Big Data and Data Protection

Richard Kemp

November 2014

KEMP IT LAW

kempitlaw.com

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

BIG DATA AND DATA PROTECTION

TABLE OF CONTENTS

BIG DATA AND DATA PROTECTION

A. CONTEXT

1. **Introduction**. Big Data – the harnessing, processing and analysis of digital data in huge and ever increasing volume, variety and velocity – has quickly risen up the corporate agenda over the last twelve months as organisations appreciate that they can gain advantage through valuable insights about their customers and users through the techniques that are currently rapidly developing in the Big Data world.

   Much Big Data, for example, climate and weather data, is not personal data – data that relates to an identifiable living individual. But for Big Data that is or could be personal data, the two plates of data protection law and Big Data analytics collide in ways that the drafters of the EU Data Protection Directive 95/46[1] (**Data Protection Directive**) and even the draft EU General Data Protection Regulation[2] of January 2012 (**draft Data Protection Regulation**) could barely have foreseen.

   The response of data protection authorities, not surprisingly, is that Big Data is no exception to any other form of data processing that falls within the scope of data protection law and so must comply with the principles and detail of the applicable rules. This in itself is the source of some tension at the moment, and this tension is likely to become more pronounced as Big Data techniques develop and use becomes ubiquitous at the same time as data protection law itself is developing quickly, particularly at the moment as the draft Data Protection Regulation continues its tortuous progress towards the statute book.

2. **Scope and aims of this white paper**. The main purpose of this paper is to provide a practical overview of the data protection law issues that arise in the world of Big Data and to provide practical pointers to how they can be addressed. This Section sets the context by exploring a number of fundamental questions - what are information, data and Big Data? What is data in legal terms? – and offering a common legal analytical framework for data law, with data protection as a central element of that framework. Section B seeks to put some flesh on the bones by briefly overviewing data protection aspects of Big Data initiatives in a number of different vertical sectors (healthcare, insurance, healthcare, air travel and public sector). Section C give a broader policy perspective by looking at recent developments around Big Data and data protection on both sides of the Atlantic. Section D considers the main data protection issues that arise in the Big Data world. Section E provides points towards how those issues may be addressed.

---

[1] Directive 95/46 of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and the free movement of such data, OJ L281 - http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML

[2] Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and the free movement of such data (General Data Protection Regulation), COM(2012)11 Final and associated draft Directive, COM(2012) 10 final, 25 January 2012, http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm

For an analysis of the legal aspects of Big Data and Big Data management more generally (which reviews the intellectual property and contractual aspects and considers how Big Data is used and can be managed within the organisation), please see our White Paper at http://www.kempitlaw.com/category/white-papers/.

3. **What is data?** A reasonable start point for any discussion about the legal and regulatory 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[3]:

   > "**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".

4. **What is Big Data?** Two descriptions of Big Data from the policy papers referred to at Section C below are helpful in understanding the attributes of Big Data, effectively next generation data mining techniques using more data, faster processing and new software First, the White House Report referred to at paragraph B.15 below:

   > "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."[4] 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."[5]

   This is complemented in the Commission Communication mentioned at paragraph B.16:

   > "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".

---

[3] 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.

[4] 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).

[5] 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.

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.

5. **What is data in legal terms?** 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 and use 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 or obligations ***in*** data but that extensive rights and obligations arise ***in relation to*** data. The UK criminal law case of Oxford v Moss[6] is generally taken as authority for the proposition that there is no property ***in*** data as it cannot be stolen. 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 will 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, 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)[7]. Current

