you haven't heard' trumpeted the Financial Times' Lex column of 27 June 2014, 'Big Data is everywhere'.[1] Over the past twenty years, the bow wave in IT has moved on from hardware and software to the data that they process, and in an increasingly competitive and data-centric world, harnessing the tides of the Big Data ocean will confer competitive advantage in enabling a company to know more about its customers and market place than its competitors.

Commenting that the business intelligence and analytics ('**BIA**') software market is worth $16bn a year and growing at 8% a year, the FT Lex column called out research from consultancy Gartner Inc.[2] who showed that the BIA market is currently undergoing an 'accelerated transformation' from retrospective BIA software - used mainly for measurement and reporting - to prospective BIA software used for prediction, forecasting and modelling. This is fuelling a race as the BIA software majors – Oracle, SAP, IBM and SAS, whose combined BIA software turnover totals around $10bn – vie with smaller, faster growing BIA specialists like QlikTech, Splunk and Tableau to bridge the gap between the oceans of available Big Data and BIA software's ability to harness it for competitive advantage in a structured, legally compliant way.

The European Commission (**Commission**) in its Communication of 2 July 2014[3], quoting a UK report, also comments on this accelerating growth:

> "Big data technology and services are expected to grow worldwide to USD 16.9 billion in 2015 at a compound annual growth rate of 40% – about seven times that of the information and communications technology (ICT) market overall. A recent study predicts that in the UK alone, the number of specialist big data staff working in larger firms will increase by more than 240% over the next five years."

It is this race for competitive advantage – knowing more than your competitor not so much about what your customers have just done as about what they are likely to do next – that is at the commercial epicentre of Big Data. But it is a race that is just beginning: Gartner also points out[4] that only 15% of Fortune 500 companies will be able to exploit Big Data for competitive advantage by the end of 2015 and that only 8% of companies are currently using Big Data analytics at all.

---

[1] http://www.ft.com/cms/s/3/525236ca-fd4f-11e3-bc93-00144feab7de.html?siteedition=uk#axzz35vtpzx2A

[2] http://www.gartner.com/technology/reprints.do?id=1-1QHKSEP&ct=140206&st=sb

[3] *Towards a thriving data-driven economy* (COM(2014) 442 Final) at https://ec.europa.eu/digital-agenda/en/news/communication-data-driven-economy

[4] http://www.gartner.com/technology/topics/big-data.jsp

2. **The US NIC's December 2012 report**. Big Data's direction of travel is well signposted in the December 2012 long range report of the US National Intelligence Council '*Global Trends 2030: Alternative Worlds*'[5] where it articulates a focus on data solutions and Big Data as a key IT driver over the next two decades:

> "Information technology is entering the Big Data era. Process power and data storage are becoming almost free; networks and the cloud will provide global access; and pervasive services; social media and cybersecurity will be large new markets."[6]

Opportunities arising through Big Data are not without their challenges and issues however:

> "Since modern data solutions have emerged, big datasets have grown exponentially in size. At the same time, the various building blocks of knowledge discovery, as well as the software tools and best practices available to organizations that handle big datasets, have not kept pace with such growth. As a result, a large - and very rapidly growing - gap exists between the amount of data that organizations can accumulate and organizations' abilities to leverage those data in a way that is useful. Ideally, artificial intelligence, data visualization technologies and organizational best practices will evolve to the point where data solutions ensure that people who need the information get access to the right information at the right time - and don't become overloaded with confusing or irrelevant information."[7]

It is these challenges and issues that the fast growing BIA software market is seeking to address.

3. **What is 'Big Data'?** As used in this White Paper, 'Big Data' is shorthand for the aggregation, analysis and increasing value of vast exploitable datasets of unstructured and structured digital information. Along with Cloud[8], mobile[9] and social computing, it is one of the four main drivers of change in information technology as it moves into new areas whose features currently include machine learning, 3D printing, virtual reality, the Internet of Things and nanotechnology.

Two recent papers, one from each side of the Atlantic, have addressed Big Data. Commenting that there was no one generally accepted definition, the White House's Executive Office of the President (EOP) in a report dated 1 May 2014[10] nevertheless gave a useful description:

> "Most definitions reflect the growing technological ability to capture, aggregate, and process an ever-greater volume, velocity, and variety of data. In other words, "data is now available faster, has greater coverage and scope, and includes new types of observations and measurements

---

[5] http://globaltrends2030.files.wordpress.com/2012/11/global-trends-2030-november2012.pdf.

[6] At page ix.

[7] At page 85.

[8] See Kemp et al, '*Cloud computing: the rise of service-based computing*' in Practical Law - http://uk.practicallaw.com/2-385-1280.

[9] See Kemp, '*Mobile payments: current and emerging regulatory and contracting issues*' (29 CLSR [2], pp. 175-179), or Practical Law at http://uk.practicallaw.com/3-523-4318?q=mobile+payments.

[10] *Big Data: Seizing Opportunities, Preserving Value*', http://www.whitehouse.gov/issues/technology/big-data-review. The report focuses on 'how big data will transform the way we live and work and alter the relationships between government, citizens, businesses, and consumers'.

> that previously were not available."[11] More precisely, big datasets are "large, diverse, complex, longitudinal, and/or distributed datasets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future."[12]

The Commission in its Communication of 2 July 2014 referred to above gives a similar description, which also covers the analytics aspects:

> "The term "Big Data" refers to large amounts of different types of data produced with high velocity from a high number of various types of sources. Handling today's highly variable and real-time datasets requires new tools and methods, such as powerful processors, software and algorithms, [g]oing beyond traditional "data mining" tools designed to handle mainly low-variety, small scale and static datasets, often manually"[13].

Big Data is therefore characterised by:

- **aggregation:**
  - ***size*** – vast volumes of digital data;
  - ***shape*** – in many variable formats (text, image, video, sound, etc.);
  - ***structure*** – in unstructured (typically, 80%) as well as structured (typically, 20%) varieties;
  - ***speed*** – arriving at a faster velocity;
- **analysis:**
  - these aggregated datasets analysed on a ***real-time*** rather than ***batch*** basis;
  - by ***quantitative analysis*** software (using artificial intelligence, machine learning, neural networks, robotics and algorithmic computation);
  - enabling a shift from ***retrospective*** to ***predictive*** insight;
- **increasing value:**
  - facilitating small but constant, fast and ***incremental business change***;
  - enhancing ***competitiveness efficiency*** *and* ***innovation*** and the value of the data so used.

4. **The policy perspective – the Commission's July 2014 Communication**. The Commission Communication of 2 July 2014 *Towards a thriving data-driven economy* referred to above sets out a number of activities it considers necessary "to be able to seize [Big Data] opportunities and compete globally in the data economy" including:

---

[11] Liran Einav and Jonathan Levin, "The Data Revolution and Economic Analysis," Working Paper, No. 19035, *National Bureau of Economic Research*, 2013, http://www.nber.org/papers/w19035; Viktor Mayer-Schonberger and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, (Houghton Mifflin Harcourt, 2013).

[12] National Science Foundation, Solicitation 12-499: *Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA),* 2012, http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.pdf.

