Available online at www.sciencedirect.com

ScienceDirect

www.compseconline.com/publications/prodclaw.htm

Computer Law & Security Review

Legal aspects of managing Big Data

Richard Kemp[*]

Kemp IT Law, London, UK

Keywords:
Big data
Information architecture
Data protection
Intellectual property rights
Contract law
Data governance

ABSTRACT

Big Data is shorthand for the currently rapidly evolving techniques of gathering and analysing for competitive advantage vast unstructured and structured sets of digital data. Big Data is currently at an early stage of development, but many organisations will be embarking on Big Data projects in the next couple of years in order to be in a position to know more about their customers than their competitors. Central to the success of these projects will be four critical factors: (i) understanding the legal framework for Big Data and how it applies to the organisation concerned; (ii) effectively bringing together the organisation's IT and legal functions in the Big Data project; (iii) a clear understanding of the organisation's objectives for its Big Data operations; and (iv) a structured approach to the strategy, policy and process aspects of Big Data governance.

1. Introduction

1.1. *'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 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 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 $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 Big Data for competitive advantage in a structured, legally compliant way.

It is this race for competitive advantage – knowing more than your competitor not so much about what your customer has just done as about what he or she is likely to do next – that is at the commercial epicentre of Big Data. Yet this is a race that is just beginning: Gartner also points out[3] that only 15% of Fortune 500 companies will be able to exploit Big Data

[*] Kemp IT Law, 21 Napier Avenue, London SW6 3PS.
E-mail address: richard.kemp@kempitlaw.com.
URL: http://www.kempitlaw.com

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

http://dx.doi.org/10.1016/j.clsr.2014.07.006
0267-3649/© 2014 Kemp IT Law. Published by Elsevier Ltd. All rights reserved.

for competitive advantage by the end of next year and that only 8% of companies are currently using Big Data analytics at all.

1.2. What is 'Big Data'?

Commenting that there was no one generally accepted definition of Big Data, the White House's Executive Office of the President in a report dated 1 May 2014[4] 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 that previously were not available."*[5] *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".*[6]

As used in this White Paper, 'Big Data' is therefore shorthand for the collation, processing, analysis and use of vast exploitable datasets of unstructured and structured digital information. Along with Cloud, mobile[7] and social computing, Big Data 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.

1.3. 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's '*Global Trends 2030: Alternative Worlds*'[8] 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.*[9]

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.*[10]

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

1.4. 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 2 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 3 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 4 considers Big Data within the organisation from the standpoint of input, processing and output operations; and
- Section 5 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 2 and 3 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

[4] '*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'.

[5] 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).

[6] 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.

[7] See Kemp, 'Mobile payments: current and emerging regulatory and contracting issues' (29 CLSR [2], pp. 175–179).

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

[9] At page ix.

[10] At page 85.

references are provided[11] to further materials where these aspects are discussed at greater length.

2. The business context: Big Data in key vertical sectors

2.1. *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.[12] 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 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.

2.2. *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[13]), which operates an open standards based enterprise IA framework, and BIAN (the Banking Industry Architecture Network[14]), which operates a banking specific standard IA based on SOA,[15] have announced[16] 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.

2.3. *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 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 motor car 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. Similarly, smart domestic sensors can help improve

[11] 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). 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).

[12] 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.

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

[14] 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.

[15] 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.

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

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 motor car, 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 in the insurance industry is Solvency II[17] 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.

2.4. *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 have generated 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 are likely to emerge to support these trends.[18] 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.

2.5. *The recorded music industry*

The recorded music industry is a $15bn global business in the course of transformation through digitisation as developing patterns of online consumption through streaming downloads and the like 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 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.

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.

2.6. *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*’,[19] 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.

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

[17] 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.

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

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

GDP), a ‘relentless’ producer of Big Data in the words of a report in the Guardian newspaper[20].

