The problem of 'personal data' in cloud computing: what information is regulated?—the cloud of unknowing*

W. Kuan Hon**, Christopher Millard***, and Ian Walden****

Introduction

This article considers 'personal data' under the EU Data Protection Directive[1] (DPD) in relation to anonymized data, encrypted data, and fragmented data in cloud computing. Many hesitate to use cloud computing due to privacy concerns.[2] Privacy law is broad,[3] and this article focuses only on certain aspects of the DPD, whose potentially wide reach may extend to non-EU entities.

The DPD aims to encourage the free movement of personal data within the European Economic Area (EEA) by harmonizing national data protection provisions, while protecting the rights and freedoms of individuals ('data subjects') when their personal data is processed 'wholly or partly by automatic means'. It requires member states to impose certain obligations on a data 'controller' (who determines purposes and means of processing personal data) provided it has the requisite EEA connection.[4] It does not apply to certain matters,[5] where member states' national implementations may, for example, allow exemptions from certain obligations. Important national differences in data protection laws exist, such as on civil liability and penalties for non-compliance.[6] We address the DPD only at a European level, although illustrative national examples will be given.

Abstract

- Cloud computing service providers, even those based outside Europe, may become subject to the EU Data Protection Directive's extensive and complex regime purely through their customers' choices, of which they may have no knowledge or control.
- This article considers the definition and application of the EU 'personal data' concept in the context of anonymization/pseudonymization, encryption, and data fragmentation in cloud computing.
- It argues that the 'personal data' definition should be based on the realistic risk of identification, and that applicability of data protection rules should be based on risk of harm and its likely severity.
- In particular, the status of encryption and anonymization/pseudonymization procedures should be clarified to promote their use as privacy-enhancing techniques, and data encrypted and secured to recognized standards should not be

* This article forms part of the QMUL Cloud Legal Project ('CLP') <http://cloudlegalproject.org >, Centre for Commercial Law Studies, Queen Mary, University of London (CCLS). The authors are grateful to Microsoft for generous financial support, making this project possible. Views herein, however, are solely the authors'.

** Research Assistant, CLP.

*** Professor of Privacy and Information Law, CCLS; Project Leader, CLP; Research Associate, Oxford Internet Institute, University of Oxford.

**** Professor of Information and Communications Law, CCLS.

1 Directive 95/46/EC 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 on the free movement of such data [1995] OJ L 281/31.

2 Eg Ponemon Institute, *Flying Blind in the Cloud—The State of Information Governance* (7 April 2010); European Network and Information Security Agency, *An SME perspective on Cloud Computing—Survey* (ENISA, November 2009).

3 Other privacy law issues are not covered: eg confidentiality; use of private information, or right to private life under European Convention of Human Rights or EU Charter of Fundamental Human Rights. On confidential information in the cloud, see Chris Reed's CLP paper, 'Information "Ownership" in the Cloud' (2010) Queen Mary School of Law Legal Studies Research Paper No. 45/2010 <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1562461> last accessed 26 August 2011.

4 We will not discuss the DPD's applicability to an entity through its having the requisite EEA connection, or transferring personal data outside the EEA.

5 Eg national security, defence—Art 3(2).

6 C Kuner, *European Data Protection Law: Corporate Compliance and Regulation* (2nd edn, OUP, Oxford 2007), ch 1, pt G.

considered 'personal data' in the hands of those without access to the decryption key, such as many cloud computing providers.

- Unlike, for example, social networking sites, Infrastructure as a Service and Platform as a Service providers (and certain Software as a Service providers) offer no more than utility infrastructure services, and may not even know if information processed using their services is 'personal data' (hence, the 'cloud of unknowing'); so it seems inappropriate for such cloud infrastructure providers to become arbitrarily subject to EU data protection regulation due to their customers' choices.

The DPD is being reviewed and a draft reform measure is expected by the end of 2011. Cloud computing has been mentioned in many European Commission documents, so it seems likely the review will seek to address the implications of cloud computing in some fashion.[7] In this article, we argue that one aspect that the reform needs to address is the current uncertainty concerning the boundaries of what constitutes the regulated sphere of 'personal data'.

Definitions

Cloud computing definitions vary, but our definition is as follows:[8]

- Cloud computing provides flexible, location-independent access to computing resources that are quickly and seamlessly allocated or released in response to demand.
- Services (especially infrastructure) are abstracted and typically virtualized, generally being allocated from a pool shared as a fungible resource with other customers.
- Charging, where present, is commonly on an access basis, often in proportion to the resources used.

Cloud computing activities are often classified under three main service models:

- Infrastructure as a Service (IaaS)—computing resources such as processing power and/or storage;[9]
- Platform as a Service (PaaS)—tools for constructing (and usually deploying) custom applications;[10]
- Software as a Service (SaaS)—end-user application functionality.[11]

These services form a spectrum, from low-level (IaaS) to high-level (SaaS) functionality, with PaaS in between. One cloud service may involve layers of providers, not always to the customer's knowledge, and perspective affects classification. For example, customers of storage provider Dropbox may consider it a SaaS; while for Dropbox, which uses Amazon's IaaS infrastructure to provide its service, Amazon provides IaaS.[12] Furthermore, PaaS may be layered on IaaS, and SaaS may be layered on PaaS or IaaS. So, for example, PaaS service Heroku is based on Amazon's EC2 IaaS.[13]

Cloud customers also increasingly combine different providers' services. Ancillary support for primary cloud services includes analytics, monitoring,[14] and cloud-based billing systems.[15] SaaS across different providers is increasingly integrated; for example, Google Apps MarketPlace enables customers of Google Apps SaaS office suite to use third party SaaS integrating with, and managed and accessed through, Google Apps.[16] These show increasing sophistication of cloud use and layering of providers. With traditional IT, organizations may install and operate different applications, while with cloud, customers increasingly integrate different cloud applications and support services, with each other and with legacy internal systems.

The DPD's broad 'processing' definition includes any operation on data, including collection or

7 European Commission, 'A comprehensive approach on personal data protection in the European Union' (Communication) COM (2010) 609 final (November 2010); Neelie Kroes, 'Cloud computing and data protection' (Les Assises du Numérique conference, Université Paris-Dauphine, 25 November 2010) SPEECH/10/686.

8 S Bradshaw, C Millard and I Walden, 'Contracts for Clouds: Comparison and Analysis of the Terms and Conditions of Cloud Computing Services' (2010) Queen Mary School of Law Legal Studies Research Paper No 63/2010 ('CLP Contracts Paper') <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1662374> last accessed 26 August 2011.

9 Eg Rackspace; Amazon's EC2 and S3.

10 Eg Google's App Engine; Microsoft's Windows Azure.

11 Eg webmail services like Yahoo! Mail, social networking sites like Facebook, Salesforce's online customer relationship management service (enterprise SaaS).

12 CLP Contracts paper (n 8), s 3, 8.

13 Heroku, 'Can I connect to services outside of Heroku?' <http://devcenter.heroku.com/articles/external-services> last accessed 26 August 2011. Heroku's acquisition by SaaS (and, increasingly, PaaS) provider Salesforce.com was completed in January 2011. Salesforce.com, 'Salesforce.com Completes Acquisition of Heroku' (2011) <http://www.salesforce.com/company/news-press/press-releases/2011/01/110104.jsp> last accessed 26 August 2011.

14 Eg Nimsoft's cloud applications monitoring and reporting.

15 Eg, for Windows Azure users, Zuora's payments system. 'Real World Windows Azure: Interview with Jeff Yoshimura, Head of Product Marketing, Zuora' (*Windows Azure Team Blog*, 11 November 2010) <http://blogs.msdn.com/b/windowsazure/archive/2010/11/11/real-world-windows-azure-interview-with-jeff-yoshimura-head-of-product-marketing-zuora.aspx> last accessed 26 August 2011.

16 'Google Apps Marketplace now launched' (*Google Apps*, 10 March 2010) <http://googleappsupdates.blogspot.com/2010/03/google-apps-marketplace-now-launched.html> last accessed 26 August 2011.

disclosure.[17] We assume 'processing' includes all operations relating to data in cloud computing, including storage of personal data.

'Personal data' definition

Relevance to cloud computing, and problems with the definition

Central to any consideration of cloud-based processing is the 'personal data' definition. The DPD only applies to 'personal data'.[18] Information which is not, or ceases to be, 'personal data', may be processed, in the cloud or otherwise, free of data protection law requirements.

In cloud computing, the 'personal data' definitional issue is most relevant in respect of anonymized and pseudonymized data; encrypted data, whether encrypted in transmission or storage; and sharding or fragmentation of data. In each case, the question is, should such data be treated as 'personal data'?

These forms of data involve applying different procedures to personal data, at different stages, and/or by different actors. They will be discussed in detail, after considering the 'personal data' definition.

'Personal data'

Data protection law uses objective definitions for personal data and sensitive personal data, unlike privacy law's subjective 'reasonable expectations'. This results, as discussed below, in a binary, 'all or nothing' perspective, and wide-ranging applicability.

The DPD defines 'personal data' as:

> any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.[19]

Stricter regulation applies to the processing of special categories of personal data deemed particularly sensitive ('sensitive data'),[20] namely personal data revealing 'racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership', 'data concerning health or sex life', and criminal offences or convictions.[21]

Information which is 'personal data' is subject to the DPD, whatever its nature or context. Similarly, sensitive data is automatically subject to stricter rules, such that information that X has an embarrassing disease, or just flu, must be treated the same way.

Not all information merits protection as 'personal data'. Some information may seem intrinsically 'non-personal', for example meteorological information periodically recorded on Mt Everest by automated equipment.[22] Leaving aside apparently 'non-personal' information, however, DPD recital 26 recognizes that information constituting 'personal data' may be rendered 'anonymous'. Unfortunately, its interpretation and application are not straightforward,[23]especially when considering how to 'anonymise' or 'pseudonymise' personal data sufficiently to take data outside the DPD.

The Article 29 Working Party (A29WP)[24] has issued guidance on the 'personal data' concept (WP136).[25] WP136 interprets the concept broadly, stating that the DPD is intended to cover all information concerning, or which may be linked, to an individual,[26] and 'unduly restricting' the interpretation should be avoided. Seemingly over-broad application of the DPD should instead be balanced out using the flexibility allowed in applying the DPD's rules.[27]

A29WP opinions are persuasive but not binding on EU member states' courts or the European Court of Justice. Data protection regulators might be less likely to pursue entities complying with A29WP opinions, but even individual regulators, not to mention courts, may have views differing from those in A29WP opinions.[28] Therefore, in practice, controllers may exercise caution when relying on such opinions.