[13] at page 4

- supporting 'lighthouse' data initiatives (like personalised medicine in healthcare, integrated regional transportation management and food chain management tracking food from farm to fork);
- focusing public research and investment 'on technological, legal and other bottlenecks';
- making 'sure that the relevant legal framework and policies, such as on interoperability, data protection, security and IPR are data-friendly, leading to more regulatory certainty for business and creating consumer trust in data technologies';
- rapidly concluding 'the legislative processes on the reform of the EU data protection framework, network and information security' and 'supporting exchange and cooperation between the relevant enforcement authorities (e.g. for data protection, consumer protection and network security)';
- accelerating 'the digitisation of public administration'; and
- using 'public procurement to bring the results of data technologies to the market'.

5. **Scope and aims of this white paper**. The main purpose of this paper is to provide a practical overview of the legal aspects of Big Data management and governance projects. In order to illustrate how Big Data and BIA software are beginning to have real impact and provide context for the discussion that follows, ***Section B*** briefly overviews Big Data initiatives and potential in a number of different vertical sectors (financial services, insurance, healthcare, air travel, music and public sector). The focus is then on providing three 'views' of Big Data from the legal perspective:

- ***Section C*** offers a common legal analytical framework for Big Data, centred on intellectual property rights in relation to data, contracting for data and data regulation;
- ***Section D*** considers Big Data within the organisation from the standpoint of input, processing and output operations; and
- ***Section E*** overviews the key aspects of Big Data management projects from the perspective of governance, addressing risk assessment, strategy, policy and processes/procedures.

The Legal and the IT Groups are likely to be the two business functions most closely associated with an organisation's Big Data management project. This paper addresses primarily the issues that will be relevant for the Legal Group rather than the IT group, but data modelling is addressed in outline at Sections B and D in view of its central importance. Detailed discussion of the technical aspects of data law and the detail of Big Data governance is outside the scope of this paper, but references are provided[14] to further materials where these aspects are discussed at greater length.

---

[14] For a more detailed review of the technical aspects of data law see Kemp et al, '*Legal Rights in Data*' (27 CLSR [2], pp. 139-151), or Practical Law at http://uk.practicallaw.com/5-504-1074?q=Big+Data+Kemp.

For a more detailed review of governance in a related area – Open Source Software – and points for consideration in strategy and policy statements and processes/procedures, see Kemp, '*Open source software*

B. THE BUSINESS CONTEXT: BIG DATA IN KEY VERTICAL SECTORS

6. **Introduction: pebbles and mountains**. This section provides an overview of current developments in a number of different vertical sectors (financial services, insurance, healthcare, air travel, music and public sector) in order to emphasise the scale of the changes that are occurring and how and where they are happening. Big Data is best seen currently as bringing about small incremental changes, but in relation to large amounts of aggregated data. Citing tests conducted by Facebook where behavioural differences of between 0.04% and 0.1% were accounted positively, the Economist recently commented[15] that constant experimentation and rapid implementation producing "small effects [having] large aggregated consequences" may be "the unspoken secret of Big Data". In a tag that has been picked up extensively by social media, the Chairman of Cloud Big Data analytics developer Applied Predictive Technologies was quoted in the Economist piece as saying that Big Data was "about building a mountain with pebbles".

7. **The banking sector**. The banking sector is one of the largest users of IT globally. Trading platforms – complex computer systems facilitating secondary trading in securities, derivatives and other financial instruments – are its beating heart and data its lifeblood. Market data – the data that these platforms generate - is a $25bn global industry, based on an ecosystem of exchanges and other data sources, index providers, data revendors, and data users on the buy-side (asset managers) and sell-side (banks and brokers). The ecosystem is held together by contract, with market practice based on contract structures that license, restrict and allocate risk around data use. From the legal perspective, these contracts constitute a stable cohesive normative framework in a market that has seen surprisingly little litigation.

   As an alphabet spaghetti of new rulebooks finally emerges from the 2008 financial crisis, the financial instrument trading regime that has applied to equities across the EU since 2007 will shortly be extended to most other asset classes by MiFID II[16]. MiFID II effectively takes MiFID I's regulatory template for public price transparency for equities and extends it to the secondary market for bonds, OTC derivatives and most structured finance products. It makes its contribution to the dawning era of Big Data by requiring pre- and post- contract price data to be disclosed and reported to the market for trades in all the securities that it regulates. As was the case for MiFID I and equities after 2007, MiFID II is likley to lead to hefty growth in the market data world.

   The degree of transformation that the new rulebooks are imposing, not just on IT platforms and data but across the whole spectrum of financial instrument trading, sets the scene for widespread adoption of Big Data techniques in the banking sector as trading operations and procedures that

---

(OSS) governance in the organisation' (26 CLSR [3] pp. 309–316), or Practical Law at http://uk.practicallaw.com/3-501-0318?q=open+source+governance.

[15] The Economist, 19 July 2014, Schumpeter, p. 66 http://www.economist.com/news/business/21607816-businesses-should-aim-lots-small-wins-big-data-add-up-something-big-little

[16] 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) and 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). MiFID II and MiFIR are scheduled to come into force on 3 January 2017.

have developed incrementally since the onset of computerised trading in the 1970s are re-written to comply with the more prescriptive requirements of the new rules.

8. **Information architecture in the banking sector: TOGAF and BIAN**. 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, TOGAF (The Open Group Architecture Framework[17]), which operates an open standards based enterprise IA framework, and BIAN (the Banking Industry Architecture Network[18]), which operates a banking specific standard IA based on SOA[19], have announced[20] cooperation so as 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. An IA's database schema – the 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 it through levels of increasing abstraction and maps it to a data model - 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.

   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.

9. **The insurance sector**. In insurance, where the insured transfers the risk of a particular loss to the insurer by paying a premium in return for the insurer's commitment to pay if the loss occurs, Big Data enables risk to be assessed much more precisely than in the past by reference to specific data about the insured and the risk insured, and hence enables the price of the policy to be calculated more accurately.

   As well as the traditional 'top down' statistical and actuarial techniques of risk calibration and pricing, insurers can now rely on actual data relating to the insured concerned. For example, in vehicle insurance, location based data from the driver's mobile can show where the insured was, and telematics data from on-board IT can show how safely they were driving, at the time of the accident.

---

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

[18] 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.

[19] 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.

[20] See e.g. https://bian.org/participate/bian-webinars/recorded-sessions/collaboration-between-bian-togaf/.

Similarly, smart domestic sensors can help improve responsiveness to the risk of fire, flooding or theft at home, and health apps and 'wearables' – body-borne small electronic devices - can provide information relevant to health and life insurance.

These examples – data sourced remotely from telematics, location based services, home sensors and wearables – are early illustrations of Big Data (and also the 'Internet of Things') in consumer insurance. They will over time have a material impact on the pricing of vehicle, home and health policies.

Big Data in insurance also points up two other common themes. First, the tension between Big Data and the privacy of the insured's personal data and its availability to business and the State – a tension that becomes greater when considering data about genetic pre-disposition to illness and the availability and price of health and life insurance; and secondly, as in the banking sector, the regulatory dimension, where an impulse towards Big Data adoption is Solvency II[21] which will regulate the amount of capital that an EU insurance company must hold against the risk of its insolvency, in turn based on likelihood of aggregated policy pay outs.

10. **The air transport industry**. The air transport industry ('**ATI**') has grown up with computerisation and standardisation as key components in getting passengers (three billion globally in 2012) and their baggage to the airport of departure, on to the plane, and to and from the airport of arrival. In doing so, airlines and other ATI companies generate and hold vast amounts of data about customers' preferences during all stages of their journey. But this data can be siloed in a particular application or airline, so as competitive pressures tend both to increase the popularity or air travel and reduce prices, Big Data techniques will emerge to support these trends[22]. Gathering, analysing and using Big Data will enable ATI players to develop insights about customers and their air travel preferences, and doing this better than its competitors will give a particular airline a competitive advantage.