2.7. *The public sector*

In common with all developed countries, HMG's database about its citizens is the largest in the UK, and of large and increasing value. Government departments like Health, Home Office, Education, HMRC and BIS have huge and growing digital databases, and were it not for the run of high profile public sector IT missteps over the past decade, the hue and cry from civil liberties groups about the risks to individual freedoms of ‘citizen on a stick’ – everything the State knows about any citizen on a memory stick – would have been much louder by now. As it is, as individual government departments now master their own data estates and central government as a whole starts to join up the dots on what each department knows about any particular individual, HMG's data estate – a term we will become much more familiar with in years to come – may well become one of the UK's most valuable national assets.

Looked at as an asset in this way, 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, safeguards against State overreaching, commercial interests, and maximising the benefits of technological progress for citizens.

3. Towards a common legal framework for Big Data

3.1. *Introduction: what is data in legal terms?*

A reasonable start point for the discussion about the legal framework for Big Data is to ask a fairly fundamental question: 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 (information and data are used interchangeably in this paper). 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 to us because of investment in IT, just as music, books and films require 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 vs Moss[21] 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[22] 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 a matter of law, developing rapidly at the moment. They are likely to develop even more quickly as Big Data techniques become more prevalent.

These rights and duties arise through intellectual property rights (‘**IPR**’), contract and regulation. They are important as (positively) 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).[23] These developments mean that ‘data law’ is emerging as a new area in its own right around IPR, contract and regulation.

3.2. *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 (see Fig. 1, towards a common legal framework for Big Data).

3.3. *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 accessing and extracting the data held in that software.[24]

3.4. *Level 2: information architecture*

The information architecture (IA) is the intermediate step between the platform infrastructure and the data itself and, as observed at Section 2.2 above, sits at the centre of networked

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

[21] [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.

[22] Your Response Ltd vs 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 there is no property in a database, just that there is no personal tangible property.

[23] 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).

[24] See for example Navitaire Inc vs 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 footnotes 11 and 23 above.

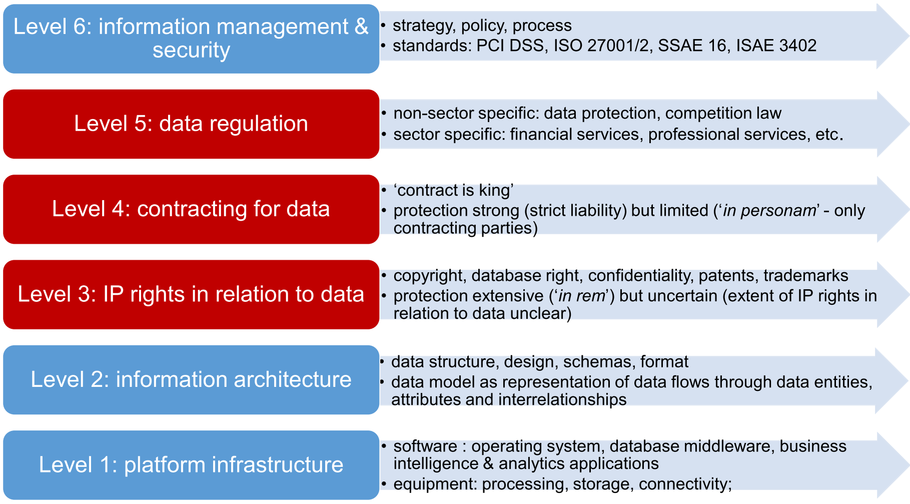

Level 6: information management & security
• strategy, policy, process
• standards: PCI DSS, ISO 27001/2, SSAE 16, ISAE 3402
Level 5: data regulation
• non-sector specific: data protection, competition law
• sector specific: financial services, professional services, etc.
Level 4: contracting for data
• 'contract is king'
• protection strong (strict liability) but limited ('*in personam*' - only contracting parties)
Level 3: IP rights in relation to data
• copyright, database right, confidentiality, patents, trademarks
• protection extensive ('*in rem*') but uncertain (extent of IP rights in relation to data unclear)
Level 2: information architecture
• data structure, design, schemas, format
• data model as representation of data flows through data entities, attributes and interrelationships
Level 1: platform infrastructure
• software : operating system, database middleware, business intelligence & analytics applications
• equipment: processing, storage, connectivity;

Fig. 1 – Towards a common legal framework for Big Data.