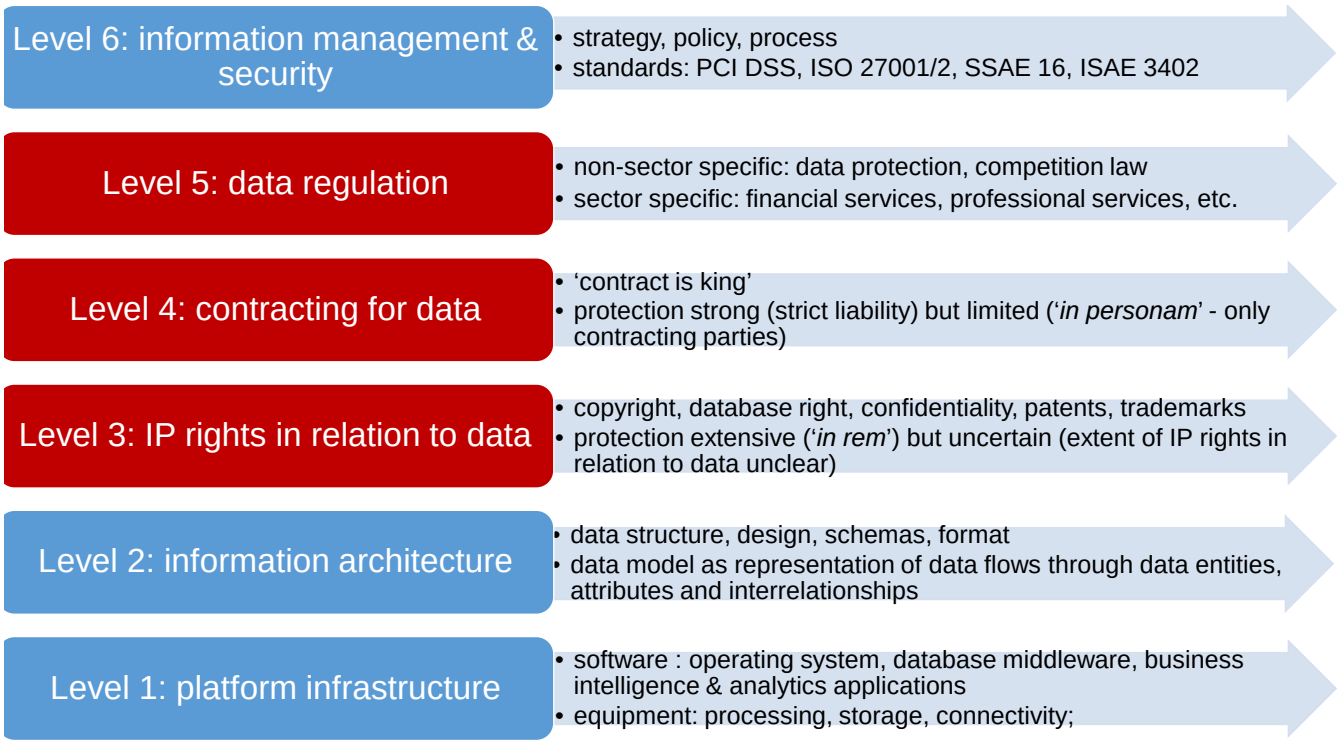
[6] [1979] Crim LR 119, where it was held that confidential information in an exam question was not 'intangible property' within the meaning of Section 4(1) of the Theft Act 1968 and so could not be stolen

[7] 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).

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.

6. **Towards a common legal analytical framework for data**. 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, Towards a common legal framework for Big Data). Data protection is the most important component of level 5, data regulation.

Figure: Towards a common legal framework for Big Data



B. THE BUSINESS CONTEXT: BIG DATA AND DATA PROTECTION IN VERTICAL SECTORS

7. **Introduction: pebbles and mountains**. This section provides an overview of current developments in a number of different vertical sectors (healthcare, insurance, air travel and public sector) by way of examples of data protection issues in Big Data. 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 has commented[8] 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

[8] 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

Data analytics developer Applied Predictive Technologies was quoted in the Economist piece as saying that Big Data was "about building a mountain with pebbles".

8. **The healthcare sector**. Healthcare is the sector both where adoption and use of Big Data is likely to have the greatest impact on people's daily lives and where information is about identifiable living individuals. In its January 2013 report '*The 'big data' revolution in healthcare*'[9], 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 GDP), a 'relentless' producer of Big Data in the words of a report in the Guardian newspaper[10].

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. 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

---

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

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

insurance. They will over time have a material impact on the pricing of vehicle, home and health policies.

Big Data in insurance especially points up 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.

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 personal data about passengers' 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[11]. 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.

    Passenger data emerges as both a key enabler and regulatory issue for Big Data in the ATI, which is seeing the development of a number of policy initiatives around the standardisation of passenger consent to use of their personal information in and around the airport and their journey more generally. Passenger data is a key enabler through increasing use of the mobile phone as data source, data store and processing point. For the airline passenger, 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. In addition, use of this data segues to m-commerce and shopping at the airport and beyond to smart cities.