    In particular, the ATI illustrates the importance to Big Data of mobile in consumer markets and m-commerce through the mobile phone's unique features as data source, data store and processing point. For the airline customer, the mobile wallet facilitates paperless ticketing and boarding passes and its NFC (near field communication) feature enables mobile check in, each improving efficiency and reducing time and costs at the point of sale and in the airport.

11. **The recorded music industry**. The recorded music industry is a $15bn global business that is being transformed by digitisation as developing patterns of online consumption through streaming and downloading continue to displace purchases of physical music product. The structure of the industry has grown up around norms based on the individual and collective licensing and management of the various and distinct copyrights that arise in a song's composition, lyrics and publication, and in its recording and performance. These copyright norms operate primarily on a national basis, as copyright is a right conferred by national law, with international harmonisation and

---

[21] 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), scheduled to come into force on 1 January 2016.

[22] http://www.sita.aero/content/big-data-big-insights.

equivalence mediated through international copyright treaties like the Berne Convention and WIPO Treaties.

The big three record companies (Universal, Sony BMG and Warner) together account for around 70% of the global recorded music market. The music track is effectively the product unit for the sector, and PPL, the UK CMO (Collective Management Organisation) for the public performance rights of its 11,500 recording rightsholder members and 79,000 performer members, operates a computerised repertoire database of 6.7 million tracks that is currently growing by 18,000 sound recordings per week. Management of data is a large part of PPL's work, driving more accurate distributions and better international collections, where the trend is towards standardising of data submission and exchange formats between country CMOs, their members and licensees.

With supply and demand increasingly operating online and on a global basis, the record industry is another sector where Big Data techniques will enable existing structured datasets relating to music to be combined with unstructured data from sources like social media and mobile so as rapidly to gain insights into consumer preferences. These insights up to now have been the particular province of record company A&R (Artiste & Repertoire) teams, and it is likely that in future Big Data will increasingly influence musical taste, fashion and trends and hence the creation of music itself in a way that has not been possible before.

12. **The healthcare sector**. Healthcare is the sector where adoption and use of Big Data is likely to have the greatest impact on people's daily lives. In its January 2013 report '*The 'big data' revolution in healthcare*'[23], consultants McKinsey & Co pointed to four changes that were creating a tipping point for innovation in healthcare around Big Data:

- demand-side pressures for better data are growing as cost pressures intensify, structural reforms continue and early movers and adopters demonstrate advantage;
- on the supply side, national collections of clinical and treatment outcome data are starting to become available in particular areas (for example cardiac in the UK);
- investment is gathering pace in technical developments for aggregating and anonymising data from individual hospitals and treatment centres and in the BIA software tools that generate insights from them; and
- governments are catalysing market change by their continuing commitment to making data publicly available and through the creation of interoperability standards that encourage private sector participation.

[23] http://www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health_care.

Although the McKinsey report focused on the USA, these change agents are even more powerful in the UK through the NHS (whose budget for 2014 is around £120bn, or 8% of UK GDP), a 'relentless' producer of Big Data in the words of a report in the Guardian newspaper[24].

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

Summer 2014 has seen the issue of data sharing within government rise up the agenda with increasing press interest[25] around policy developments following publication by the Cabinet Office Data Sharing Policy Team on 9 April 2014 of their Initial Discussion Document[26]. This advocates an open policy making approach to balancing the delivery of better public services through the removal of barriers to sharing or linking different datasets with potential concerns of citizens and safeguarding people's privacy. Ideas put forward in the discussion document include developing the December 2012 proposals of the Administrative Data Taskforce[27] for two models – the Trusted Third Party and the Firewall Single Centre - that would each allow data sharing for cross-linked research on de-identified data whilst restricting access to and use of identity data to the extent needed to cross-link the datasets concerned. Structural safeguards proposed include accreditation and registration of projects and individuals having access to de-identified data; a formal process to be carried out by the UK Statistics Authority to accredit the four Administrative Data Research (**ADR**) Centres that form part of the ADR Network[28], HMG's vehicle for public sector Big Data; and compliance with the Data Sharing[29] and Anonymisation[30] Codes of Practice published by the Information Commissioner's Office (**ICO**), the UK Data Protection regulator.

---

[24] http://www.theguardian.com/healthcare-network/2013/apr/25/big-data-nhs-analytics.

[25] See e.g. the Daily Telegraph of 3 August 2014 - http://www.telegraph.co.uk/news/11009405/Revealed-Ministers-blueprint-to-share-private-data.html

[26] http://datasharing.org.uk/current-proposals/

[27] Report of the Administrative Data Taskforce (a collaborative initiative between the Economic and Social Research Council, the Medial Research Council and Wellcome Trust) on Improving Access for Research and Policy - http://www.esrc.ac.uk/_images/ADT-Improving-Access-for-Research-and-Policy_tcm8-24462.pdf.

[28] See the UK Data Service's news article of 25 June 2014 at http://ukdataservice.ac.uk/news-and-events/newsitem/?id=3835

[29] http://ico.org.uk/for_organisations/data_protection/topic_guides/data_sharing. See below paragraph 28.

[30] http://ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation. See below paragraph 28.

C. TOWARDS A COMMON LEGAL FRAMEWORK FOR BIG DATA

14. **Introduction: what is data in legal terms?** A reasonable start point for the discussion about the legal framework for Big Data is to ask: what is the nature of information and data? For present purposes, information is that which informs and 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[31]:

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

Unlike real estate for example, information and data as expression and communication are limitless and it would be reasonable to suppose that subjecting information to legal rules about ownership would be incompatible with its nature as without boundary or limit. Yet digital information is only available because of investment in IT, just as music, books and films require investment in creative effort.

This equivocal position is reflected in the start point for the legal analysis, which is that data is funny stuff in legal terms. This is best explained by saying there are no rights ***in*** data but that extensive rights and obligations arise ***in relation to*** data. The UK criminal law case of Oxford v Moss[32] is generally taken as authority for the proposition that there is no property ***in*** data as it cannot be stolen; and a recent case in the UK Court of Appeal[33] has confirmed that a lien (a right entitling a person in possession to retain it in certain circumstances) does not subsist over a database. However, the rights and duties that arise ***in relation to*** data are both valuable and potentially onerous and, as an area of law, developing rapidly at the moment. They are likely to develop even more quickly as Big Data techniques become more prevalent.

---

[31] 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.