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 copyright protection in the normal way; and the database 'schema' or formal structure (as distinct from the data content of a database) is protectible by copyright in the EU under Chapter II, Article 3 of the Database Directive.[25] 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 SSO (Standards Setting Organisation), TC (Technical Committee) or individual organisation that manages the standard concerned.

3.5. *Level 3: intellectual property rights in relation to data*

The main IP rights in relation to data are copyright, database right and confidentiality. 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, but again, generally not in relation to the actual data.

Copyright protects the form or expression of information and not the underlying information itself. It arises by operation of law in the EU (i.e. it does not require to be registered) and is a formal remedy that does what it says on the tin and stops unauthorised copying (or the unauthorised carrying out of other acts that copyright, a bundle of rights in this respect, protects). It applies to software, certain databases, literary works, music, films, videos and broadcasts. A successful claim for copyright infringement will need to show (i) that copyright subsists in the work – generally, that it is original, where the UK standard is low and normally that the work concerned has not been copied from elsewhere; (ii) that the claimant owned or could otherwise sue on that copyright; (iii) that the work was within copyright (life plus seventy years in the case of software, databases and other literary works); and (iv) that the copyright had been infringed – for example, a qualitatively substantial part of the work had been reproduced without authorisation and where a copyright permitted act exception did not apply. In the UK, database copyright is subtly different from copyright in software and other written work as the standard for originality is higher – the selection and arrangement of the database's contents must have been the author's 'own intellectual creation' and not just original.

Database right sits alongside but is separate from copyright. It arises in a database (essentially for this purpose, a searchable collection of independent works) in whose 'obtaining, verifying or presentation' the maker has made a substantial investment. The right lasts for fifteen 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 on a one-off basis, or repeatedly and systematically of insubstantial parts. At best an 'eggshell' right giving a thin layer of protection, a number of judgments of the European Court of Justice in Luxembourg have drawn its teeth as a powerful right, at least for the moment.[26]

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. Protection can extend to aggregation of datasets even where parts of the data are in the public domain and so not otherwise confidential, and to second generation data derived from the initial confidential data.

IPR in relation to data is of uncertain scope at the moment, and the law in this area is likely to continue to develop in the

[25] 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.

[26] See Kemp et al, '*Database right after BHB vs William Hill: enact and repent at leisure*' (22 CLSR [6], pp 493–498).

coming years: historically, IPR development has followed the commercialising of innovation and as the value of Big Data rises, so likely will the IP rights underpinning it. Whilst of uncertain scope, as rights '*in rem*' – enforceable against the whole world, not depending on a pre-existing relationship – they are extensive, with powerful infringement remedies, from temporary and permanent injunctions (court orders requiring the infringement to terminate whose breach are sanctioned through contempt of court) to damages and account of profits.

3.6. *Level 4: contracting for data*

The converse of IPR is true for contractual rights in relation to data. Contract law confers strong, enforceable rights and imposes strong, enforceable obligations since liability is strict: liability will arise if breach is proved as a question of fact on a balance of probabilities. Contract rights in relation to data are technically entirely separate from IPR and 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.*[27]

If the good news is that data contracts are strong, the bad 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 duties under a contractual wrapper, so contract IPR and IPR 'proper' also need to be considered separately). It is however the strength of contract law that underpins durable ecosystems like market data referred to at Section 2.1 above.

3.7. *Level 5: data regulation*

The third legal area of increasing importance for Big Data is regulation, where the law is derived from statute. Data protection – conferring rights and imposing obligations on the processing of personal data – is the most important and 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 1.2 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.

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.

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 2 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 the carrying out of their regulatory functions. These requirements are tending to become more intrusive as regulatory authorities obtain wider powers to obtain information, investigate business practices and conduct and audit organisations under their charge.

3.8. *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, is another area where work can 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 operate a range of Data Security Standards (DSS). More generically, the International Standards Organisation (ISO) has published the 27,000 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 16 and ISAE 3402).