    From the regulatory point of view, however, the use passenger name record (**PNR**) data to combat terrorism and serious crime has had a chequered history since the agreement between the European Commission and the USA for the transfer of passenger data to the USA was struck down by the European Court of Justice in May 2006 and the 2011 draft PNR Directive[12] has moved slowly through the EU institutions.

11. **The public sector**. As with 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.

---

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

[12]

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 saw the issue of data sharing within government rise up the agenda with increasing press interest[13] around policy developments following publication by the Cabinet Office Data Sharing Policy Team on 9 April 2014 of their Initial Discussion Document[14]. 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 their privacy. Ideas put forward in the discussion document include developing the December 2012 proposals of the Administrative Data Taskforce[15] 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[16], HMG's vehicle for public sector Big Data; and
- compliance with the Data Sharing[17] and Anonymisation[18] Codes of Practice published by the Information Commissioner's Office (**ICO**), the UK data protection regulator.

C. BIG DATA AND DATA PROTECTION: RECENT DEVELOPMENTS

12. **The Draft European General Data Protection Regulation**[19]. The draft Data Protection Regulation is outside the scope of this White Paper, but currently, in the last quarter of 2014, the vocal and extensive debate that is still taking place about the detail and underlying policy objectives of the draft Regulation overhangs any discussion about Big Data. Although the draft Data Protection Regulation

---

[13] 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

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

[15] 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.

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

[17] http://ico.org.uk/for_organisations/data_protection/topic_guides/data_sharing.

[18] http://ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation. See below

[19] Citation at footnote 2 above.

does not mention Big Data by name, nonetheless it contains many provisions that would affect the use of personal data in Big Data, including many of the points that are discussed at Section D below. Paragraphs 117 to 126 on pages 39 to 42 of ICO's July 2014 Report (see paragraph C.17 below) also comments on the Big Data aspects of the draft Data Protection Regulation.

13. **The US National Intelligence Council's December 2012 Report**[20]. 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*'[21] 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"[22]

and points up the continuing tradeoffs that individuals and organisations will need to make between utility and privacy of personal data.

14. **HMG's October 2013 Strategy Paper**[23]. HMG's paper presents a positive view of the UK's ability to seize the data opportunity and the government's determination to position the UK to make the most of what it calls the "data revolution", noting in the foreword that Big Data's potential impact "is so significant that it could transform every business sector and every scientific discipline". The section on privacy and data protection at the end of the paper states the UK's aim as:

> "a clear and pragmatic policy on data privacy and confidentiality which ensures public trust in the confidentiality of their data, while increasing the availability of data to maximise its economic and social value."

The paper does however give an indication of the conflicting views around the draft Data Protection Regulation and is much less guarded than (for example) ICO in its publicly expressed views about the need to avoid over-prescriptive rules:

> "The UK supports the need to bring data protection rules in line with the reality of the 21st century. The government does not believe that the European Commission's proposals for reform to the data protection framework strike the right balance between privacy and innovation. We should also be careful about overly prescriptive regulation that increases red tape and costs for businesses, the public sector, and for regulators themselves."

15. **The Executive Office of the US President's May 2014 Big Data Report**[24]. The White House's report from May 2014 focuses on "how big data will transform the way we live and work and alter the

---

[20] 'Global Trends 2030: Alternative Worlds', http://globaltrends2030.files.wordpress.com/2012/11/global-trends-2030-november2012.pdf

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

[22] At page ix.

[23] 'Seizing the data opportunity – A strategy for UK data capability', 30 October 2013, https://www.gov.uk/government/publications/uk-data-capability-strategy

[24] 'Big Data: Seizing Opportunities, Preserving Value', May 1, 2014, http://www.whitehouse.gov/issues/technology/big-data-review

relationships between government, citizens, business and consumers". It contains a useful section (pages 15 to 22) on the evolution of US privacy laws and how they fit into international privacy frameworks, before considering Big Data and privacy in the public sector (pages 32 to 38) and private sector (pages 43 to 47). The US approach since the 1970s has been characterised by 'notice and consent' – focusing on obtaining user permission prior to collecting data – and the paper concludes on Big Data and privacy with the realistic assessment that:

> "trends [in Big Data] may require us to look closely at the notice and consent framework that has been a central pillar of how privacy practices have been organized for more than four decades. In a technological context of structural over-collection, in which re-identification is becoming more powerful than de-identification, focusing on controlling the collection and retention of personal data, while important, may no longer be sufficient to protect personal privacy. In the words of the President's Council of Advisors for Science & Technology, "The notice and consent is defeated by exactly the positive benefits that big data enables: new, non-obvious, unexpectedly powerful uses of data."