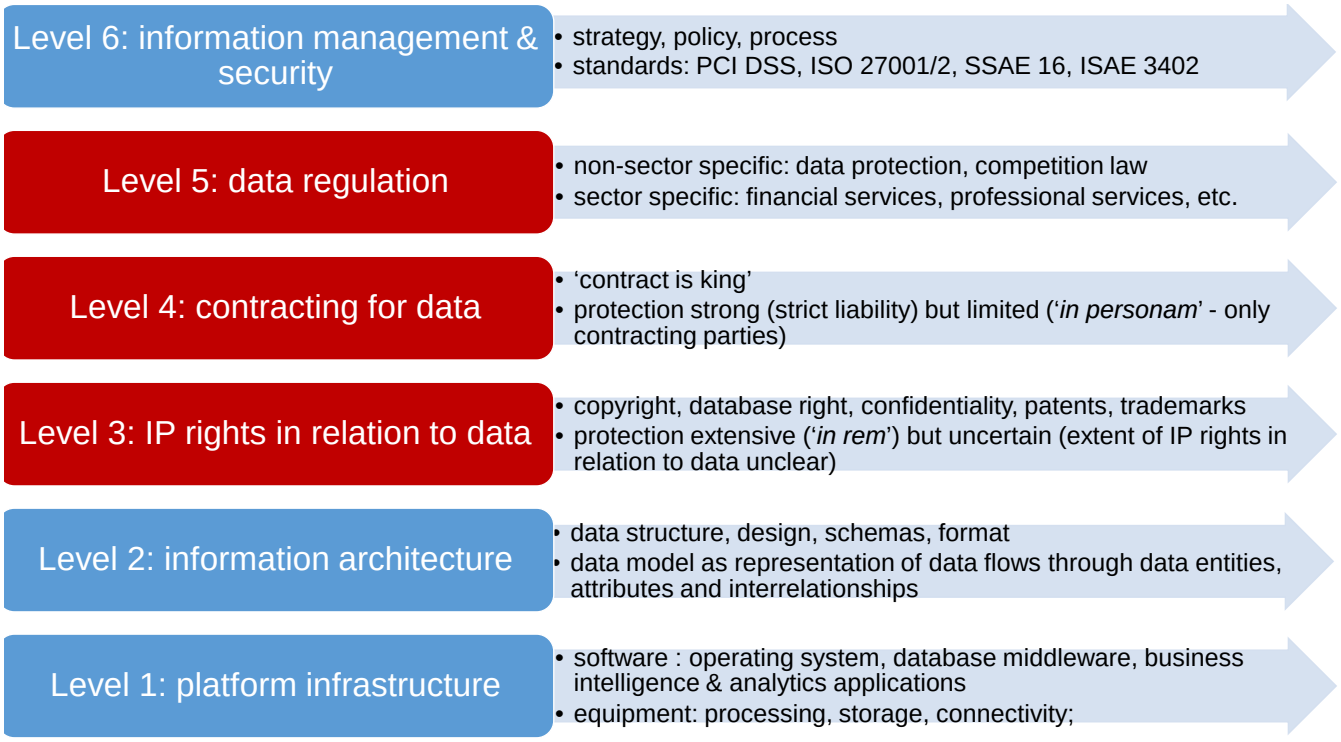
[32] [1979] Crim LR 119, where it was held that confidential information in an exam question was not 'intangible property' within the meaning of Section 4(1) of the Theft Act 1968 and so could not be stolen

[33] Your Response Ltd v Datateam Business Media Ltd, judgment of the Court of Appeal 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 under English law is traditionally a possessory remedy available only in respect of 'things' (or 'choses') in possession – i.e. personal tangible property. A database on the other hand is a 'thing' (or chose) in action – i.e. something capable ultimately of enjoyment only through court action – so that this case should not be taken as authority for the proposition that that there is no property in a database, just that there is no personal tangible property.

These rights and duties arise through intellectual property rights ('**IPR**'), contract and regulation. They are important as (positively, in the case of IPR and contract) they can increasingly be monetised and (negatively) breach can give rise to extensive damages and other remedies (for IPR infringement and breach of contract) and fines and other sanctions (breach of regulatory duty)[34]. Current developments in each of these areas mean that 'data law' is emerging as a new area in its own right around these three constituents of IPR, contract and regulation.

15. **The 6 level data stack**. IPR, contract and regulation in the Big Data context can be conceptualised in a legal analytical model as the middle three layers of a 6 layer stack, sandwiched between platform infrastructure and information architecture below and information management and security above (see Figure 1 below, towards a common legal framework for Big Data).

Figure 1: towards a common legal framework for Big Data



Level 1: Platform Infrastructure

16. **Level 1: platform infrastructure**. This level consists of the platform's physical infrastructure – servers, storage, user devices, routers, local network, internet connectivity, etc. - and the software that resides on the platform – operating system, middleware data access and connectivity software and applications like BIA referred to above. The legal analysis at this level tends to be around traditional software copyright issues (rights in computer languages, software 'look and feel', etc.) and the interrelationships between copyright and database right in relation to database software and

[34] For a more detailed review of the technical aspects of data law see Kemp et al, '*Legal Rights in Data*' (27 CLSR [2], pp. 139-151), or Practical Law at http://uk.practicallaw.com/5-504-1074?q=Big+Data+Kemp.

accessing and extracting the data held in that software.[35] The increasing degree of interoperability in a world that is ever more interconnected is also focusing legal attention increasingly on how the technical standards that have been adopted to achieve these network effects are used.

Level 2: Information Architecture

17. **Level 2: information architecture**. The information architecture or IA is the intermediate level between the platform infrastructure and the data itself and, as observed at Sections B.8 and B.9 above, sits at the centre of networked and therefore standardised data flows. The IPR position of the IA itself is easily overlooked in practice, and is worth calling out for attention. Here the documentation describing and specifying the architecture will attract traditional literary copyright protection in the normal way; and the database 'schema' or formal structure (as distinct from the data content of a database) will be protectible by copyright in the EU under Chapter II, Article 3 of the Database Directive.[36] In the context of a standardised IA the question how the IPR in it will be licensed will normally be determined by the IPR policy applicable to the relevant SSO (Standards Setting Organisation), TC (Technical Committee) or individual organisation that manages the standard.

Level 3: IP Rights in Relation to Data

18. **Level 3: intellectual property rights in relation to data (i) - introduction**. The main IP rights in relation to data are copyright (paragraph 19), database right (paragraph 20) and confidentiality (paragraph 21), which are now briefly overviewed in the data context. 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.

19. **Level 3: IP rights in relation to data (ii) – copyright**.

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

Ingredients for a successful copyright infringement claim. A successful claim for copyright infringement will need to show:

---

[35] See for example Navitaire Inc v Easyjet Airline Company and Bulletproof Technologies, Inc - http://www.bailii.org/ew/cases/EWHC/Ch/2004/1725.html. This case is discussed in the paper referred to at footnote 14 above.

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

- that copyright subsists in the work – generally, that it is original (where the usual UK standard is low and normally that the work concerned has not been copied from elsewhere) and sufficient to warrant copyright protection (where the English courts typically take the pragmatic line that 'what is worth copying is worth protecting');
- that the claimant owned or could otherwise sue on that copyright;
- that the work was within copyright (life plus seventy years in the case of software, databases and other literary works); and
- that the copyright had been infringed – for example, a qualitatively substantial part of the work had been reproduced without authorisation in circumstances where a copyright permitted act exception did not apply.

Copyright and data. In the context of data, traditional literary copyright will subsist in documentation – for example, publications relating to research[37] and stock market analysis[38], and the technical and user documentation relating to computer software and (as mentioned at paragraph 17 above) information architecture. Computer programs and preparatory design material for a computer program have been subject to literary work copyright protection in the UK since 1985 and 1993 respectively. Moral rights (for example the rights to be identified as author and to object to derogatory treatment of the work) apply to literary work copyright but not to software.

Database copyright. Database copyright is subtly different from copyright in software and other written work. This is the result of the changes to Sections 3 and 3A of the UK Copyright, Designs and Patents Act 1988[39] ('**CDPA**') that were made in 1998 to accommodate the introduction into English law of database right (see paragraph 20 below) by:

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

Database copyright and the Football Dataco cases. The new database copyright raised two central questions under English law: first, the relationship between the database and its contents; and secondly the new 'author's own intellectual creation' originality standard as it applied to content selection or arrangement. These questions were considered in relation to football fixtures in the

---

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

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

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

Football Dataco Ltd v *Brittens Pools Ltd/Yahoo UK Ltd* cases in the UK High Court and Court of Appeal (**CoA**) and the European Court of Justice (**ECJ**) between 2010 and 2012[40].