WP136 also emphasizes that whether information is 'personal data' is a question of fact, depending on

17 Art 2(b).

18 Including 'special category' or 'sensitive personal data'. See nn 20 and 21 and associated text.

19 Art 2(a).

20 Some other data types are also regulated more stringently, eg 'traffic data' and 'location data' under Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector OJ L 201/37, 31.07.2002 ('ePrivacy Directive').

21 Art 8. The DPD refers to 'special category' data. Such data are generally called 'sensitive data' or 'sensitive personal data'. Stricter requirements may include 'explicit' data subject consent to processing.

22 That is not strictly correct, as we see later. Even seemingly non-personal information can be 'personal data'.

23 Particularly as many national law definitions of 'personal data' also differ from the DPD definition—Kuner (n 6), ch 2.82 (tables of comparative definitions).

24 Established under DPD art 29, comprising national EU data protection regulators and the European Data Protection Supervisor (who supervises EU institutions' compliance with data protection requirements).

25 Opinion 4/2007 on the concept of personal data, WP136 (2007).

26 Even information about things can be 'personal data' if linkable to an individual—WP136 part 2.

27 WP136, 4–6.

28 Perhaps not surprising, as A29WP decisions are by simple majority—Art 29(3).

context. For example, a common family name may not single someone out within a country, but probably identifies a pupil in a classroom.[29]

Information which is not 'personal data' in the hands of one person (for example a cloud user) may, depending on circumstances, become 'personal data' when obtained or processed by another[30] (such as a cloud provider, if it tries to process it for its own purposes). Indeed, information not originally being 'personal data' may become so, if its holder processes it for other purposes, such as to identify individuals.

Similarly, when considering identifiability, account must be taken of 'all the means likely reasonably to be used by the controller or any other person' to identify them.[31] This test is dynamic. Methods 'likely reasonably to be used' may change as re-identification technology improves and costs decrease. Accordingly, the intended storage period of information is also relevant.[32]

Finally, whether information is 'personal data' may (where the processing's purpose is not identification) be affected by technical and organizational measures to prevent identification.[33] More effective measures make information more likely to be anonymous data.

Anonymization and pseudonymization

Cloud users and/or providers may process information free of the DPD if it is not 'personal data', but 'anonymous'. Also, personal data may be 'anonymised' to facilitate future cloud processing.

Anonymized or pseudonymized data result from actions deliberately taken on personal data attempting to conceal or hide data subjects' identities. Users may perform anonymization or pseudonymization procedures on datasets before processing resulting data in the cloud. Also, providers may anonymize personal data stored with them, in order to then use, sell, or share the anonymized data.[34] Some US health data storage providers anonymize and sell health data,[35] while the UK service HipSnip[36] states that it may 'share, rent, sell, or trade aggregated and anonymised data (eg which age groups prefer a particular service, but never individual information)'.[37]

Anonymization or pseudonymization as 'processing'

Processing anonymous or pseudonymized data involves two steps:

1. anonymizing or pseudonymizing 'personal data'; then
2. disclosing or otherwise processing the resulting data.

If step one itself constitutes 'processing', the DPD would apply, for example requiring data subjects' consent to anonymization or pseudonymization procedures, including explicit consent for sensitive data. WP136 did not discuss whether these procedures are 'processing'. Uncertainties regarding their status may, unfortunately, discourage their use as privacy-enhancing techniques, or the production and use of anonymized or pseudonymized data for socially desirable purposes such as medical research.[38]

This article assumes there are no problems with anonymization or pseudonymization procedures.

Common anonymization and pseudonymization techniques

Methods to 'anonymize' personal data, particularly before publishing statistical information or research results, include:

- deleting or omitting 'identifying details', for example names;
- substituting code numbers for names or other direct identifiers (this is pseudonymization, effectively);
- aggregating information, for example by age group, year, town;[39] and

29 WP136, 13, and the contextual nature of 'personal data' has been recognized eg in *Common Services Agency v Scottish Information Commissioner (Scotland)* [2008] UKHL 47, (*CSA*), [27].

30 WP136, 20, and the examples in ICO, *Determining what is personal data (Data Protection Technical Guidance)* (2007), 5.2.

31 Recital 26. WP136 considers a 'mere hypothetical possibility' to single out someone is not enough to consider the person 'identifiable'. The difficulty is, when does a 'possibility' exceed merely 'hypothetical'?

32 WP136, 15—information meant to be stored for a month might not be 'personal data' as identification may be considered impossible during its 'lifetime'. But for information to be kept for 10 years, the controller should consider the possibility of identification in say year 9 making it 'personal data' then.

33 WP136, 17: here, implementing those measures are '. . . a condition for the information precisely not to be considered to be personal data and its processing not to be subject to the Directive'.

34 Such use may indeed be key to its business model. Miranda Mowbray, 'The Fog over the Grimpen Mire: Cloud Computing and the Law' (2009) 6:1 SCRIPTed 129, 144–5. Sometimes personal data are not even anonymized before 'processing', eg automated software scanning social networking profiles (or web-based emails) to display content-based advertisements.

35 K Zetter, 'Medical Records: Stored in the Cloud, Sold on the Open Market' (*Wired*, 19 October 2009) <http://www.wired.com/threatlevel/2009/10/medicalrecords/> last accessed 26 August 2011.

36 Which enables mobile phone users to 'snip'/save eg consumer offers.

37 HipSnip, *HipSnip Legal Statement* <http://hipsnip.com/hip/legal> last accessed 26 August 2011.

38 I Walden, 'Anonymising Personal Data' [2002] 10(2) International Journal of Law and Information Technology 224. Some consider consent *should* be required for anonymization, eg where anonymized data will be used for medical research in areas where a data subject has moral objections—ibid., fn 33. In the UK, a tribunal has held that anonymization is 'processing'—*All Party Parliamentary Group on Extraordinary Rendition v The Information Commissioner & The Ministry of Defence*, [2011] UKUT 153 (AAC) (*APGER*), [127].

39 WP136, 22. Aggregation into a group attempts to make it harder to single out individuals.

- barnardization[40] or other techniques introducing statistical noise, for example differential privacy techniques with statistical databases;[41]

—or some combination of these methods.

Many anonymization and pseudonymization techniques involve amending only part of a dataset, for example disguising names, including applying cryptography to identifiers. Other information in the dataset, such as usage information or test results associated with names, remains available to those having access to resulting data.[42]

WP136 notes[43] that identification[44] involves singling someone out, distinguishing them from others:

- direct identification includes identification by name or other 'direct identifier';
- indirect identification includes identification by reference to identification number or other specific personal characteristic, including identification by combining different information (identifiers), held by controllers or others, which individually might not be identifying.

Omitting or deleting direct identifiers, such as names, while leaving indirect identifiers untouched, may not render information sufficiently non-personal. Identification numbers or similar unique identifiers may in particular enable linking of disparate information, associated with the same indirect identifier, to the same physical individual, to identify them. Nevertheless, deleting direct identifiers is often considered adequate to prevent identifiability.[45] Proposed guidance on minimum standards for de-identifying datasets, to ensure patient privacy when sharing clinical research data, recommends deleting direct identifiers including names, email addresses, biometric data, medical device identifiers, and IP addresses. If the remaining information includes at least three indirect identifiers, such as age or sex, the authors recommend independent review before publication. Thus, they consider three or more indirect identifiers presents sufficient risk of identification to require independent consideration of whether the risk is 'non-negligible'.[46]

Pseudonyms, involving substituting nicknames, etc for names, are indirect identifiers.[47] WP136 describes[48] pseudonymization as 'the process of disguising identities', to enable collection of additional information on the same individual without having to know his identity, particularly in research and statistics. There are two types:

- *Retraceable/reversible pseudonymization* aims to allow 'retracing' or re-identification in restricted circumstances.[49] For example, 'key-coded data' involves changing names to code numbers, with a 'key'[50] mapping numbers to names. This is common in pharmaceutical trials. Another example is applying two-way cryptography to direct identifiers.
- *Irreversible pseudonymization* is intended to render re-identification impossible, for example 'hashing', applying one-way cryptography (hash functions) to direct identifiers.[51]

Retraceably pseudonymized data may be 'personal data', as the purpose is to enable re-identification, albeit in limited circumstances.

If each code is unique to an individual, identification is still a risk, so pseudonymized information remains 'personal data'.[52] However, if pseudonymization reduces

40 A statistical technique aiming to anonymize statistical counts, and ensure individuals cannot be identified from statistics, while still indicating actual numbers. 0, +1 or −1 are randomly added to all non-zero counts in table cells, recalculating row/column totals accordingly. *CSA* (n 29) [8], [15].
After *CSA*, the Scottish Information Commissioner found that barnardization would not be adequate to anonymize data, but broader aggregation would. Statistics requested by age range 0–14, for each year from 1990–2003, within the Dumfries and Galloway area by census ward, were considered too identifying, even if barnardized. However, disclosure was ordered of aggregated statistics for the whole area for each year from 1990 to 2001. *Collie and the Common Services Agency for the Scottish Health Service, Childhood leukaemia statistics in Dumfries and Galloway* [2010] UKSIC 021/2005 ref 200500298.

41 C Dwork, 'Differential Privacy: A Survey of Results, Theory and Applications of Models of Computation', in Manindra Agrawal, Dingzhu Du, Zhenhua Duan and Angsheng Li (eds), *Theory and Applications of Models of Computation* (Springerlink 2008). This aims to provide accurate statistical information when querying databases containing personal information, without compromising privacy.

42 Social networking sites share 'anonymized' data, and individuals are re-identifiable from anonymized social graphs (network of individuals they're connected to). Arvind Narayanan and Vitaly Shmatikov, 'De-anonymizing Social Networks' (Proceedings of the 2009 30th IEEE Symposium on Security and Privacy, IEEE Computer Society Washington, DC, USA, 17–20 May 2009) 173.

43 12–15.

44 WP136 analysed all 'personal data' definitional building blocks. We consider only 'identified or identifiable'.

45 Eg Joined Cases C 92/09 and C 93/09 *Volker und Markus Schecke (Approximation of laws)* OJ C 13/6, 15.1.2011 (not yet published in ECR) assumes deleting names, etc would adequately anonymize recipients of certain funds.