16. **The European Commission's July 2014 Communication**[25]. The European Commission (**Commission**) Communication sets out a number of activities it considers necessary for the EU "to be able to seize [Big Data] opportunities and compete globally in the data economy" including:

> "[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; [and]
>
> [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)."

17. **The UK Information Commissioner's Office's July 2014 Report**[26]. ICO's Big Data and Data Protection paper of 28 July 2014 (the **ICO Paper**) added to the growing literature on the topic. In applying the relevant principles of the Data Protection Act (**DPA**) to the different aspects of Big Data and providing useful practical pointers on how to address them, it is the most detailed consideration to date of these issues. The ICO Paper is considered further in Sections D and E below.

18. **The Article 29 Working Party's September 2014 Statement**[27]. At the time of writing, one of the most recent policy pieces is the three page statement of the EU Article 29 Working Party of 16 September 2014, two paragraphs of which are particularly worth calling out:

---

[25] 'Towards A Thriving Data-Driven Economy', COM(2014) 442 Final, 2 July 2014, https://ec.europa.eu/digital-agenda/en/news/communication-data-driven-economy

[26] 'Big data and data protection', 28 July 2014, http://ico.org.uk/for_organisations/data_protection/topic_guides/big_data

[27] 'Statement of the Article 29 Working Party on the impact of the development of big data on the protection of individuals with regard to the processing of their personal data in the EU, 16 September 2014', http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/index_en.htm. The Article 29 Working Party also considered big data at Annex 2 (pages 45 to 48) of its Opinion 03/2013 of 2 April 2013 on purpose limitation (see previous hyperlink).

> "- The Working Party acknowledges that the challenges of big data might require innovative thinking on how some of the key data protection principles are applied in practice. However, at this stage, it has no reason to believe that the EU data protection principles, as they are currently enshrined in Directive 95/46/EC, are no longer valid and appropriate for the development of big data, subject to further improvements to make them more effective in practice. It also needs to be clear that the rules and principles are applicable to all processing operations, starting with collection in order to ensure a high level of data protection.
>
> - In fact, the Working Party strongly believes that complying with this framework is a key element in creating and keeping the trust which any stakeholder needs in order to develop a stable business model that is based on the processing of such data. It also believes that compliance with this framework and investment in privacy-friendly solutions is essential to ensure fair and effective competition between economic players on the relevant markets. In particular, upholding the purpose limitation principle is essential to ensure that companies which have built monopolies or dominant positions before the development of big data technologies hold no undue advantage over newcomers to these markets."

D. BIG DATA AND DATA PROTECTION: THE MAIN ISSUES

19. **Introduction**. This section addresses by reference to the DPA the main substantive data protection law issues that arise. These issues, each considered in the ICO Paper, tend to track the DPA's data protection principles, particularly fairness (principle 1), consent (Schedule 2, paragraph 1), purpose limitation/repurposing (principle 2), data minimisation (principles 3 and 5), overseas transfer (principle 8) and proving harm and liability for loss. These issues are each now considered. As we will see, they give rise in the data protection/Big Data debate to a number of terms loaded with meaning, like 'unexpected', 'intrinsic', 'compatible' and 'N ≠ all'.

20. **Issue 1 – fairness (data protection principle 1)**. Under data protection principle 1, personal data must be processed fairly and lawfully in circumstances where one of the conditions in DPA Schedule 2 is met. Schedule 1 Part II sets out how this principle is to be interpreted, broadly by looking at how the personal data was obtained; whether the provider was misled or deceived as to the purpose of the processing; and whether the data processor provided to the data subject at the right time details of its identity and the processing purpose and any "further information which is necessary, having regard to the specific circumstances in which the data are or are to be processed, to enable processing in respect of the data subject to be fair".

    The ICO Paper points up (paragraphs 47 to 49) the importance of fairness, transparency and meeting reasonable expectations in Big Data processing:

    > "Fairness is partly about how personal data is obtained. The 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. This means that transparency about how the data will be used is an important element in assessing whether big data analytics comply with the data protection principles. …
    >
    > It is also necessary to consider the effect of the processing on the individuals concerned. … Fairness involves a wider assessment of whether the processing is within the reasonable expectations of the individuals concerned."