Briefly, Football Dataco, which had been appointed by the English and Scottish professional football leagues as their agent to license football fixture lists, 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. Accordingly, it held in the database copyright part of the case 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 Football Dataco were of no relevance 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.

As regards the originality threshold itself, the 'author's own intellectual creation' standard in relation to the structure of the database was met when the author expressed creative ability in an original manner by making free and creative choices – in effect when the author put their personal touch on the work. It followed when the case went back to the CoA on 20 November 2012 that football fixture lists did not benefit from database copyright.

20. **Level 3: IP rights in relation to data (iii) – database right**.

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

Ingredients for a successful database right infringement claim. Database right arises in a database (which bears the same meaning as under the CDPA – see paragraph 19 above) in whose 'obtaining, verifying or presentation' 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 obtaining, verifying or presenting its contents. The right lasts for fifteen years from initial creation, effectively refreshed wherever 'any substantial change' is made. It is infringed

---

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

[41] Council Directive 96/9/EC of 11 March 1996 on the legal protection of databases (OJ L 77/1996 20, http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:31996L0009

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

by 'extraction and/or re-utilization' of a substantial part of the database contents on a one-off basis or repeatedly and systematically of insubstantial parts.

Subsistence of 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 November 2004 of which the ***BHB*** case[43] is the most important. Here the ECJ considered what was meant in the 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:

> "[t]he expression 'investment in …the obtaining …of the contents' of a database in … [the 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 narrowed down the scope of database right considerably, especially for real time databases in the financial services industry for example where the creation of underlying data (like securities trades), their collation into a database and their verification are effectively instantaneous.

Subsistence of database right and the Football Dataco cases. That this principle is not free from difficulty was shown in the CoA judgment of 6 February 2013[44] in another case involving Football Dataco, this time where the counterparties were Sportradar GmbH and Stan James plc. Here, the subject of the dispute was Football Dataco's '*Football Live*' service which published live and online factual match information (like goals, scorers, substitutions and red and yellow cards). Defendants Sportradar published a competitive service '*Sport Live Data*' which they 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 Football Dataco's resources went into collecting the data generated from the football matches, not creating that data, and upheld the first instance judgment, again of Floyd J[45], that '*Football Live*' was protected by database right. The CoA judgment in ***Football Dataco*** v ***Sportradar*** thus 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.

---

[43] 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 and repent at leisure*' (22 CLSR [6], pp 493-498).

[44] ***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

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

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

On the other hand, indirect as well as direct acts can constitute extraction and re-utilisation; exhaustion of rights (the EU term for the first sale doctrine in the USA) does not apply to re-utilisation (***BHB***); and re-utilisation covers any distribution of any part of the database, and can take place in any EU country where the alleged infringer intends to target members of the public (***Sportradar*** in the ECJ[46]).

21. **Level 3: IP rights in relation to data (iv) – confidentiality and trade secrets**.

Data and confidentiality. Copyright and database right both protect expression and form rather than the substance of information. This means, somewhat counterintuitively, that equitable rules protecting confidentiality of information ('equity will intervene to enforce a confidence') very often provide the best form of IPR-type protection as they can protect the substance of data that is not generally publicly known. There is a long line of cases in the UK[47] showing that protection can extend to aggregation of datasets even where parts of the data are in the public domain and so not otherwise confidential. Protection may also extend to second and subsequent generation data derived from the initial confidential data.

The draft EU Trade Secrets directive. On 28 November 2013, the Commission published[48] a draft directive to harmonise trade secret protection across the EU by setting common standards to protect trade secrets against unlawful acquisition, use and disclosure. On 26 May 2014, the Council of the European Union (**Council**) agreed a general approach to the proposed directive supporting the Commission's proposal it whose principal features are:

- "a minimum harmonisation of the different civil law regimes, whilst allowing member states to apply stricter rules;

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

[47] *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).

[48] http://eur-lex.europa.eu/legal-content/EN/ALL/;ELX_SESSIONID=7RQpT0fHnwkbhB2TLQPMWYh2dwjKQV8BgjGck9MNn9QTxhkYYBkl!-563378897?uri=CELEX:52013PC0813

- the establishment of common principles, definitions and safeguards, in line with international agreements, as well as the measures, procedures and remedies that should be made available for the purpose of civil law redress;
- a limitation period of six years for claims or bringing actions before courts;
- the preservation of confidentiality in the course of legal proceedings, while ensuring that the rights of the parties involved in a trade secret ligation case are not undermined;
- the establishment of a favourable regime to employees in what concerns their liability for damages in case of violation of a trade secret if acting without intent".[49]

It is anticipated that the European Parliament (**Parliament**) will give its opinion on the draft directive later in 2014. Once implemented into national law, the directive will bring EU law more closely into line with Article 39 of the WTO TRIPS Agreement[50] (which gives IPR protection to trade secrets as undisclosed information) and the US Uniform Trade Secrets Act[51]. It will provide in the business context another line of attack or defence in addition to established rules on confidentiality to those seeking to enforce or disapply the confidentiality or secrecy of data.

22. **Level 3: IP rights in relation to data (v) – likely direction of travel**. IP rights in relation to data are of uncertain scope at the moment, and the law in this area will continue to develop in the coming years as Big Data gathers pace. Historically, IPR development has followed the commercialising of innovation, and as the value of Big Data rises, so will the value of IP rights underpinning it and case law around database right, database copyright and (once enacted) the Trade Secrets Directive will grow as Big Data aggregation, analysis and value grow. Whilst uncertain, IPR are nevertheless extensive as rights '*in rem*' (enforceable against the whole world, not depending on a pre-existing relationship) with powerful infringement remedies, from temporary and permanent injunctions (court orders requiring termination of the infringement whose breach is sanctioned through contempt of court) to damages and account of profits.

Level 4: Contracting for Data

23. **Level 4: contracting for data (i) – introduction**. Contract rights in relation to data are technically entirely separate from IPR. Their value was confirmed in a UK High Court case in 2006 where the judge said that an owner of data:

> "is entitled in principle to impose a charge for use of its data by users whether or not it has IP rights in respect of that data".[52]

Conversely to IPR law, contract confers rights and imposes obligations that the law recognises as certain, strong and enforceable. If the good news is that data contracts are strong, the less good

---

[49] http://ec.europa.eu/internal_market/iprenforcement/trade_secrets/index_en.htm#maincontentSec1

[50] 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

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

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

news is that they operate '*in personam*' – unlike IP rights, they are only enforceable against a party to the agreement concerned and not against the whole world. Confusingly, contract can impose IPR-type obligations under the contractual wrapper, so contract IPR and IPR 'proper' also need to be considered separately. It is however fair to say that 'contract is king' in the world of data and it is this strength of contract law that underpins durable ecosystems like market data referred to at Section B.7 above.