46 I Hrynaszkiewicz and others., 'Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers' (2010) 340 British Medical Journal c181.

47 Eg, in Germany a pseudonymous person may seek access to information online service providers hold regarding his pseudonym. Kuner (n 6), ch 2.10.

48 WP136, 17.

49 Eg, with pseudonymized medical trials data, to identify individuals who may need follow-up treatment, or enable regulators to audit trials.

50 Accessible only to a restricted set of individuals.

51 Ross Anderson, *Security Engineering: A Guide to Building Dependable Distributed Systems* (2nd edn, John Wiley & Sons, 2008), Ch 5.3.1.

52 WP136, 19 suggests risks of key hacking or leakage are a factor when considering 'means likely reasonably to be used'.

the risks for individuals, data protection rules could be applied more flexibly, and the processing of pseudonymized data subjected to less strict conditions, than the processing of information regarding directly identifiable individuals.[53] Austria's DPD implementation[54] illustrates this less strict approach. Information is 'indirectly personal data' if its controller, processor, or recipient cannot identify individuals using legally permissible means. Indirectly personal data are, effectively, pseudonymous data: identities can be retraced, but not legally. Key-coded pharmaceutical trials data are considered to be 'indirectly personal data'. Such information, presumably because privacy risks are considered lower, has less protection than 'personal data'. It can, for example, be exported without regulatory approval.[55]

If the same code number is used, such as for all individuals in the same town, or all records for the same year, WP136 considers that the identification risk might be eliminated to render data anonymous. This effectively involves aggregating data; in WP136's examples, the aggregation is by town or year, respectively.

Key-coded medical trials data may be 'personal data'; but WP136 also recognizes:[56]

> This does not mean, though, that any other data controller processing the same set of coded data would be processing personal data, if within the specific scheme in which those other controllers are operating re-identification is explicitly excluded and appropriate technical measures have been taken in this respect.[57]

Therefore, key-coded data may be non-personal data when held by another person specifically not intended to identify individuals, on taking appropriate measures to exclude re-identification (for example, cryptographic, irreversible hashing).[58] Furthermore, WP136 considers[59] information may not be 'personal data' in that person's hands even if identification is theoretically possible in 'unforeseeable circumstances', such as through 'accidental matching of qualities of the data subject that reveal his/her identity' to a third party,[60] whereupon the third party would have accessed 'personal data'.

The European Commission considers transferring key-coded data to the USA (without transferring or revealing the key) is not personal data export subject to Safe Harbor principles.[61] WP136 considers itself consistent with this view as recipients never know individuals' identities; only the EU researcher has the key.[62]

We now consider 'irreversibly pseudonymised' data and aggregated data. In discussing pseudonymized data, WP136 focused on changing names or other perceived unique identifiers into code numbers, ie key-coded data, rather than attempts to pseudonymize data irreversibly by deleting direct identifiers or one-way encrypting them. It only touched on aggregation.

WP136 seems initially to equate irreversible pseudonymization with anonymization.[63] However, WP136 then states[64] that whether information is truly anonymous depends on the circumstances, looking at all means likely reasonably to be used to identify individuals. It considers this particularly pertinent to statistical information where, although information is aggregated, the original group's size is relevant. With a small group, identification is still possible through combining aggregated information with other information.

Deleting or irreversibly changing direct identifiers leaves untouched other information originally associated with that identifier. If information comprises name, age, gender, postcode, and pharmacological test results, and only names are deleted or changed, information about age, gender, etc remains. Indeed, usually the deletion or change is intended to disguise identities while enabling disclosure of other information. That

53 WP136, 18–19.

54 Datenschutzgesetz 2000.

55 Kuner (n 6), ch 2.12; Peter Fleischer, 'Austrian insights' (Peter Fleischer: Privacy . . . ?, 22 February 2010) <http://peterfleischer.blogspot.com/2010/02/austrian-insights.html> last accessed 26 August 2011.

56 Page 20.

57 Other controllers processing that data may not be processing 'personal data' because only the lead researcher holds the key, under a confidentiality obligation, and key-coding is to enable only him/her or authorities to identify individuals if necessary, while disguising trial participants' identities from recipients of pseudonymized data. Typically, recipients include research sponsors/funders or, when publishing research containing key-coded data, readers.

58 The UK Data Protection Act 1998 (DPA) 'personal data' definition differs from the DPD's, causing disagreement about how personal data may be anonymized and released. The Scottish court in *Craigdale Housing Association & Ors v The Scottish Information Commissioner* [2010] CSIH 43, [2010] SLT 655 [19] observed that the 'hard-line' interpretation, 'under which anonymised information could not be released unless the data controller at the same time destroyed the raw material from which the anonymisation was made (and any means of retrieving that material)', was 'hardly consistent' with recital 26. The Tribunal in *APGER* (n 38) considered that anonymized personal data remained 'personal data' to the controller who held the key, but could be released as it thereupon lost its 'personal data' character. A subsequent court decision, *Department of Health v Information Commissioner, [2011] EWHC 1430 (Admin)* held that a controller who anonymized personal data could disclose the resulting data, which was anonymous data even if the controller had the key to identifying individuals concerned. It noted the adverse impact a contrary ruling would have on the publication of medical statistics.

59 Page 20.

60 How accidental matching could happen was not detailed.

61 The Safe Harbor is one method enabling export of personal data to the USA. Commission Decision 2000/520/EC [2000] OJ L 215/7.

62 European Commission, *Frequently Asked Questions relating to Transfers of Personal Data from the EU/EEA to Third Countries* (2009).

63 Page 18.

64 Page 21.

purpose would be defeated if age and certainly test results had to be deleted before disclosure.

However, age, gender, etc can be identifying, when combined with each other and perhaps information from other sources.[65] Information is increasingly linkable, and individuals increasingly identifiable.[66] With automated fast data mining over large datasets, different information, perhaps from different sources, is linkable to the same individual for analysis. Over time, more information becomes linkable, increasingly enabling identification, whether data are key-coded, irreversibly-pseudonymized, aggregated, or barnardized.

To summarize, when processing data in the cloud, note that:

- Retraceably pseudonymized data, such as key-coded data, may remain personal data.
- However:
  - aggregating pseudonymized data, for example through non-unique codes, may render data 'anonymous', with dataset size being one relevant factor, enabling cloud processing of anonymous data free of the DPD, and
  - even retraceably pseudonymized data may be anonymous data in the hands of another person operating within a scheme where re-identification is explicitly excluded and appropriate measures are taken to prevent re-identification by them, even if, theoretically, others could 'accidentally' re-identify individuals.
- Critically, whether information is 'personal data' depends on the circumstances, considering all means likely reasonably to be used to identify individuals, including, for anonymized or pseudonymized data, the strength of 'anti-identification' measures used.
- Anonymization/pseudonymization procedures may themselves be 'processing'.

The A29WP was criticized for 'deficient' understanding, on the basis that dataset size, rather than quality and effectiveness of measures used, determines effectiveness of pseudonymization or anonymization procedures.[67] However, dataset size should not be the only determinant; quality and effectiveness of measures are also major factors to consider, all in the particular circumstances. If, for example, strong encryption is applied to the whole dataset, dataset size may not matter.

The DPD and WP136 certainly recognized indirect identification was possible through combining information, and WP136 mentioned dataset size and linkability, noting that deanonymization techniques would improve. However, they did not anticipate the pace of technological advances, and do not deal adequately with the implications.

Re-identification methods are progressing,[68] reinforcing the reality that current techniques, such as removing identifiers and/or aggregation, may not effectively anonymize data irreversibly.[69] Indeed, any information linkable to an individual is potentially 'personal data', because it can identify them if combined with enough other information.[70]

Encryption

Are encrypted data in the cloud 'personal data'?[71] Cryptographic applications may be used to transform or convert an entire dataset for security purposes, by applying an 'algorithm' to it, like translating information into another language, so that only those understanding that language can read it.[72]

One-way cryptography ('hashing') applies one-way functions (cryptographic hash functions) to data, producing fixed-length 'hashes' or 'hash values'. It is intended to be irreversible. Two-way cryptography ('encryption') is reversible, enabling reconstitution of the original dataset, but only by certain persons or in certain circumstances.[73] As with anonymization, applying cryptography to personal data may be 'processing'

65 Eg a person's Internet search queries can identify them, especially when different queries by the same person are recorded against the same code number, and therefore can be combined. See the AOL search results release incident, summarized in Paul Ohm, 'Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization' (2010) 57 UCLA Law Review 1701.

66 See the examples in Ohm (n 65).

67 Douwe Korff, *European Commission Comparative study on different approaches to new privacy challenges, in particular in the light of technological developments—Working Paper No. 2: Data protection laws in the EU: The difficulties in meeting the challenges posed by global social and technical developments* (European Commission 2010), 48.

68 Even differential privacy is attackable. See Graham Cormode, 'Individual Privacy vs Population Privacy: Learning to Attack Anonymization' (2010) <http://arxiv4.library.cornell.edu/abs/1011.2511> last accessed 26 August 2011.

69 Research also continues on anonymization—eg the Purdue project on anonymizing textual data, and its impact on utility <http://projects.cerias.purdue.edu/TextAnon/>.

70 Even meteorological data collected automatically on Mt Everest may be linkable to researchers publishing the data.

71 On encryption generally, see Anderson (n 51), ch 5.

72 Similarly, encrypted data are intended for use only by whoever holds and can use the decryption key—typically, whoever knows the password or passphrase required to generate and use the key, itself usually encrypted.

73 Anderson (n 51).

requiring, for example, consent or other justification. The arguments above apply equally here.

Whether encrypted data are 'personal data' depends on the circumstances, particularly 'means likely reasonably to be used' (fair or foul) to re-identify individuals. Factors affecting encrypted data's security against decryption include strength of encryption method (the algorithm's cryptographic strength); encryption key length (longer keys are generally more secure against attacks); and key management, such as security of decryption key storage, and key access control.[74] Some encryption methods have been 'broken' or 'cracked'.[75] We consider information 'strongly' encrypted if secure against decryption for most practical purposes most of the time in the real world;[76] in particular, if it is encrypted, and decryption keys secured, to recognized industry standards and best practices.[77]

Cryptography may be applied to data in an electronic file, folder, database, parts of a database, etc. Users may apply cryptography to parts or, perhaps more commonly, the whole of a dataset, recorded in whatever form, before storing it in the cloud. One-way or two-way cryptography may be applied to identifiers within personal data (for example, only names) but other data left readable as 'plaintext'. This overlaps with pseudonymization/anonymization. Alternatively, two-way cryptography may be applied to the whole dataset. Data may be encrypted within the user's computer prior to transmission, using the user's own software, or the provider's.[78] Even if users intend to process data unencrypted in the cloud, the provider may choose to encrypt all or part of the data it receives, before using or selling anonymized or pseudonymized data (for example, applying cryptography to identifiers), or to store data more securely (applying two-way cryptography to the full dataset). Transmissions may themselves be encrypted or unencrypted, usually depending on how providers set up their systems.