3.9. *Legal rights in relation to Big Data – a complex picture*

Legal rights in relation to Big Data present 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 on data concurrently: a particular dataset – say PNR (passenger name record) data from the ATI – can be subject to IPR as database right or copyright (in the system of the airline); contractual rights and duties (between the airline and a travel agent); and data protection regulation as personal data relating to the passenger.

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

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 at each stage as incremental investment is made and value added. (In the case of market data in the financial sector, these processes happen at great speed, so that the evidential burden in formal dispute resolution in showing what happened when can be time consuming and costly, and this is perhaps another reason why there has been little litigation in this area). A bank subject to regulatory information security and audit scrutiny may seek contractually to impose those requirements in turn on its IT vendors so as to ensure it is not beholden to its regulator without being able to enforce what it would see as back to back compliance from its own suppliers.

Fourth, IPR rule sets operate differently in different countries as they are national rights conferred by national law and enforceable (primarily and initially) in national courts. These differences vary from the relatively 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 more major (for example, database right is ‘made in Europe’ and does not apply to databases made in the USA, and some countries operate a registration requirement for copyright, whilst in others copyright arises by operation of law without the need for 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. (This is one reason for the EU's recent preference for Regulations, which are directly applicable in national Member State law without the need for transposition).

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

4. Big data operations inside the organisation

4.1. *Introduction*

The 6 level stack at Section 3 provides this paper's first ‘view’ of Big Data as a common legal analytical framework. Section 4 overlays on to that view the organisation's Big Data operations – the input into, processing within and output from, the Big Data ‘engine’ (see Fig. 2 – The Big Data engine – input, processing and output operations).

4.2. *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, in the quote from the White House EOP report Section 1.2 above, it consists of “large, diverse, complex, longitudinal, and/or distributed datasets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources”. 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.

4.3. *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 1.3 above points out, there is a large gap to be bridged before Big Data can be harnessed effectively. This gap arises “between the amount of data that organizations can accumulate and organizations' abilities to leverage those data in a

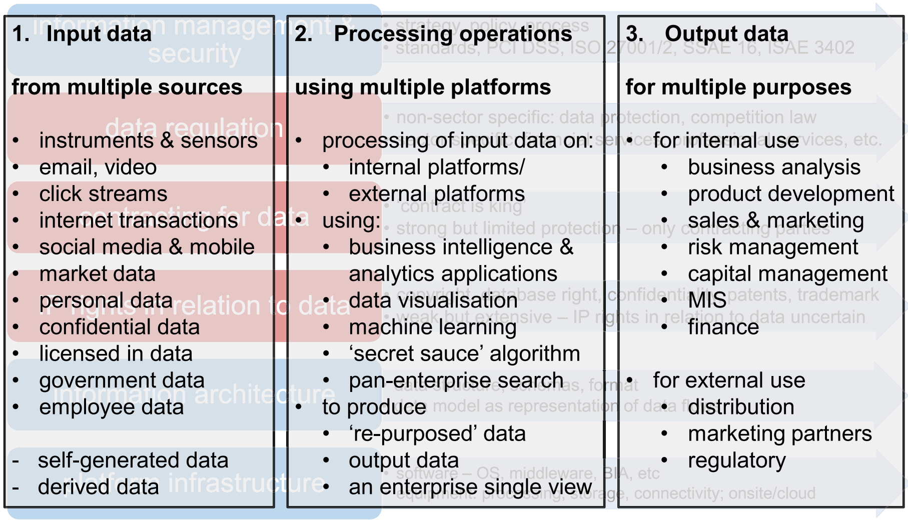
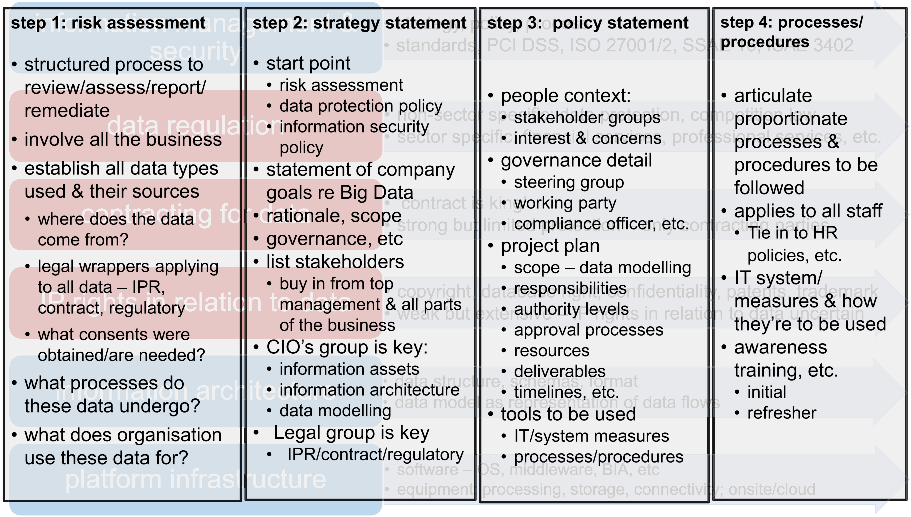