    The central rationale of Big Data is to discover hitherto unobserved correlations between different datasets and to provide actionable insights from these (by definition) 'unexpected' results.

Reconciling the broad concept of fairness in data protection terms with this central feature of Big Data will lie at the heart of an organisation's data protection compliant Big Data use.

The ICO Paper illustrates how unexpected Big Data correlations may be by quoting the well-known example of US retailer Target whose marketing department's analytics found a pattern between the purchase dates of certain products by expectant women and their due date. When Target sent marketing literature for baby-related products to the daughter, who was still in high school, of a Minneapolis father as fitting this pattern, the father complained to Target about her receiving this kind of marketing. The father was not then aware of his daughter's pregnancy and subsequently apologised to the store.

Less graphically, ICO states that issues may arise where personal data obtained in relation to providing a particular service is then used for a purpose that does not necessarily fit with the original purposes or that is not 'intrinsic' to that provision of that service. It continues (paragraph 50) that use for market research by a retailer of loyalty card data would be on the right side of the line but by a social medial provider would not. This concept of where 'intrinsicness' starts and ends is likely to develop with Big Data, and we will hear a lot more about the nexus between the purposes that are 'intrinsic' and those that are 'non-intrinsic' to particular services where personal data has been obtained.

21. **Issue 2 – consent (DPA Schedule 2, paragraph 1)**. By DPA Schedule 2, paragraph 1, the processing of personal data with the consent of the data subject meets the condition of data protection principle 1 for fair processing. The bar for consent has risen markedly in recent years, and how high it is to be set is a key point in the deliberations for the draft Data Protection Regulation, but for present purposes the formula is that consent must be freely given, specific and informed.

The first point is that just because Big Data is complex is not an excuse for not getting it where it is required:

> "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)

Where an organisation is relying on consent in the Big Data context means that:

> "people must be able to understand what the organisation is going to do with their data and there must be a clear indication that they consent to it. If an organisation has collected personal data for one purpose and then decides to start analysing it for completely different purposes (or to make it available for others to do so) then it needs to make its users aware of this. This is particularly important if the organization is planning to use the data for a purpose that is not apparent to the individual because it is not obviously connected with their use of a service. For example, if a social media company were selling on the wealth of personal data of its users to another company for other purposes" (paragraph 56).

But consent must be appropriate as well as reasonable:

> "It may be reasonable for organisations to use consent as a condition for processing in a big data context but they have to be sure that it is the appropriate condition. Furthermore, if people do not have a real choice and are not able to withdraw their consent if they wish, then the consent would not meet the standard required by the DPA."

How appropriate – or 'real' - is the consent that may be given for Big Data processing is another key issue alongside the 'unexpectedness' of Big Data insights and the nexus between intrinsic and non-intrinsic purposes.

22. **Issue 3 – purpose limitation/repurposing (data protection principle 2)**. Data protection principle 2 is a two-step approach that personal data must be obtained for specified, explicit and legitimate purposes (the original purpose) and then must not be further processed for any other purpose (the new purpose) that is incompatible with the original purpose.

    The ICO Paper puts the repurposing issue into context by saying that principle 2 does not bar the new purpose, or say that the new purpose and original purpose must be the same, but that the new and original purposes must not be incompatible. It then emphasises fairness as a yardstick for determining compatibility:

    > "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).

23. **Issue 4 – data minimisation (data protection principles 3 and 5)**. The DPA encompasses the principle of data minimisation through the combination of data protection principles 3 and 5: personal data must be 'adequate, relevant and not excessive' (principle 3) and not 'kept for longer than is necessary' (principle 5) in relation to the purposes for which they are processed. Big Data on the other hand 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 the outset why they need to collect and process particular datasets" (paragraph 73).

    As data storage costs reduce, and the costs of destroying data outweigh those of keeping it, a similar point arises in relation to the length of time for which personal data are kept in the world of Big Data where the bigger the datasets analysed the better.

24. **Issue 5 – overseas transfers (data protection principle 8)**. Big Data adds further complexity to the data protection aspects of international transfers of personal data, supplementing current issues around the US/EU Safe Harbor programme, the EU model clauses and Cloud computing. Big Data does not necessarily add to them but rather represents a further, international and non EU dimension to substantive fairness, consent, repurposing and minimisation issues discussed above. This is probably why the ICO Paper (at paragraph 96) defers to its guidance on Cloud Computing which it states "covers the international transfer issues raise by" Big Data technology.