Regarding one-way cryptography, WP136 stated,[79] 'Disguising identities can also be done in a way that no reidentification is possible, e.g. by one-way cryptography, which creates in general anonymised data'. This suggests data containing one-way encrypted identifiers would not be 'personal data', and it does seem that 'accidental' re-identification is less likely with data anonymized through one-way cryptography. However, WP136 then discusses the effectiveness of the procedures, so in reality the key issue is the reversibility of the one-way process. Even one-way cryptography may be broken, and original data reconstituted.[80] The more secure the cryptography method, the less likely that information will be 'personal data'. If a cryptography technique employed to 'anonymise' data is cracked, to maintain 'non-personal data' status data may require re-encryption using a more secure method.[81]

Furthermore, as previously discussed, irreversibly hashing direct identifiers cannot prevent identification through indirect identifiers, other information in the dataset, and/or other sources. Thus, personal data where identifiers have been deleted or one-way hashed may, after considering such 'means likely reasonably to be used', remain 'personal data'—and their storage or use by providers subject to the DPD.

74 Matt Blaze *et al*, *Minimal key lengths for symmetric ciphers to provide adequate commercial security* (US Defense Technical Information Center 1996). The publication of secret US embassy cables on Wikileaks illustrates the importance of restricting key access. While the data were stored securely, the overall measures were not conducive to security because it seems too many people had access—Greg Miller, 'CIA to examine impact of files recently released by WikiLeaks', *The Washington Post* (22 December 2010).

75 Encryption techniques found vulnerable have required replacing by more secure algorithms. Also, technological advances—including cloud computing—facilitate decryption via 'brute force' attacks, whereby numerous computers rapidly try different keys to find what 'fits'. Eg messages encrypted using Data Encryption Standard (DES) were decrypted by the US Electronic Frontier Foundation—see <http://w2.eff.org/Privacy/Crypto/Crypto_misc/DESCracker/> last accessed 26 August 2011.

76 When weaknesses are discovered in cryptographic systems, the system will not become suddenly insecure. However, practical attacks using the techniques discovered will probably be possible someday. So, such discoveries impel migration to more secure techniques, rather than signifying that everything encrypted with that system is immediately insecure. Bruce Schneier, 'Cryptanalysis of SHA-1' (*Schneier on Security*, 18 February 2005) <http://www.schneier.com/blog/archives/2005/02/cryptanalysis_o.html> last accessed 26 August 2011.

77 See eg US Code of Federal Regulations 45 CFR Part 170, Health Information Technology: Initial Set of standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology, US Department of Health and Human Services (Federal Register, July 28, 2010). §170.210 stipulates standards to which §170.302 generally requires electronic health information to be encrypted: 'Any encryption algorithm identified by the National Institute of Standards and Technology (NIST) as an approved security function in Annex A of the Federal Information Processing Standards (FIPS) Publication 140–2 . . .'.

78 Eg Mozilla Weave, now Firefox Sync—Christopher Soghoian, 'Caught in the Cloud: Privacy, Encryption, and Government Back Doors in the Web 2.0 Era' [2010] Journal on Telecommunications and High Technology Law 8(2) 359, which suggests why cloud security measures are relatively lacking, eg encryption, and offers possible solutions.

79 Page 18.

80 Cloud computing, in the form of Amazon's EC2 GPU cluster instances, was used to find 14 passwords of 1–6 characters long from their SHA-1 hashes in under an hour, for about US$2. Thomas Roth, 'Cracking Passwords in the Cloud: Amazon's New EC2 GPU Instances' (*Stacksmashing.net*, 15 November 2010) <http://stacksmashing.net/2010/11/15/cracking-in-the-cloud-amazons-new-ec2-gpu-instances> last accessed 26 August 2011.

81 At least, within a reasonable period.

We now consider two-way encryption. WP136 focused mainly on onw-way cryptographic hashing;[82] notably, scrambling direct identifiers, supposedly irreversibly, to anonymize or pseudonymize personal data. However, users often want to store data for future use, but, to ensure security and confidentiality, apply two-way encryption to the full dataset (not just one component like names). The user can read encrypted data using its secret decryption key; others are not intended to decipher it.

The 'personal data/not personal data' debate concentrated on anonymizing or pseudonymizing parts of a dataset. However, WP136 applies equally to two-way encryption of full datasets. Under WP136, 'anonymised' data may be considered anonymous in a provider's hands if 'within the specific scheme in which those other controllers [eg providers] are operating re-identification is explicitly excluded and appropriate technical measures have been taken in this respect'. On that basis, we suggest that if you cannot view data, you cannot identify data subjects, and therefore identification may be excluded by excluding others from being able to access or read data. By analogy with key-coded data, to the person encrypting personal data, such as a cloud user with the decryption key, the data remain 'personal data'. However, in another person's hands, such as a cloud provider storing encrypted data with no key and no means 'reasonably likely' to be used for decryption,[83] the data may be considered anonymous.

This may arguably remove cloud providers from the scope of data protection legislation, at least where data have been strongly encrypted by the controller before transmission, and the provider cannot access the key.[84] Consider a SaaS provider using a PaaS or IaaS provider's infrastructure to offer its services. The PaaS or IaaS provider may not have any keys, even if the SaaS provider does—so the information may be 'personal data' to the SaaS provider, but not other providers. In SaaS involving mere storage where users encrypt data before transmission, even the SaaS provider may not have the key.

When encrypting a full dataset, size should not matter; unlike with key-coded data, no data remain available 'in the clear' as potentially linkable, indirectly identifying information. Encrypted data, transformed into another form, differs qualitatively from, and arguably poses fewer risks than, key-coded data or aggregated data, so there is a stronger argument that fully-encrypted data are not 'personal data' (to those without the key).

The issue again is security against decryption by unauthorized persons. Stronger 'anti-identification' measures applied to data make it more likely the data will be anonymous. Again, encryption strength is important, as is the effectiveness of other measures such as key management. If personal data were not encrypted before transmission to the cloud, or only weakly encrypted, or if the key was insecurely managed, data stored might be 'personal data', and the provider a 'processor'.[85] However, if personal data were encrypted strongly before transmission, the stored data would be unlikely to be 'personal data' in the provider's hands.

However, why should whether a user decides to encrypt data,[86] or the effectiveness of their chosen encryption method or other security measures, determine whether encrypted data hosted by providers constitute 'personal data'? Generally, 'pure' cloud storage providers[87] cannot control in what form users choose to upload data to the cloud.[88] Nor would providers necessarily know the nature of data users intend to store, hence the 'cloud of unknowing' in this article's title. Yet

82 Passwords are often 1-way 'hashed' by applying a 1-way cryptographic 'hash function' to the password, and the resulting hash value stored. A password later entered is similarly hashed. If the two hash values are identical, the password is accepted. This avoids insecurely storing 'cleartext' passwords. Comparing hashes also enables integrity checking, ie detecting changes or corruption to original data. Hashes can be transmitted and/or stored with the original data. If integrity is compromised, the hashes will differ. WP136 discusses 1-way cryptography, not for password authentication or integrity checking, but to scramble irreversibly identifiers within a larger dataset.

83 Bearing in mind that cloud computing may itself increasingly be 'means likely reasonably' to be utilized to decrypt data!—'Cracking Passwords in the Cloud: Breaking PGP on EC2 with EDPR' (*Electric Alchemy*, 30 October 2009) <http://news.electricalchemy.net/2009/10/cracking-passwords-in-cloud.html> last accessed 26 August 2011.

84 Dropbox (text to n 12) had to clarify that it held keys and could access users' encrypted data—Ryan Singel, 'Dropbox Lied to Users about Data Security, Complaint to FTC Alleges' (*Wired*, 13 May 2011) <http://www.wired.com/threatlevel/2011/05/dropbox-ftc/> last accessed 26 August 2011.

85 Assuming a provider storing personal data 'processes' data for customers, and so is a 'processor'—see W Kuan Hon, Millard and Walden, 'Who is Responsible for "Personal Data" in Cloud Computing? The Cloud of Unknowing, Part 2' (2011) <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1794130> last accessed 26 August 2011.

86 Disadvantages to storing data in encrypted form include the inability to index and therefore search encrypted data. This may lead some users not to encrypt data. However, searchable encryption is being investigated, and may someday become feasible—Seny Kamara and Kristin Lauter, 'Cryptographic Cloud Storage', in Radu Sion and others (eds), *FC'10 Proceedings of the 14th International Conference on Financial Cryptography and Data Security* (Springer-Verlag Berlin, Heidelberg 2010), 136.

87 Ie providers of IaaS or PaaS for storage by users of data on the provider's infrastructure, or providers of SaaS as 'storage as a service', where the service is limited to data storage and tools to upload, download, and manage stored data.

88 They can control it if they build it into their systems. Eg Mozy's procedures involve encrypting data on a user's local computer, using the user's key, before transfer to Mozy's servers via encrypted transmissions. Mozy, Inc, 'Is MozyHome secure?' (2010) <http://docs.mozy.com/docs/en-user-home-win/faq/concepts/is_it_secure_faq.html> last accessed 26 August 2011.

the status of data stored with providers, which affects the provider's status as 'processor' (or not) of data stored by users, will vary with each user's decisions and actions—and may differ for different users, even for the same user storing different kinds of data, or the same data at different times. This seems unsatisfactory.

Regarding providers' knowledge, an Italian court has considered that, to impose criminal data protection liability on a data host for not monitoring or pre-screening uploaded sensitive data, there must be knowledge and will on its part—even with a SaaS provider considered more than a passive storage provider.[89]

With data in transmission ('data in flight' or 'data in motion'), the connection for transmitting unencrypted or encrypted data may itself be unencrypted or encrypted,[90] normally depending on how the provider set up its systems. That is generally within the provider's control. But if users transmit unencrypted personal data, even via secure channels, providers will still receive personal data as such.