24. **Level 4: contracting for data (ii) – key areas for Big Data**. Key areas for Big Data contracting are similar to those in other areas of data contracting. Particular attention should be focused on:

- scope of rights being licensed
  - internal use/onward dissemination;
  - territoriality – where the rights licensed may be used, etc;
  - combination/use with other data;
  - treatment of derived data;
- what purposes can the data be used for?
  - check whether anticipated analysis of data is expressly permitted;
  - what are the mechanisms for re-purposing/adding new purposes?
  - particularly for use of social media data, check that the standard terms of the provider expressly permit anticipated uses;
- ownership of underlying rights and rights to derived data;
- (mutual?) warranties of compliance with laws and regulation – data protection; sector specific regulation; audit/investigation;
- risk allocation:
  - reliance on data being provided – 'as is' or reasonable skill and care?
  - supplier and customer indemnity and liability positions;
- duration, suspension and termination of supply; and
- post-term use of data supplied in-term, derived data, etc.

Level 5: Data Regulation

25. **Level 5: data regulation (i) - introduction**. The third legal area of increasing importance for Big Data is regulation. Data protection the most important, but not the only area of regulation, and competition law and sector specific regulation are also likely to become increasingly important. General consumer regulation may also apply to Big Data but is not considered further here.

26. **Level 5: data regulation (ii) – Data Protection.** Data protection – conferring rights and imposing obligations on the processing of personal data as data relating to an identified or identifiable individual - continues to attract most attention. As the draft EU Data Protection Regulation continues its tortuous progress towards the statute book, it is becoming clearer that requirements for explicit, informed consent on the part of the individual to the use and processing of personal data about him

or her are likely to be become more generally applicable. Managing compliance with these requirements in future will play a large part in Big Data management projects involving data harvested from the expanding range of available digital sources that the White House EOP report mentioned at Section A.3 above is so concerned about. Many organisations will already have an established data protection governance structure and policy and compliance framework in place and these can be helpful as pathfinders towards structured Big Data governance.

27. **Level 5: data regulation (iii) – ICO's 28 July 2004 Report on Big Data and Data Protection.** On 28 July 2014, ICO published a paper on Big Data and Data Protection[53]. The paper applies the relevant principles of the Data Protection Act (**DPA**) to the different aspects of Big Data and provides useful practical pointers on how to address them.

Big Data and Data Protection – issues. The paper focuses particularly on:

- ***fairness (DPA Principle 1)***: pointing out that fairness is partly about how personal data is obtained, the paper notes that "processing is unlikely to be fair if people are deceived or misled about how their data will be used at the point they are providing it" so that transparency about how the data will be used (and hence the organisation's privacy notice) will be important in determining compliance with DPA principles (paragraph 48).
- ***consent (DPA Schedule 2, paragraph 1)***: an organisation which has collected data for one purpose needs to obtain users' consent before it starts analysing it for a different purpose that is not apparent to the individuals concerned:

  > "the apparent complexity of big data analytics should not become an excuse for failing to seek consent where it is required. Organisations must find the point at which to explain the benefits of the analytics and present users with a meaningful choice - and then respect that choice when they are processing their personal data" (paragraph 60).

- ***purpose limitation/repurposing (DP Principle 3)***: fairness is also relevant to deciding whether the new purpose is incompatible with the original purpose:

  > "If, for example, information that people have put on social media is going to be used to assess their health risks or their credit worthiness, or to market certain products to them, then unless they are informed of this and asked to give their consent, it is unlikely to be either fair or compatible. Where the new purpose would be otherwise unexpected, and it involves making decisions about them as individuals, then in most cases the organisation concerned will need to seek specific consent, in addition to establishing whether the new purpose is incompatible with the original reason for processing the data" (paragraph 69).

- ***data minimisation (DP Principles 3 and 5)***: Big Data analytics involves collecting as much data as possible ('N = all') and this causes tension with DPA data minimisation requirements:

  > "Big Data may discover unexpected correlations, for example between data about people's lifestyles and their credit worthiness, but that does not necessarily mean that any information that can be obtained about those individuals is necessarily relevant to the purpose of assessing credit risk. Finding the correlation does not retrospectively justify obtaining the data in the first place. Organisations therefore need to be able to articulate at

[53] http://ico.org.uk/for_organisations/data_protection/topic_guides/big_data

the outset why they need to collect and process particular datasets" (paragraph 73).

It would have been helpful for ICO to have expressed its views on the technical legal questions of quantifying the harm that individuals may suffer, and the corresponding liability that may arise, as a result of using non-DPA compliant Big Data analytics, and it is to be hoped that ICO will shed light on this before too long.

Practical pointers towards addressing Big Data Data Protection issues. Having illustrated how tension arises between the DPA and Big Data, the ICO paper also suggests pointers that organisations should address when considering Big Data analytics:

- **anonymisation**[54]: Although data is of course no longer personal data if fully anonymised[55], the growing power of Big Data means that absolute anonymisation may not be possible, so that organisations "should focus on mitigating the risks [of re-identification] to the point where the chance … is extremely remote" (paragraph 42) using "solutions proportionate to the risk [which] may involve a range and combination of technical measures such as data masking, pseudonymisation, aggregation and banding, as well as legal and organisational safeguards" (paragraph 43) and privacy by design (paragraphs 102 to 104). This formulation is rather bland, however, and the report shies away from more contentious technical considerations about the ability to re-identify anonymised and pseudonymised data.
- ***privacy Impact assessments***: the report advocates the privacy impact assessment[56] as a tool to be used before processing begins to assess how Big Data analytics is likely to affect the individuals whose data is being processed and whether processing is fair;
- ***building trust***: citing IBM and Nectar loyalty card operator Aimia, "some evidence" is noted of companies "developing an approach to Big Data that focuses on the impact of the analytics on individuals" (paragraph 137) with companies looking:

  > "to place big data in a wider and essentially ethical context. In other words, they are asking not only "can we do this with the data?", ie does it meet regulatory requirements, but also "should we do this with the data?" ie is it what customers expect, or should expect?".

  ICO comment favourably on this approach in terms of fairness and transparency:

  > "adopting an ethical approach of the type outlined in these examples will also go some way towards ensuring that the analytics complies with data protection principles" (paragraph 138).
- ***Information governance***: finally, and swelling the theme of 'a trust based ethical approach' ICO notes a growing emphasis on the issue of data quality and information governance in relation to

---

[54] See also http://ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation for ICO/s Anonymisation Code of Practice.

[55] where "it is not possible to identify an individual from the data itself or from that data in combination with other data, taking account of all the means that are reasonably likely to be used to identify them" (paragraph 40)

[56] http://ico.org.uk/for_organisations/data_protection/topic_guides/privacy_impact_assessment

Big Data analytics citing a report from Forrester Research from August 2013[57] (see Section E paragraph 38 below).

28. **Level 5: data regulation (iv) – competition law**. Privacy and data protection are by no means the only aspect of data regulation however. At the non-sector specific level, national and EU competition authorities have over the last five or so years been showing increasing interest in analysing through the lens of competition law business patterns, licensing and contracting for data in a number of sectors, particularly financial market data[58].

29. **Level 5: data regulation (v) – sector specific regulation**. Data regulation is also deepening in many vertical industry sectors. This is not necessarily a new thing – the rules on the confidentiality of client information and privilege have been cornerstones of the legal profession for generations. The digitisation of data is however changing the picture fundamentally, as shown by the examples from Section B above in the financial sector (MiFID II transparency requirements), insurance (Solvency II), the Air Travel Industry (specific rules on PNR – passenger name record – data about an airline customer's itinerary) and healthcare (rules about aggregating anonymised clinical outcome patient data).