25. **Issue 6 – proving harm and liability for loss**. 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 questions of compensation and liability that may arise, as a result of using non-DPA compliant Big Data analytics. As Big Data is all about 'building mountains from pebbles' – small, fast, incremental changes – one view would be that any single Big Data DPA breach may not be material. Such a view would clearly undermine the substantive application of the DPA to Big Data if its effect were to be that remedies were restricted. On the other hand, the trend in data protection law is for the rights of data subjects and others to expand. Individual rights introduced or extended by the DPD and DPA (for example the rights at DPA sections 7 to 14 in relation to subject access, prevention of processing, compensation and rectification, blocking and destruction of personal data) will be expanded further in the Data Protection Regulation. Similarly, regulators' powers are increasing also, and will be extended further by the Regulation. It is to be hoped that ICO will provide guidance as to how it see the DPA liability regime operating in the Big Data world.

E. ADDRESSING DATA PROTECTION ISSUES IN BIG DATA: PRACTICAL POINTERS

26. **Getting the right consent to the right processing at the right time**. From Section D above it will be observed that, even before the draft Data Protection Regulation has been passed, the issue of consent becomes complex in the Big Data world. Consent must be 'reasonable', 'appropriate', 'real', 'meaningful' and 'able to be withdrawn' as well as 'freely given, specific and informed', all in an area where finding previously unobserved and unexpected correlations is at a premium. The ICO Paper however does give at least one pointer to what may suffice:

> "It may be possible to have a process of graduated consent. A study commissioned by the International Institute of Communications found that some users wanted to be able to give consent (or not) to different uses of their data throughout their relationship with a service provider, rather than having a simple 'binary' choice at the beginning. For example, they could give an initial consent to opt in to the system and then separate consent for their data to be shared with other parties. Furthermore, they wanted a value exchange, ie to receive some additional benefit in return for giving their consent." (paragraph 59).

27. **Transparency and notice**. Swelling its theme of transparency, the ICO Paper refers to its Privacy notices code of practice[28] at paragraph 106 after stating how privacy notices can support transparency:

> "Organisations carrying out big data analytics … need to think about promoting transparency at an early stage. The DPA contains a specific transparency requirement, in the form of a 'fair processing notice', or more simply a privacy notice. This is where the organisation tells people what it is going to do with their data when it collects it. It should state the identity of the organisation collecting the data, the purposes for which they intend to process it and any other information that needs to be given to enable the processing to be fair.

[28] http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/docum

28. **Anonymisation**[29]. Data is of course no longer personal data if it is fully anonymised, i.e.:

> "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).

However, the very nature of Big Data means that absolute anonymisation may not be possible. Organisations using anonymised data therefore:

> "should focus on mitigating the risks [of re-identification] to the point where the chance … is extremely remote" (paragraph 42)

and:

> "need to be able to demonstrate that they have carried out this robust assessment of the risks of re-identification and have adopted solutions proportionate to the risk. This may 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)

Appendix 2 (pages 51 to 53) of ICO's Anonymisation Code of Practice[30] describes these techniques in greater detail:

- ***data masking***: this involves stripping out names and other obvious personal identifiers from a piece of information to create a data set in which none are present;
- ***pseudonymisation***: attaching a coded reference or pseudonym to a record to associate the record and an individual without identification of the individual;

ICO calls data masking and pseudonymisation 'relatively high risk techniques' as the anonymised data still exists at the level of the individual

- ***aggregation***: where data is displayed as totals, with no data relating to or identifying any individual being shown.
- ***derived data and banding***: involves deriving a set of values, typically banded, from the original data which reflect the character but not the exact values of the original data.

Again, ICO characterises aggregation and banding as 'relatively low risk techniques' because data matching is 'more difficult' or 'impossible'.

29. **Privacy Impact Assessments**. The report advocates the privacy impact assessment[31] and privacy by design as tools to be used before processing begins in order to assess how Big Data analytics is likely to affect the individuals whose data is being processed and whether processing is fair.