Transmission and longer-term storage may merit different treatment. Transmission durations, and therefore possible interception windows, may be relatively short. Therefore in many cases perhaps 'data in motion' need not be as strongly encrypted, to make transmitted information non-personal.[91] However, stronger encryption may be necessary for data in persistent storage to be considered non-personal.[92]

The DPD forbids exports to 'third countries' without an 'adequate level of protection'. Art 25(2) requires assessing adequacy in light of all circumstances surrounding the export operation or set of operations, giving particular consideration to matters including the proposed processing operation(s)' duration. This implies that if an operation, and therefore presence of personal data in the country, is of shorter duration, risks are lower, and less stringent protective measures may be considered adequate, than if data were located there for longer. Similarly, information exposed unencrypted for relatively short periods for transient processing operations should arguably not lose 'anonymous' status thereby.

However, suppose military-grade security measures are applied to data transmitted for storage, rendering the data non-'personal' in the provider's hands. For applications to process data subsequently, such as for sorting or analysis, the data requires decrypting first.[93] Research continues on enabling secure operations on encrypted data, for example inside encrypted 'containers'. Hopefully secure computation[94] will become practicable.[95] However, currently such operations would take an unfeasibly long time.[96]

Therefore, currently, to run applications on (originally personal) data stored encrypted in the cloud, the user must first decrypt the data, necessarily involving processing 'personal data'.[97] If users download data before decryption, providers would not be involved in users' local processing of decrypted personal data, but users would lose cloud processing power. If users run cloud applications on decrypted data in the provider's servers, the provider could become a 'processor'. However, as with transmissions, such operations may be transient; data may remain in encrypted form for longer than in decrypted 'personal data' form. Must all the DPD's requirements nevertheless be applied to those operations, or would more limited application be sensible?

In summary, to try to render information stored in the cloud non-'personal data' in the provider's hands, the best course seems to be to encrypt it strongly before transmission. However, the matter is unclear, and even personal data encrypted for storage as anonymous data must be decrypted for operations, which will therefore be on 'personal data'.

Uncertainties regarding whether, and when, encrypted data are considered 'personal data', and to what extent users' own encryption or other decisions affect data's status in the cloud, cause practical con-

89 Liability was imposed for other reasons. The judgment is being appealed. Judge Oscar Magi, Milan, Sentenza n 1972.2010, Tribunale Ordinario di Milano in composizione monocratica 92–96.

90 The TLS/SSL protocol is used to secure transmission of, eg, credit card details between web browser and remote server (https). Connections may also be secured using virtual private networks (VPNs). Eg, IaaS provider Amazon offers VPN connections between 'Virtual Private Clouds' on its infrastructure and users' own data centres.

91 Although, if transmissions between particular sources/destinations are actively monitored, brevity of individual transmissions may be irrelevant.

92 Where, therefore, the potential attack window will be longer.

93 Mowbray (n 34), 135–6.

94 Eg N Smart and F Vercauteren, 'Fully Homomorphic Encryption with Relatively Small Key and Ciphertext Sizes' in Phong Q Nguyen and David Pointcheval (eds), *Public Key Cryptography—PKC 2010*, 420 (Springer, Berlin/Heidelberg 2010).

95 R Chow and others, 'Controlling data in the cloud: outsourcing computation without outsourcing control' in Radu Sion and Dawn Song (chairs), *Proceedings of the 2009 ACM workshop on Cloud computing security* (ACM, Chicago, Illinois 2009). Other approaches to secure cloud processing include data obfuscation—Miranda Mowbray, Siani Pearson and Yun Shen, 'Enhancing privacy in cloud computing via policy-based obfuscation' (online 31 March 2010) The Journal of Supercomputing DOI: 10.1007/s11227–010–0425-z.

96 Daniele Catteddu and Giles Hogben, *Cloud Computing—Benefits, Risks and Recommendations for Information Security* (European Network and Information Security Agency, 2009) 55, V10.

97 Even if information is 'personal data' ephemerally, whilst operated on in decrypted form, and then encrypted and saved as such. Information which is 'personal data' ephemerally is not exempt, but, based on WP136, arguably poses lower risks to data subjects, as with data in flight.

cerns. It is hoped the revised DPD clarifies the position on encrypted data and encryption procedures.[98]

Sharding

We now consider cloud data storage mechanics and implications. In this section and the next, we do *not* deal with strongly encrypted data—only data which users have not encrypted, or secured only weakly. This is because strongly-encrypted 'personal data' should already be considered 'anonymous' in the hands of a provider without key access.

IaaS, PaaS, and SaaS can store data as unencrypted 'plaintext'.[99] With cloud applications beyond simple storage, often data are stored unencrypted, or providers may access users' keys or their own secondary decryption keys. This enables value-added services, such as indexing and full-text searching of data, retention management or data format conversion, not possible with encrypted data.[100]

Cloud computing typically uses virtual machines (VMs) as 'virtual servers' (hence, 'server virtualisation'). VMs simulate physical computers, existing only in the RAM of a physical server hosting multiple VMs. If data operated on within a VM are not saved to persistent storage[101] before the VM's termination or failure, generally the data are lost. Depending on the service, even VM instances appearing to have attached storage may lose 'stored' data on 'taking down' the instance, unless actively saved to persistent storage first.

Providers often offer persistent data storage on non-volatile equipment, enabling data retrieval after the current VM instance terminates.[102] To 'scale out' flexibly, adding (often commodity-priced) physical equipment whenever more processing power or storage space is needed, providers employ 'storage virtualisation'. Stored data appear as one logical unit (or several) to the user. The provider's software automatically handles physical storage using distributed file systems, distributed relational databases such as MySQL, and/or distributed non-relational 'NoSQL' databases, which can store files over different hardware units or 'nodes', even in different locations.[103] To maximize efficient resource utilization or for operational reasons, stored data may 'move' or be copied between different hardware, perhaps in different locations: a much-publicized feature of cloud computing.

For backup/redundancy, performance, availability, and tolerance to failure of individual nodes, stored data are normally 'replicated' automatically in two[104] or three[105] data centres.

'Sharding'[106] or fragmentation, also known as 'partitioning', involves providers' software automatically splitting data into smaller fragments ('shards'), distributed across different equipment, possibly in different locations, based on the provider's sharding policies, which vary with space constraints and performance considerations.[107] Applications' requests for operations on data are automatically sent to some or all servers hosting relevant shards, and results are coalesced by the application. Sharding assists availability—retrieving smaller fragments is faster, improving response times. While 'sharding' most commonly refers to fragmenting databases, data not within a structured database may also be fragmented for storage or operations.[108] We use 'sharding' to mean all forms of data fragmentation.

98 See IaaS provider Rackspace US, Inc.'s submission, *International transfer of personal data (Consultation Paper on the Legal Framework for the Fundamental Right to Protection of Personal Data)* (2009) <http://ec.europa.eu/justice/news/consulting_public/0003/contributions/organisations_not_registered/rackspace_us_inc_en.pdf> last accessed 26 August 2011.

99 This obviously includes 'storage as a service' SaaS. It may also include other SaaS like webmail where, as well as providing cloud application software, in this case email client software, the provider also stores the data used in relation to that application, eg in this example emails and contacts data.

100 Tim Mather, Subra Kumaraswamy and Shahed Latif, *Cloud Security and Privacy: An Enterprise Perspective on Risks and Compliance* (O'Reilly 2009), 62.

101 Eg ultimately one or more hard disk drives, or SSD flash storage.

102 Storage systems include Amazon's S3, SimpleDB or Elastic Block Storage, and Windows Azure's SQL Azure, blob or table storage or XDrive.

103 Eg Facebook uses relational database MySQL and NoSQL database Cassandra—Sean Michael Kerner, 'Inside Facebook's Open Source Infrastructure' (*Developer.com*, 22 July 2010) <http://www.developer.com/features/article.php/3894566/Inside-Facebooks-Open-Source-Infrastructure.htm> last accessed 26 August 2011. NoSQL databases, while not relational databases, scale easily. Other NoSQL databases are Google's BigTable, Amazon's Dynamo and (open source) HBase.

104 Google Apps—Rajen Sheth, 'Disaster Recovery by Google' (*Official Google Enterprise Blog*, 4 March 2010) <http://googleenterprise.blogspot.com/2010/03/disaster-recovery-by-google.html> last accessed 26 August 2011.

105 Microsoft Windows Azure—Microsoft, 'Introduction to the Windows Azure Platform' (*MSDN Library*), <http://msdn.microsoft.com/en-us/library/ff803364.aspx> last accessed 26 August 2011.

106 Users may partition or 'shard' cloud databases logically, eg into different 'domains' they create on Amazon's SimpleDB NoSQL database system; we use 'sharding' to mean only automatic data fragmentation by providers' systems. Users have no say in such automated sharding, although some providers, eg Amazon and Azure, allow users to confine to broad geographically circumscribed regions, eg EU or Europe, the storage (and presumably other processing) of the resulting shards. CLP Contracts paper (n 8).

107 Eg to equalize workload across different servers and/or data centres. LA Barroso and U Hölzle, *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines* in Mark D Hill (ed.), *Synthesis Lectures on Computer Architecture* (Morgan & Claypool 2010) 16.

108 Eg data stored using Amazon's Elastic Block Store (EBS), appearing to users as physical hard drives, are, before being stored as EBS snapshots on Amazon's S3 storage service, first broken into chunks whose size depends on Amazon's optimizations. Amazon, *Amazon Elastic Block Store (EBS)* (2011) <http://aws.amazon.com/ebs/> last accessed 26 August 2011.

Where unencrypted personal data are automatically sharded[109] for distributed storage, the user uploading such data, perhaps running cloud applications on data and receiving coalesced results, is clearly processing personal data.

However, are individual shards, in distributed storage on providers' equipment, 'personal data' in the provider's hands? Key-coded data may be anonymous data in another's hands, provided only the researcher has the key and adequate measures against re-identification are taken. With cloud computing, much depends on sharding or replication methods, and measures to restrict shard 'reunification' and allow only authorized users to access reunified data.