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

Level 6: Information Management and Security

30. **Level 6: information management and security**. At the top of the Big Data common legal framework, at level 6, sits information management and security. 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

[57] http://www.ibmbigdatahub.com/whitepaper/big-data-needs-agile-information- and-integration-governance

[58] See the Art. 102 TFEU Commission Decisions of 15.11.2011, Case COMP/39/592 – Standard & Poor's - http://europa.eu/rapid/press-release_IP-11-1354_en.htm?locale=en; Commission Decision of 20.12.2012, Case AT.39654 – Reuter Instrument Codes: http://ec.europa.eu/competition/elojade/isef/case_details.cfm?proc_code=1_39654; the Credit Default Swaps investigation, where a Statement of Objections was issued on 01.07.2013 - http://europa.eu/rapid/press-release_IP-13-630_en.htm;

standards on how service companies should report on their information security and other compliance controls (for example SSAE 16 and ISAE 3402).

31. **The legal framework for Big Data – a complex picture**. The legal framework for Big Data presents a complex picture. First, IPR (and within IPR, each of copyright, database right and confidentiality), contract and regulation are discrete sets of norms each with their own technical (and sometimes mutually inconsistent) rules.

    Second, IPR, contract law and regulation act concurrently on each element of the data stack. A particular dataset – say PNR (passenger name record) data from the ATI – will also be subject to IPR as database right or copyright (in the IT system of an airline); contractual rights and duties (between the airline and a travel agent); and data protection regulation (as personal data relating to the passenger).

    Third, legal rights and duties arise in a multi-layered way. Data going through several database systems between creation and end use may be subject to a thin sliver of different database right owned by different actors at each stage as incremental investment is made. A bank subject to regulatory information security and audit duties may seek contractually to impose those requirements on its IT vendors in order to ensure that it is not beholden to its regulator without being able to enforce compliance from suppliers.

    Fourth, the computer processes by which data is created – for example financial market data – take place at great speed, so that the evidential burden in formal dispute resolution in showing what happened when is time consuming and costly.

    Fifth, IPR rule sets are national rights conferred by national law and enforceable (primarily and initially) in national courts and so operate differently in different countries. Differences vary from the minor (for example, the USA has a generic 'fair dealing' exception to copyright infringement, whereas the UK has a long list of specific 'permitted act' exceptions) to the major (database right is 'made in Europe' and does not apply to databases made in the USA; some countries operate a copyright registration requirement, whilst in others copyright arises by operation of law with no possibility of registration). In the area of regulation, directives in EU law are binding as to the objective to be achieved but leave implementation to each Member State, leading to significant differences in national approach.

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

D. BIG DATA OPERATIONS INSIDE THE ORGANISATION

32. **Introduction**. The 6 level stack at Section C provides this paper's first 'view' of Big Data, as a common legal analytical framework. This Section briefly overlays on to that view the organisation's Big Data operations – the input into, processing within, and output from, the Big Data 'engine' (see Figure 2 below, The Big Data engine – input, processing and output operations).

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

1. Input data	2. Processing operations	3. Output data
from multiple sources	**using multiple platforms**	**for multiple purposes**

• instruments & sensors
• email, video
• click streams
• internet transactions
• social media & mobile
• market data
• personal data
• confidential data
• licensed in data
• government data
• employee data
 	
- self-generated data
- derived data
	• processing of input data on:
	  • internal platforms/
	  • external platforms
	• using:
	  • business intelligence & analytics applications
	  • data visualisation
	  • machine learning
	  • 'secret sauce' algorithm
	  • pan-enterprise search
	• to produce
	  • 're-purposed' data
	  • output data
	  • an enterprise single view
		• for internal use
		  • business analysis
		  • product development
		  • sales & marketing
		  • risk management
		  • capital management
		  • MIS
		  • finance
		
		• for external use
		  • distribution
		  • marketing partners
		  • regulatory

33. **Data input operations**. Data comes into the Big Data engine from an increasingly wide variety of sources. The data can be structured – for example, a real-time feed of market data from an exchange or a bought (licensed) in marketing database; it can be confidential or publicly available; it can be personal data relating to individuals; and it can be one or more of these things at the same time. Increasingly, however, it consists of unstructured data - in the words of the White House EOP report Section A.3 above:

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

like social media data, location and other data from mobile and data from home sensors and 'wearables'. It is this capturing of 'ever-greater volume, velocity and variety of data' that, if harnessed effectively, provides the organisation with its Big Data opportunity.

34. **Data processing operations**. Although Big Data is growing exponentially and computer processing power and data storage tend over time to nil cost, nevertheless, as the '*Global Trends 2030*' report mentioned at Section A.2 above points out, there is a gap that must be bridged for Big Data to be harnessed effectively. This gap arises "between the amount of data that organizations can accumulate, and organizations' abilities to leverage those data in a way that is useful". In software applications terms, the gap is between traditional (retrospective) reporting and measurement BIA software and effective predictive forecasting and modelling (prospective) software techniques. This gap explains why 85% of the Fortune 500 are currently unable to leverage Big Data effectively, why the pace of growth in the $16bn BIA software market is so rapid, and why investment in business and artificial intelligence and analytics software tools and techniques is growing quickly. Processing operations are at the heart of Big Data – in addition to BIA, 'secret sauce algorithms', data visualisation and machine learning techniques will assist organisations in unlocking the 'unspoken secret of Big Data' – producing 'small effects' with 'large aggregated consequences', or turning pebbles into mountains in the words of the Chairman of Applied Predictive Technologies (see Section B.6).

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

36. **The 'pan-enterprise' view**. The picture presented here by this conceptualisation of the Big Data engine is of course over simplified: data input is rarely at the moment coordinated on an enterprise wide basis; processing operations are likely to be carried out at the desktop as well as at the on-premise or (public or private) Cloud server centre; and each department can have its own systems and IT requirements. There are also many ways in which an organisation can utilise Big Data: for example, it may obtain all or part of its Big Data and its BIA software as a service, on a bought in basis, rather than make the investment in capital and operating expenses itself, or it may carry out some of these activities in house and some externally. Equally, the organisation may choose to host its Big Data operations on-site, at a data centre (on a private cloud basis) or in the public cloud. As Big Data operations proliferate within the enterprise and within SMEs, the range of Big Data offerings and market places will expand. Nonetheless, looking at the Big Data engine holistically across the enterprise for input, processing and output operations remains one of the key objectives in order to harness Big Data most effectively, efficiently and compliantly.

E. MANAGEMENT AND GOVERNANCE OF BIG DATA PROJECTS

37. **Introduction**. The third view of Big Data – balancing legally compliant Big Data use with effective use of the organisation's Big Data assets – is superimposed on the first two, the Big Data legal analytical framework and the Big Data 'engine'. Here, the objective is a structured approach to managing Big Data projects with the aim of achieving legally compliant data use across the organisation in a technically enhanced and practical way that allows the business to gain maximum advantage from its data assets.

    Big Data governance does not arise in a vacuum. Large organisations will typically already have in place governance activities for all or part of their data activities, ranging from data protection and privacy governance frameworks (increasingly widely in place) to more detailed governance and management structures focused on Information Architecture, data accuracy, security, and regulatory compliance. However, the rise of Big Data and particularly the benefits arising from BIA software are fuelling a 'democratisation' of the benefits of Big Data utilisation, with many operational departments outside the CIO's group looking to use new BIA capabilities and features. A 'top down' approach to Big Data governance may result in a lack of responsiveness and flexibility, whilst a 'bottom up' approach driven by operational usage may be overly fragmented and not sufficiently address legal, regulatory and business risk in a way consistent with good governance.