> "[i]t is particularly important to assess, before processing begins, to what extent it is likely to affect the individuals whose data is being used. The tool to use for this analysis is a privacy impact assessment. Our code of practice on conducting privacy impact assessments gives

---

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

[30] http://ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation

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

practical advice on how to do this, and it links the privacy impact assessment to standard risk management methodologies.

Assessing privacy risk involves being clear at the outset about the benefits and aims of the big data project, as well as the impact on individuals' privacy. In many cases, the benefits in question are benefits to the organisation that is proposing to process the personal data, but it is important to factor in also benefits that may accrue to individuals or to society more broadly. When solutions to mitigate privacy risk have been identified, it is necessary to assess whether the final impact on those individuals, after those solutions have been applied, is proportionate to the aims of the project.

It will be important that a range of people of involved in big data projects understand PIAs. The organisation's data protection officer may need to co-ordinate the process but other staff, such as data scientists, need to understand how to apply PIA techniques to their work. For a PIA to be effective in a big data environment those who have the technical expertise in designing and applying algorithms must have an understanding of privacy impact."

30. **'Building Trust'**. Citing IBM and Nectar loyalty card operator Aimia, the ICO Paper notes "some evidence" 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).

31. **Information Governance**. On 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 Big Data analytics. It cites an August 2013 report by consultants Forrester Research commissioned by IBM entitled 'Big Data Needs Agile Information and Integration Governance'[32] that 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.

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

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 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."[33]

Many organisations already have in place a structured approach to data protection so the link to data governance in the Big Data context, which is likely to involve the organisation's data protection officer, is unlikely to be a large one.

32. **New regulatory techniques needed?** The Big Data data protection issues discussed at Section D above and the practical pointers to addressing them considered in this Section E are premised on what is, as a matter of practical reality, the only course currently open to ICO as the UK's data protection regulator – fitting Big Data into the DPA's legislative framework, and then developing compliance techniques from that standpoint. There is, however, a more radical view that is based on the premise that the evolution of Big Data techniques will massively scale what we know as personal data by combining in new ways datasets not hitherto connected in order to enable living individuals to be identified. This view is espoused by Ira S Rubinstein of the New York University School of Law and was articulated in a 2013 paper[34] where he argued that Big Data, while promising significant economic and social benefits, also raises serious privacy concerns and in particular challenges the Fair Information Practices that underpin all modern privacy law:

> "[The draft Data Protection] Regulation, in seeking to remedy some longstanding deficiencies with the [Data Protection Directive] as well as more recent issues associated with targeting, profiling, and consumer mistrust, relies too heavily on the discredited informed choice model, and therefore fails to fully engage with the impending Big Data tsunami. My contention is that when this advancing wave arrives, it will so overwhelm the core privacy principles of informed choice and data minimization on which the [Data Protection Directive] rests that reform efforts will not be enough. Rather, an adequate response must combine legal reform with the encouragement of new business models premised on consumer empowerment and supported by a personal data ecosystem."

---

[33] Page 6. Emphasis added.

[34] Big Data: The End of Privacy or a New Beginning, Ira S. Rubinstein, International Data Privacy Law, 2013, Vol. 3, No.2 at p. 74, http://idpl.oxfordjournals.org/content/early/2013/01/24/idpl.ips036

In all the hubbub around the draft Data Protection Directive it is to be hoped that legislators can also take a longer term view which sets data protection law in the context of a world where Big Data is much more developed, powerful and intrusive than appears to be the case today.

F. CONCLUSION

33. **Conclusion**. The respective positions in the public consciousness that data protection concerns and Big Data possibilities currently hold mean that the areas where these two plates collide are likely to generate heat as well as light. Big Data and data protection engage in two principal ways:

- first, through the application to the rapidly evolving Big Data world of established and developing data protection techniques; and
- secondly, because the capacity ever more effectively to mine and analyse datasets of ever increasing volume, variability and velocity causes personal data to increase exponentially also: Big Data techniques combine datasets which together enable living individuals to be identified in ways not hitherto possible.

In areas of law and regulation impacted by innovation, regulators and legislators typically have to respond to technology change through by the evolution of the established laws at their disposal. So ICO and other data protection authorities are addressing Big Data through developing techniques like notice and consent, anonymisation and Privacy Impact Assessments. These techniques address the first way in which Big Data and data protection engage. However, they do not yet address the bigger question of the massive increase in personal data that Big Data will produce over time and the regulatory implications of that change.

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