If a shard contains sufficient information to identify an individual, its access or replication involves 'processing' personal data. Even a tiny fragment may contain personal data, as the Google Streetview incident illustrated.[110] However, what matters is not shard size, but content, and intelligibility. Even if a shard contains personal data, if that data are only intelligible to the user, duly logged in, arguably it is 'personal data' only to that user. Providers have different sharding systems, so further analysis is difficult without exact details.[111]

Some fear data 'in the cloud' may be seized or accessed by local law enforcement authorities or others in the jurisdiction where storage equipment is located.[112] However, if a third party physically seizes hard drives or other equipment containing a shard, would it thereby retrieve 'personal data' stored by the targeted user? Not if the seized equipment holds only an incomprehensible shard, and it cannot access equipment holding the remaining shards, for example because it is in another jurisdiction. Related shards may be stored in the same jurisdiction, data centre or equipment; but they may not be. Where all shards are retrievable, it may not be able to reunify them intelligibly without the provider's cooperation (or indeed even with it, depending on the system's design and whether the data were encrypted).[113]

Operations on data may be distributed, split into smaller sub-operations running simultaneously in different nodes, perhaps in different locations, each processing a different shard or dataset sub-set. Sub-operation results are combined and sent to the user. When running cloud applications on data, such distributed processing may be employed automatically. While the application operates on data, shards may be stored in the provider's equipment, usually ephemerally, irrespective of whether the user intends original or resulting data to be stored permanently in the cloud. Similar issues would thus arise regarding whether such shards include intelligible personal data, and whether, as with transient operations on decrypted data, such temporary operations or storage merits full application of all the DPD's requirements.

In summary, the position on storing or operating on shards is unclear. Detailed operational aspects, varying with services and/or users, may determine whether stored information is 'personal data'. Again, this seems unsatisfactory. More transparency by providers as to sharding and operational procedures would help inform the debate.

Provider's ability to access data

We argued that strongly-encrypted data should not be 'personal data' to those without the key, as individuals cannot be identified without decryption. What about cloud data stored unencrypted, or only weakly encrypted?

Even with personal data stored unencrypted in the cloud, re-identification of individuals through stored

109 Automated sharding, as with anonymization, may itself be 'processing'—3.3.1. However, just as the DPD should permit or encourage anonymization, arguably sharding into non-personally identifying fragments should be allowed, at least where each shard is too small to contain personal data.

110 Google's vehicles travelled streets globally, collecting data for its Street View online mapping service. It was discovered that they also captured data transmitted over open (non password-protected) wi-fi networks, eg some consumers' home networks. That data included 'payload' data (content of transmitted data), as well as routing information. Various data protection authorities' investigations found 'while most of the *[captured]* data is fragmentary, in some instances entire emails and URLs were captured, as well as passwords'. Alan Eustace, 'Creating stronger privacy controls inside Google' (*Official Google Blog* 22 October 2010) <http://googleblog.blogspot.com/2010/10/creating-stronger-privacy-controls.html> last accessed 26 August 2011. Combining such data with geolocation to discover from whose home data originated, correlating location with residents' identities, could allow association of usernames/passwords with individuals.

111 Eg provider Symform offers distributed storage on users' own computers. Users' files are split into blocks, each block is encrypted, further fragmented, and distributed for storage. Even for very small files within a single data fragment:
[T]he information about which file is associated with which data fragment and where that data fragment is located is stored separate from the data fragment itself—in Symform's cloud control. So, an attacker would have to identify a file and break into Symform to find out where its fragments are located. After this, they would have to actually have access to at least one of those fragments to be able reconstruct the encrypted contents. Last, and certainly not least, they would have to break the 256-AES encryption. (Symform, 'How Symform Processes and Stores Data' (Symform) <http://symform.com/faq-how-symform-processes-and-stores-data.aspx>) last accessed 26 August 2011.

112 On law enforcement issues in cloud computing see another CLP paper: Ian Walden, 'Law Enforcement Access in a Cloud Environment' (2011) Queen Mary School of Law Legal Studies Research Paper No. 72/2011 <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1781067> last accessed 26 August 2011.

113 See further 3.6.

data is achievable only by someone who can access re-unified data shards in intelligible form. What if the user is the only person who can access reunified shards of their stored data, through logging into their account with the provider, and the provider's scheme excludes anyone else from being able to access the account, and thence access intelligible data? Arguably the data may be 'personal data' to the user, but not to anyone else.[114]

In other words, effectiveness of measures to prevent persons other than the user from accessing a user's stored unencrypted personal data, may affect whether data are 'personal data' as regards those persons. A key factor will be, how effective is the provider's access control system, which typically only allows authenticated and authorized users to access a particular cloud account? By logging into their account, the user can access and operate on the full set of any personal data stored. That does not mean, however, that others can access the set.[115] More effective and restrictive access control measures would make it more likely that re-identification by others will be excluded, and therefore that stored data will not constitute 'personal data'.[116]

Another key factor is whether 'backdoors' exist allowing providers to sign in to users' accounts or otherwise access users' re-unified data. One advantage of cloud computing is that providers usually maintain and update automatically the log-in software and (for SaaS) application software. While convenient for users, this also means providers can, unbeknownst to users, build in and use backdoors, or even introduce backdoors at the behest of law enforcement or other authorities. While equipment containing fragmentary personal data may be seized, as discussed above, undesired access to cloud data may be more likely to occur through the provider's ability to access,[117] or allow third parties[118] to access, re-unified shards of stored data, by accessing users' accounts, wherever its data centres are located (even other countries).

Currently, many providers' contract terms expressly reserve rights to monitor users' data and/or data usage.[119] Where users are consumers, the provider is probably controller of any personal data collected regarding users. This may include personal data consumers provide during sign-up, as well as, for instance, metadata[120] generated regarding their ongoing service usage. However, we do not cover monitoring of user-related personal data. We consider only any personal data, perhaps relating to others, processed by users using providers' facilities.

Now, having rights and technical ability to access data, is not the same as actually accessing data. Unless and until a provider accesses data, it may not even know data are 'personal data'. Should not a provider who restricts access to very few employees, in tightly controlled circumstances, for example only as necessary for maintenance and proper provision of services, and who takes other measures, such as regularly auditing access logs, be exposed to fewer liabilities than providers who, say, allow all employees access to users' data anytime for any purpose?

Providers who enable any internal or external access to users' accounts or data face a difficulty, even with strictly controlled access. WP136 does not envisage limited re-identification incidental to accessing data, for example to investigate service issues, rather than to identify data subjects. The scheme must exclude re-identification before data may be considered non-'personal' for providers. Thus, it seems, if a provider can access users' unencrypted stored personal data, its scheme does not exclude identification—so data stored unencrypted with it would be 'personal data'.

114 Leaving aside for now that individual shards may contain intelligible personal data. A German data protection regulator has reported the head of Google's cloud services in Central and Northern Europe as saying that if anyone broke into Google's 'top-secret computer center', the intruder would find '"absolutely nothing" usable, only "meaningless bits and bytes" because Google uses a proprietary file system'. The regulator however then expressed the view that security by transparency, with state-of-the-art security measures, is preferable to security by obscurity. Thilo Weichert, 'Cloud Computing and Data Privacy' (The Sedona Conference 2011) <https://www.datenschutzzentrum.de/cloud-computing/20100617-cloud-computing-and-data-privacy.pdf>, 10–11, last accessed 26 August 2011.

115 Subject to any 'backdoors', discussed below.

116 Users' login password strength may also affect access control measures' effectiveness. Again, something within users' control rather than the provider's, ie user password selection, affects whether information held by the provider is 'personal data'.

117 Google's Site Reliability Engineers had 'unfettered access to users' accounts for the services they oversee'. In 2010 one such engineer was dismissed for accessing minors' Google accounts without consent, including call logs and contact details from its Internet phone service, instant messaging contact lists, and chat transcripts. Adrian Chen, 'GCreep: Google Engineer Stalked Teens, Spied on Chats (Updated)' (*Gawker* 14 September 2010) <http://gawker.com/5637234/> last accessed 26 August 2011. See, regarding allegedly 'universal' employee access to users' accounts on social networking site Facebook, Soghoian (n 78), fn 99.

118 Eg, law enforcement authorities in the country of the provider's incorporation, or who otherwise have jurisdiction over it, or private parties under court judgments made against it—Soghoian (n 78). See also Google's Government Requests Report—David Drummond, 'Greater transparency around government requests' (*Google Public Policy Blog*, 20 April 2010) <http://googlepublicpolicy.blogspot.com/2010/04/greater-transparency-around-government.html> last accessed 26 August 2011. Concerns have recently been raised about US providers handing users' data to US authorities if compelled by US law, eg the US PATRIOT Act, including data located outside the USA. See, for example, Jennifer Baker, 'EU upset by Microsoft warning about US access to EU cloud' (NetworkWorld 5 July 2011) <http://www.networkworld.com/news/2011/070511-eu-upset-by-microsoft-warning.html> last accessed 26 August 2011.

119 CLP Contracts paper (n 8), s 4.11, 30.

120 On metadata generated by providers, see Reed (n 3), 9.

Many SaaS services go beyond 'pure' storage. Providers may access stored unencrypted personal data (for example, to run advertisements against content), and/or personal data are exposed to widespread, even public view, such as social networking or photo sharing sites.[121] Thus, excluding identification may be impossible. So too with SaaS 'passive' storage services, IaaS, or PaaS where, although stored unencrypted data are meant to be accessible only to the user, to investigate problems the provider's engineers need the ability to login to users' accounts or view stored data, and accordingly may see any identifying information therein.[122] Similarly where comprehensible 'personal data' shards remain temporarily on providers' equipment, pending automatic overwriting by other data after users decide to delete data or terminate accounts, at least where the provider can read shards marked for deletion.[123] Therefore, currently, it seems some data stored by these cloud services must be treated as 'personal data'.

This appears inevitable from WP136's focus on preventing identification, rather than assessing risks to privacy in context. However, there may be policy reasons for encouraging the development of cloud infrastructure services and recognizing infrastructure providers' more neutral position. More flexible application of the DPD's requirements may be appropriate in some cloud situations, for example imposing fewer requirements on infrastructure or passive storage providers, than SaaS providers who actively encourage or conduct personal data processing.

Even with publicly-accessible unencrypted personal data, one twist merits exploration. The DPD's prohibition on processing sensitive data does not apply to the processing of data 'which are manifestly made public by the data subject . . . .'[124] Such sensitive data may be processed without explicit consent; the data subject's publicization justifies others' further processing. The data remain subject to the DPD—so consent or another justification is still needed. Arguably even non-sensitive personal data publicized by the data subject should become free of the DPD's requirements, although the DPD did not so provide. Related difficulties include determining when data are 'made public',[125] especially with social networks, and the role of data subjects' intention to publicize data.