38. **The Forrester Research IIG Report**. Data governance and management is therefore rising up the corporate agenda alongside Big Data itself. For example, in the August 2013 report commissioned by IBM entitled 'Big Data Needs Agile Information and Integration Governance'[59], consultants Forrester Research presented the results of an online survey conducted in summer 2013 of 512 respondents to evaluate their approaches, practices and perceptions around data governance. The key recommendations of the report centre around three guiding principles for what Forrester calls agile Information and Integration Governance (**IIG**):

    - developing Agile IIG incrementally in stages – focusing on 'quick return on insight' and matching the level of IIG with the level of analytical sophistication;
    - prioritising IIG around data types, data sources and data use through 'context-driven IIG zones' - focusing on data security, trust, validation and management efforts; and
    - incorporating and testing IIG - like any other aspect of a pilot Big Data project.

    Underlying these recommendations is the concept of four 'context-driven IIG data zones' - controlled (highly governed), casual (somewhat governed), validated (standardised) and chaos (no governance):

    > "The concept of zones is the foundation of agile IIG. Understanding the source of data as well as the type allows organizations to classify the data within the contexts of business use and value. Data may be ***tightly governed*** when used in business processes, decision-making, or meeting regulatory requirements. ***Casual governance*** may be present for data coming into the organization but not used frequently or widely. ***Validation*** can act to ensure a baseline of

[59] http://www-01.ibm.com/software/data/information-integration-governance/

conformity. And a ***chaotic state*** of governance may be allowed if data is not ready to be incorporated into business use. Less mature IIG tends toward a policy of controlled data for all data. However, with big data, organizations use these zones and apply varying degrees of governance to focus on what matters. Interestingly, no data type or source is left out of some type of applied IIG. Data available in big data initiatives all goes through some aspect of controlled, causal, or validated governance effort. Chaos is clearly not an acceptable state."[60]

Practical, incremental management can be built into a structured approach to Big Data governance projects based around four steps – risk assessment, strategy statement, policy statement, and process and procedures[61] - whose key content is shown in Figure 3 below, Towards a structured approach for managing Big Data projects.

Figure 3: Towards a structured approach for managing Big Data projects

step 1: risk assessment	step 2: strategy statement	step 3: policy statement	step 4: processes/ procedures

• structured process to review/assess/report/ remediate
• involve all the business
• establish all data types used & their sources
  • where does the data come from?
  • legal wrappers applying to all data – IPR, contract, regulatory
  • what consents were obtained/are needed?
• what processes do these data undergo?
• what does organisation use these data for?
	• start point
	  • risk assessment
	  • data protection policy
	  • information security policy
	• statement of company goals re Big Data
	• rationale, scope
	• governance, etc
	• list stakeholders
	  • buy in from top management & all parts of the business
	• CIO's group is key:
	  • information assets
	  • information architecture
	  • data modelling
	• Legal group is key
	  • IPR/contract/regulatory
		• people context:
		  • stakeholder groups
		  • interest & concerns
		• governance detail
		  • steering group
		  • working party
		  • compliance officer, etc.
		• project plan
		  • scope – data modelling
		  • responsibilities
		  • authority levels
		  • approval processes
		  • resources
		  • deliverables
		  • timelines, etc.
		• tools to be used
		  • IT/system measures
		  • processes/procedures
			• articulate proportionate processes & procedures to be followed
			• applies to all staff
			  • Tie in to HR policies, etc.
			• IT system/ measures & how they're to be used
			• awareness training, etc.
			  • initial
			  • refresher

39. **Step 1: risk assessment**. The first step or work stream in a Big Data management and governance project is the risk assessment as to how the business is currently using its data, carried out along the normal lines of review > assess > report > remediate.

[60] Page 6. Emphasis added.

[61] For a more detailed review of governance in a related area – Open Source Software – and points for consideration in strategy and policy statements and processes/procedures, see Kemp, '*Open source software (OSS) governance in the organisation*' (26 CLSR [3] pp. 309–316), or Practical Law at http://uk.practicallaw.com/3-501-0318?q=open+source+governance.

The review will consider the sorts of issues outlined at paragraph C.24 above and focus particularly on where data is sourced from, the terms under which it is supplied and how it is being used.

The next stage will assess whether use is consistent with contractual and licence terms, etc. and whether all necessary consents have been obtained (including where the data is personal data) for the uses carried out.

The review and assessment will be part of a report to senior management. The review will normally also include recommendations both by way of remediation plan to put right any areas of non-compliance that may have been identified in the assessment and that are forward looking to the strategy and policy aspects of data governance.

40. **Step 2: strategy statement**. The strategy statement is the high level articulation of the organisation's rationale, goals and governance for Big Data. It should prepared by an inclusive group consisting of senior management, the legal team, the CIO's team and all other stakeholders – which will in practice include many parts of the business. Identification and inclusion of all stakeholders, and articulating the prime objective of each in relation to Big Data and how that objective will be achieved, will be critical to successful Big Data governance and management.

    The Big Data strategy statement will need to align with high level corporate objectives and with strategy statements in related areas like data protection and privacy, information security and other aspects of data management, as well as intellectual property management. Organisations are therefore likely to be able to build on work already done in these areas to avoid reinventing the wheel.

    The Big Data strategy statement will need to align with the organisation's Information Architecture and its data methodologies, as well as with corporate policy on data acquisition, usage and compliance. It will also need to consider whether, and if so how, to follow the sorts of risk-based approaches to data zoning and incrementalism that are currently gaining traction in the area of data governance. There are therefore key roles in Big Data governance are for the CIO's (Chief Information Officer's) team and the legal team.

    As part of its focus on the 'people context' of Big Data governance, the strategy statement and work going towards it will generally settle the detail of establishing the institutional framework – the steering group, working party or task force, whether there will be a Big Data compliance officer (who may also be the current Data Protection compliance office for example).

41. **Step 3: policy statement**. The working party or task force will be responsible for the third work stream or step of preparing of the Big Data policy. Building on and implementing the strategy statement, the policy statement is essentially a project plan setting out scope, responsibilities, authority levels, approval processes, dependencies, deliverables and timelines for the project, as well as the IT/system aspects of the project.

    The working group/task force and policy statement are where the legal considerations around compliant Big Data use across the organisation and the technical considerations around the organisation's information architecture come together. Central to this work is data modelling and, at the policy level, how the IA will implement the organisation's policy choices about Big Data use.

42. **Step 4: processes and procedures**. The policy statement will drill down to the level of the fourth step or work stream, the detailed processes and procedures around project methodology and the data modelling to be used. Here, more precise processes and procedures will be developed in the context of the data model used in the IA to decide how the data entities are to be tagged for any type of data the organisation uses or may want to use.

    The processes and procedures will also tie into the organisation's HR policies and provide for awareness training – the key 'do's' and 'don'ts' of compliant Big Data usage.

F. CONCLUSION

43. **Conclusion**. As gaining unique competitive insight from Big Data becomes an increasingly important strategic goal of larger businesses, the effort and resources applied to Big Data projects are set to grow significantly over the next few years. A sound analytical legal model for understanding the rights and duties that arise in relation to Big Data in order to manage risk, and the development of a structured approach to legally compliant Big Data input, processing and output will be essential for successful Big Data projects and their governance and management.

Richard Kemp,
Kemp IT Law,
London,
September 2014
richard.kemp@kempitlaw.com
Tel: 020 3011 1670