Fig. 2 – The Big Data engine – input, processing and output operations.


security
data regulation
platform infrastructure
• software – OS, middleware, BIA, etc
• equipment processing, storage, connectivity; onsite/cloud
• data model as representation of data flows

step 1: risk assessment

- 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?

step 2: strategy statement

- 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

step 3: policy statement

- 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

step 4: 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

Fig. 3 – Towards a structured approach for managing Big Data projects.

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. These gaps explain why currently 85% of the Fortune 500 is 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.

4.4. *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 and external to the organisation where it can be most effectively used. This 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, by product development to consider new product offerings and for marketing purposes.

4.5. *The 'pan-enterprise' view*

In practice, the picture presented by this conceptualisation of the Big Data engine is again much more complex: 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-premises or Cloud server centre, and each department can have its own systems and IT requirements. Nonetheless, looking at the Big Data engine holistically across the enterprise for all input, processing and output operations remains one of the key objectives in order to harness Big Data most effectively, efficiently and compliantly.

5. Management and governance of Big Data projects

5.1. *Introduction*

The third view of Big Data from the legal perspective – balancing legally compliant Big Data use and 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 way that allows the business to gain maximum advantage from its data assets. As Fig. 3 shows, the governance structure is based on four steps – risk assessment, strategy statement, policy statement, and process and procedures.[28]

5.2. *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 along the normal lines of review > assess > report > remediate:

- ***reviewing:***
  - ○ where all data is sourced from;
  - ○ (where the data is structured) the terms under which it is supplied, how it is being used and whether all use is consistent with contractual and licence terms, etc.; and
  - ○ (where the data is unstructured) whether all necessary consents have been obtained (including but not limited

[28] 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).

to where the data is personal data) for all uses carried out;
- ***assessing*** whether and if so where and how the business is acting outside the scope of any contract or licence terms or non-compliantly with IP or regulatory duties for any of that data;
- ***reporting*** to senior management; and
- preparing a ***remediation plan*** to put right any areas of non-compliance that have been identified in the assessment.

The key roles here are for the legal team and the CIO's (Chief Information Officer's) team. Here, organisations can take as a possible template for the Big Data risk assessment the work they have already done in the data protection and information security areas.

5.3. *Step 2: strategy statement*

In parallel with the risk assessment the second step is the formulation of the company's Big Data strategy. The strategy statement is a high level articulation of the organisation's rationale, goals and governance for Big Data prepared by an inclusive group consisting of senior management, the legal team, the CIO's team and all other stakeholders. Identification and inclusion of all stakeholders, and articulating the prime objective of each in relation to Big Data and how that objective will be achieved, will be critical to successful Big Data governance and management.

5.4. *Step 3: policy statement*

The strategy statement will generally name a steering group who will be responsible for the third work stream or step, preparation of the Big Data policy statement. This is essentially a project plan setting out scope, responsibilities, dependencies, deliverables and timelines for the project and the tools that it will be using.

The policy group and statement is 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.

5.5. *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 include awareness training – the key ‘do's’ and ‘don'ts’ of compliant data usage.

6. 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 will 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 and software enhanced Big Data input, processing and output will be essential factors for successful Big Data projects and their governance and management.