Consumers are posting unprecedented amounts of information online, including personal data, yet are not necessarily aware of possibly extensive consequences, both for themselves and others whose data they post. Regulators and legislators are increasingly concerned about consumer protection, and are focusing on issues such as the importance of more privacy-friendly default settings.[126] Many social networking sites' default settings make posted information available to a wider group, sometimes even publicly; whereas some consumers may believe posted data are only available to a limited group. Policy makers may therefore be reluctant to endorse free processing of personal data publicized by data subjects. Nevertheless, there may still be scope for relaxing the DPD requirements applicable to such data.

In summary, effectiveness of access control restrictions and any means for a provider to access personal data stored unencrypted with it may affect whether data are 'personal data' in the provider's hands, even if the provider only has limited incidental access. However, arguably the DPD's rules should not be applied in full force, or at all, to infrastructure providers such as pure storage providers, who may not know the nature of the data stored in their infrastructure 'cloud of unknowing'.[127]

The way forward?

We suggest a two-stage approach. First, the 'personal data' definition should be based on a realistic likelihood of identification. Secondly, rather than applying all the DPD's requirements to information determined to be 'personal data', it should be considered, in context, which requirements should apply, and to what extent, based on a realistic risk of harm and likely severity.

The 'personal data' concept currently determines, in a binary, bright line manner, whether the DPD regulates information. If information is personal data, all the DPD's requirements apply to it, in full force; if not, none do. However, as WP136 and courts have recognized, 'personal data' is not binary, but analogue. Identifiability falls on a continuum. There are degrees of

121 CLP Contracts paper (n 8), s 4.9, 27.

122 Similarly with storage services offering indexing and searching facilities or other value-added services requiring the provider or its software to be able to access data—Mather, Kumaraswamy, Latif (n 100).

123 CLP Contracts Paper (n 8), s 4.8, 23. Where providers do not promptly delete such data, including duplicates, and retained data contain personal data, would that affect the provider's status?

124 Art 8(2)(e).

125 Peter Carey, *Data Protection: a Practical Guide to UK and EU Law* (OUP, 2009), 86: a TV interview statement is public; what about a personal announcement to friends?

126 Eg A29WP and Working Party on Police and Justice, *The Future of Privacy—Joint contribution to the Consultation of the European Commission on the legal framework for the fundamental right to protection of personal data, WP 168* (2009) [48], [71]; A29WP, *Opinion 5/2009 on online social networking, WP 163* (2009), 3.1.2 and 3.2.

127 Hon, Millard and Walden (n 85).

identifiability; identifiability can change with circumstances, who processes information, for what purpose; and as information accumulates about someone, identification becomes easier. On this basis, almost all data is potentially 'personal data', which does not help determine applicability of the DPD to specific processing operations in practice. Similarly, whether particular information constitutes 'sensitive data' is often context-dependent. As the UK Information Commissioner (ICO) stated,[128] while many view health data as 'sensitive', is a management file record that an employee was absent from work due to a cold, particularly sensitive in any real sense?

Advances in re-identification techniques have surely put paid to justifications for applying the DPD's requirements in an 'all or nothing' fashion. The ICO also considers that the 'personal data' definition needs to be clearer and more relevant to modern technologies and the practical realities of processing personal data within both automated and manual filing systems.[129]

The ICO pointed out that any future framework must deal more effectively with new forms of identifiability, but suggested different kinds of information, such as IP address logs, might have different data protection rules applied or disapplied.[130] Its overall view[131] was that 'a simple "all or nothing" binary approach to the application of data protection requirements no longer suffices, given the breadth of information now falling within the definition of "personal data"'.

In its Communication[132] the European Commission did not consider changing or eliminating the 'personal data' definition. However, it noted,[133] as regards this definition, 'numerous cases where it is not always clear, when implementing the Directive, which approach to take, whether individuals enjoy data protection rights and whether data controllers should comply with the obligations imposed by the Directive'. It felt certain situations, involving processing specific kinds of data,[134] require additional protections, and will consider how to ensure coherent application of rules in light of new technologies and the objective of ensuring free EU-wide circulation of personal data.

It may be time to consider focusing less on protecting data as such,[135] and more on properly balancing, in specific processing situations, protection of individuals' rights regarding their data, with free data movement within the EEA and, where appropriate, beyond. In particular, rather than considering solely whether information is personal data,[136] it may make more sense to consider risk of identification and risk of harm[137] to individuals from the particular processing, and the likely severity of any harm. Processing should then be tailored accordingly, taking measures appropriate to those risks.[138]

Arguably the A29WP's underlying approach is already risk-based; consider, for example, WP136's suggestion that pseudonymous data may involve fewer risks. Regarding anonymization, the ICO has pointed out different levels of identifiability and argued for a more nuanced and contextual approach to protecting 'personal data' and 'anonymised' data.[139]

Austria's treatment of 'indirectly personal data' has been mentioned.[140] Sweden's 2007 changes to its implementation of the DPD[141] also illustrate a risk-based

128 ICO, *The Information Commissioner's response to the Ministry of Justice's call for evidence on the current data protection legislative framework* (2010) ('MoJ').

129 Ibid.

130 Eg, for IP logs, requiring security measures but not subject access rights or consent to log recording—ICO, *The Information Commissioner's response to the European Commission's consultation on the legal framework for the fundamental right to protection of personal data* (2010), 2; and MoJ (n 128), 7.

131 MoJ (n 128), 7; ICO (ibid.), 2; ICO, *The Information Commissioner's (United Kingdom) response to a comprehensive approach on personal data protection in the European Union—A Communication from the European Commission to the European Parliament, the Council, the Economic and Social Committee and the Committee of the Regions on 4 November 2010* (2011).

132 See n 7.

133 Ibid., 5.

134 Eg location data; even key-coded data.

135 The GSMA has noted that EU data protection rules are based on distinguishing individuals from others, 'even though an organisation collecting and processing such information has no intention of using it to target or in any way affect a specific individual'. Martin Whitehead, *GSMA Europe response to the European Commission consultation on the framework for the fundamental right to the protection of personal data* (GSMA Europe, 2009) <http://ec.europa.eu/justice/news/consulting_public/0003/contributions/organisations/gsma_europe_en.pdf> last accessed 26 August 2011.

136 As the ICO pointed out regarding manual records (MoJ, n 128), reliance on the 'personal data' definition has resulted in attempts to avoid regulation by trying to structure situations so that data fall outside it.

137 As per APEC's Privacy Framework (APEC Secretariat, 2005), and see Christopher Millard, 'Opinion: The Future of Privacy (part 2)—What might Privacy 2.0 look like?' (2008) 5(1) Data Protection Law & Policy, 8–11. Technology multinational Cisco supports 'a more explicit link between harm and the necessary data protection requirements', noting that the DPD 'tends towards seeing all personal data as worthy of protection regardless of the privacy risk'. This approach's divergence from one based on privacy risk 'is exacerbated by the broad definition of personal data, which encompasses any data that can be linked to an individual'. Cisco Systems, *Cisco response to the consultation on the legal framework for the fundamental right to protection of personal data* (2009) <http://ec.europa.eu/justice/news/consulting_public/0003/contributions/organisations/cisco_en.pdf> s 3.3, 5, last accessed 26 August 2011.

138 If you are identified as belonging to a group with 100 members, the chances (and risks) of identifying you are 1 in 100. That may entail more precautions with that data, than with data which would only identify you as belonging to a group with1 billion members.

139 MoJ (n 128), 8.

140 Text to n 54.

141 Personuppgiftslag (1998:204) (Swedish Personal Data Act).

approach, with reduced regulation of personal data contained in unstructured material such as word processing documents, webpages, emails, audio, and images, presumably based on risks being considered lower there. Sweden relaxed requirements for processing such material, provided it is not included or intended for inclusion in a document management system, case management system or other database. Such unstructured data are exempt from most of the DPD's requirements, such as the restrictions on export. However, security requirements apply, and processing of such personal data must not violate data subjects' integrity (privacy).[142] Thus, the focus is on preventing harm. In a report on lacunae in the Council of Europe's Convention 108[143] arising from technological developments, the author argued it was increasingly less relevant to ask whether data constituted personal data—rather, one should identify risks relating to the use of data posed by technologies in a particular context, and respond accordingly.[144]

Whether information is 'personal data' already requires consideration of specific circumstances. Assessing risks of identification/harm posed by a particular processing operation would not seem more difficult than determining whether particular information is 'personal', yet may be more successful in making controllers consider the underlying objective: protecting privacy.

The 'personal data' definition is currently the single trigger for applying all DPD requirements. But, on this definition, given scientific advances, almost all data could qualify as such. In considering whether particular information should trigger data protection obligations, the key factor ought to be, what is the realistic risk of identification? Only where risk of identification is sufficiently realistic (for example, 'more likely than not'),[145] should information be considered 'personal data'.

Where identification risk is remote or highly theoretical, for example due to technical measures taken, we suggest information should not be 'personal data'. In particular, encrypted data should be recognized as non-personal data in cloud computing, at least where strongly encrypted.[146] Clarification is also needed regarding anonymized data and anonymization and encryption procedures. The current law partly recognizes this ('means likely reasonably to be used', and WP136's reference to theoretical risks). However, in today's environment it may make sense for the threshold to be higher, based on realistic risks (such as 'more likely than not'). The boundary should be clearer.

Criteria for triggering the application of the DPD's requirements should be more nuanced, not 'all or nothing'. It may be appropriate to apply all of the DPD's requirements in some situations, but not in others. A better starting point might be to require explicit consideration of risk of harm to living individuals from intended processing, and its likely severity, balancing interests involved, with appropriate exemptions.

This would require an accountability-based approach, proportionate to circumstances including those of controllers, data subjects, and any processors.[147] More sensitive situations, with greater risk of resulting harm and/or greater severity of likely harm, would require greater precautions. That would accord with the European Commission's desire to require additional protections in certain circumstances, yet allow fewer or even no DPD requirements to be applied to strongly-encrypted data[148] held by someone without the key.

Could such a broad, flexible approach produce uncertainty and exacerbate lack of cross-EU harmonization? Arguably no more so than currently.[149] Indeed, it may better reflect practical realities in many member states.

142 Swedish Ministry of Justice, *Personal Data Protection—Information on the Personal Data Act* (2006).

143 Convention 108 for the Protection of Individuals with regard to Automatic Processing of Personal Data—on which the DPD was largely based.

144 Jean-Marc Dinant, 'Rapport sur les lacunes de la Convention no. 108 pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel face aux développements technologiques' T-PD-BUR(2010)09 (I) FINAL (Conseil de l'Europe 2010), 8.

145 The UK tribunal in *APGER* (n 38) did not think that 'appreciable risk' of identification was the statutory test; however, in deciding whether, on the facts, certain information was 'personal data', they concluded that its publication would not render individuals identifiable 'on the balance of probabilities'. Thus, it seems that they applied a 'more likely than not' test, in practice. *APGER* [129].

146 The European Commission supports developing 'technical commercial fundamentals' in cloud computing, including standards (eg APIs and data formats). Neelie Kroes, 'Towards a European Cloud Computing Strategy' (World Economic Forum Davos 27 January 2011) SPEECH/11/50. Encryption and security measures also need standardization, with a view to appropriate legal recognition of accepted standards.

147 We use 'accountability' as per the Canadian Personal Information Protection and Electronic Documents Act 2000 (PIPEDA). Views differ on 'accountability'. The A29WP in *Opinion 1/2010 on the principle of accountability, WP 173* (2010) seems to consider 'accountability' involves taking measures to enable compliance with data protection requirements and being able to demonstrate measures have been taken. However, PIPEDA's broader approach treats accountability as end-to-end controller responsibility (eg PIPEDA Sch 1, 4.1.3).

148 At least where that is reasonable, assuming industry standards would require stronger encryption for sensitive data. If encryption under recognized standards is broken, stronger encryption might be expected to be substituted within a reasonable time.

149 Also, given the recognized importance of harmonization and of doing more to foster a unified approach, if the Commission thinks fit the Data Protection Directive reforms could provide for guidance by the A29WP or Commission to bind member states and courts, and/or empower the A29WP to assess adequacy/consistency of national implementations and issue binding rulings.

Interestingly, the A29WP's response[150] to the European Commission's December 2009 DPD consultation did not mention the 'personal data' definition. Instead it concentrated, we believe rightly, on broader practical structural protections, such as privacy by design (PbD)[151] and accountability.

It has been suggested[152] that because, in today's environment, re-identification is increasingly easier and fully anonymizing personal data near impossible, 'the basic approach should be to reduce the collecting and even initial storing of personal data to the absolute minimum'.[153] PbD and privacy-enhancing technologies (PETs)[154] will assist data minimization,[155] and we consider them essential in the revised DPD. However, data minimization alone cannot protect personal data already 'out there'.[156] The processing of such personal data must still take account of risks to living individuals, and their likely severity.

What about sensitive data? Under current definitions, all data may potentially be sensitive, depending on context. Arguably, the 'personal data'/'sensitive data' distinction is no longer tenable. The European Commission is considering whether other categories should be 'sensitive data', for example genetic data; and will further clarify and harmonize conditions for processing sensitive data.[157] The ICO has pointed out[158] that a fixed list of categories can be problematic:[159] sensitivity can be subjective/cultural, set lists do not take account sufficiently of context and may even exclude data which individuals consider to be sensitive, while non-EU jurisdictions may have different lists, causing possible difficulties for multinationals. The ICO also suggested:

> [A] definition based on the concept that information is sensitive if its processing could have an especially adverse or discriminatory effect on particular individuals, groups of individuals or on society more widely. This definition might state that information is sensitive if the processing of that information would have the potential to cause individuals or [sic] significant damage or distress. Such an approach would allow for flexibility in different contexts, so that real protection is given where it matters most. In practice, it could mean that the current list of special data categories remains largely valid, but it would allow for personal data not currently in the list to be better protected, eg financial data or location data. Or, more radically, the distinctions between special categories and ordinary data could be removed from the new framework, with emphasis instead on the risk that particular processing poses in particular circumstances.[160]

This indicates possible support for a risk-based approach to personal data generally, under which 'sensitive data' as a special category may not be necessary—sensitivity of particular data in context being one factor affecting how they may or should be processed.

We suggest a two-stage, technologically-neutral, accountability-based[161] approach to address privacy concerns targeted by the 'personal data' concept:

1. *risk of identification*—appropriate technical and organizational measures should be taken to minimize identification risk. Only if the resulting risk is still sufficiently high, should data be considered 'personal data', triggering data protection obligations;
2. *risk of harm and likely extent*—the risk of harm and its likely severity should then be assessed, and appro-

150 *The Future of Privacy* (n 126).

151 Embedding privacy-protective features when designing technologies, policies, and practices, pioneered by Ontario's Information & Privacy Commissioner Dr Ann Cavoukian—NEC Company, Ltd. and Information and Privacy Commissioner, Ontario, Canada, *Modelling Cloud Computing Architecture Without Compromising Privacy: A Privacy by Design Approach* (Privacy by Design 2010). PbD is increasingly supported by regulators and legislators, eg 32nd International Conference of Data Protection and Privacy Commissioners, 'Privacy by Design Resolution' (27–29 October 2010, Jerusalem, Israel); *The Future of Privacy* (n 126); Communication (n 7).

152 LRDP Kantor and Centre for Public Reform, *Comparative study on different approaches to new privacy challenges, in particular in the light of technological developments—final report to European Commission* (European Commission 2010) [121].

153 Data minimization is an existing Principle—Art 6(1)(c). The suggestion was to focus on it primarily, or more.

154 PETs are not necessarily the same as PbD. Views differ on defining PETs. The ICO considers PETs are not limited to tools providing a degree of anonymity for individuals, but include any technology that exists to protect or enhance privacy, including facilitating individuals' access to their rights under the DPA—ICO, *Data Protection Technical Guidance Note: Privacy enhancing technologies (PETs)* (2006). London Economics, *Study on the economic benefits of privacy-enhancing technologies (PETs): Final Report to The European Commission DG Justice, Freedom and Security* (European Commission 2010) noted the complexity of the PETs concept, stating that security technologies are PETs if used to enhance privacy, but they can also be used in privacy-invasive applications. It suggested that more specific terminology, eg 'data protection tools', 'data minimisation tools', 'consent mechanisms', etc, was preferable in many cases.

155 Eg IBM's Idemix, Microsoft's technologies incorporating Credentica's U-Prove, the Information Card Foundation's Information Cards. Touch2ID pilots smartcards proving UK holders are of drinking age without revealing other data—Kim Cameron, 'Doing it right: Touch2Id' (*Identity Weblog*, 3 July 2010) <http://www.identityblog.com/?p=1142> last accessed 26 August 2011.

156 Case C-73/07 *Tietosuojavaltuutettu v Satakunnan Markkinaporssi and Satamedia Oy (Approximation of laws)* OJ C 44/6, 21.2.2009; [2008] ECR I-9831.

157 Communication (n 7), 2.1.6.

158 MoJ (n 128), 10.

159 The draft US Privacy Bill demonstrates the difficulties with defining 'sensitive personal data' using a list of categories, rather than by reference to impact, or potential impact, of processing on the individual—MoJ (n 128).

160 MoJ (n 128), 11.

161 For example, users remain accountable and should consider the risk of providers inadvertently deleting encrypted data, and take steps to protect data accordingly, eg by saving copies locally or with other providers.

priate measures taken regarding the personal data, with obligations being proportionate to risks.

Accordingly, if a controller has successfully implemented appropriate measures to minimize identification risk so that information is not considered 'personal data' (such as strong encryption), the risk of harm need not be addressed. However, if there is a sufficient risk of identification in specific circumstances, for example with pseudonymization or aggregation performed in certain ways, then risk of harm and its likely severity should be assessed, and appropriate measures taken.

In two situations, less restricted or even free processing of originally 'personal' data might be permissible:

1. where data are so depersonalized that they are no longer 'personal', such as through strongly encrypting full datasets; and
2. where data subjects intentionally make public their personal data (raising more difficult policy issues).

Account should be taken not just of what is done to data, but who does it: data subject, cloud user, or cloud provider.

Concluding remarks

We have advanced proposals for data protection laws to cater for cloud computing and other technological developments in a clearer, more balanced way.

The data protection regime should be more nuanced, proportionate, and flexible, based on an end-to-end accountability approach (rather than binary distinctions). The threshold inherent in the 'personal data' definition should be raised, basing it instead on realistic risk of identification. A spectrum of parties processing personal data should be recognized, having varying data protection obligations and liabilities, with risk of identification and risk of harm (and its likely severity) being the key determinants, and with appropriate exemptions. Such an approach should result in lighter or no regulation of cloud infrastructure providers, while reinforcing obligations of cloud providers who knowingly and actively process personal data, to handle such data appropriately.

The status of encrypted and anonymized data (and encryption and anonymization procedures) should be clarified so as not to deter their use as PETs. This could be done, for example, by stating that such procedures are not within the DPD, are not 'processing', or are authorized, or that fewer obligations apply to the resulting data. This would enable more data to fall outside the regulated sphere for 'personal data' in appropriate situations. In particular, we suggest that data strongly encrypted and secured to industry standards (including on key management) should not be considered 'personal data'. As regards anonymized data, it is important to clarify when anonymization may produce non-'personal data'. Likelihood of identification should be the main determinant. For example, information should be treated as 'personal data' where it 'more likely than not' would identify individuals, but not where the realistic risk of identification is insufficient. For sensitive data, a similar risk of harm approach should be considered, with definitions as suggested by the ICO. The position in relation to personal data manifestly made public should be clarified for non-sensitive as well as sensitive data, such as applying fewer data protection rules to such data.

Providers, especially infrastructure providers, should consider developing and implementing measures to minimize the likelihood of cloud services being regulated inappropriately by EU data protection laws; for example, by implementing encryption on the user's equipment using keys generated by and available only to the user. More transparency on sharding and other operational procedures would assist regulators to treat cloud services more appropriately, as would industry standards on matters such as encrypting data for cloud storage, including privacy by design. Emphasizing standards, while facilitating more flexible and pragmatic regulation of cloud ecosystem actors, should also help shift the regulatory focus back to protecting individuals.

The DPD was proposed in 1990 and adopted in 1995. Technologies, in particular Internet-related technologies, have evolved significantly since. It is time to make the DPD fit for the twenty-first century.

doi:10.1093/idpl/ipr018
Advance Access Publication 14 September 2011