Title

The (UK) Freedom of Information Act's disclosure process is broken: where do we go from here?

Henry Pearce, University of Portsmouth, Senior Lecturer in Law

Portsmouth Law School,

University of Portsmouth,

Richmond Building,

Portland Street,

Portsmouth,

Hampshire,

PO1 3DE

Email: henry.pearce@port.ac.uk

Abstract

This article builds on previous literature in the data protection and freedom and information field, which has argued that the "release and forget" disclosure model utilised by the Freedom of Information Act 2000 (FOIA) is unfit for purpose in the context of personal data that have been subject to a process of anonymisation, and that reform is necessary. Rather than outlining a detailed proposal for reform, the article intends to stoke debate in this area by highlighting a range of issues and factors that could help inform discussions regarding what shape any reform of the FOIA's disclosure model should take. The article argues that the notions of privacy and data protection by design, data licensing, risk, contextual controls, metadata, and privacy enhancing technologies, should all have a role to play in respect of improving how anonymised data are disclosed under the FOIA.

Key Words

Data Protection, Freedom of Information, Anonymisation, Data Protection by Design, Privacy by Design, Data Licensing, Risk, Metadata, Privacy Enhancing Technologies

Introduction

As has been argued elsewhere, due to advances in the field of anonymisation, the "release and forget" model of disclosure utilised by the Freedom of Information Act 2000 (FOIA) is unfit for purpose in the context of anonymised data and is in need of reform.[1] This article lays the groundwork for the construction of a new model for disclosing anonymised data under the FOIA, and considers what form such a model should take. Specifically, the article argues that a new approach to disclosing anonymised data under the FOIA, based on the notions of data protection by design, data licensing, and risk management, represents a promising way forward, and examines what issues and factors would need to be taken into account were such an approach to become a reality. The purpose of the article, therefore, is not to outline a detailed proposal for reform per se, but to highlight a range of important factors and ideas that could help inform reform in this area.

The article takes the following structure. First, the article starts by restating the nature of the problem faced, and the underlying rationales and purposes of legislation pertaining to freedom of information and data protection. This is done to highlight and reiterate why reform to the FOIA's "release and forget" approach to disclosing anonymised data is necessary, and to illuminate what any reform of this area of law should aim to achieve. This section of the article argues that the process for disclosing anonymised data under the FOIA requires remodelling so to incorporate control mechanisms that impose restrictions on post-disclosure uses of anonymised data. Second, the article examines the notions of privacy and data protection by design, and argues that an effective way of incorporating post-disclosure control mechanisms into the FOIA's disclosure process would be through the "building in" of various technical and organisational safeguards and procedures. The third, fourth and fifth sections consider how data licensing represents a promising example of a post-release control mechanism that could be "built in" to the FOIA disclosure process. Here, the article outlines licenses as a legal concept, before then considering their use in the context of information, data, and other intangible material. The article then argues that the introduction of a system of licensing based on the notion of risk has the potential to remedy some of the deficiencies inherent in the current FOIA's current "release and forget" approach to disclosure. The remaining sections of the article address a range of issues and factors that could help inform the development and design of a risk-based licensing approach to disclosing anonymised data under the FOIA. To this end, the sixth section of the article comprises of a detailed examination of the nature of risk-based regulation, its inherent issues and challenges, and how risk assessments relating to the disclosure of anonymised data as a part of any

[1] H Pearce and S Stalla-Bourdillon, 'Rethinking the "release and forget" ethos of the Freedom of Information Act 2000: Why developments in the field of anonymisation necessitate the development of a new approach to disclosing data' [2019] 10(1) European Journal of Law and Technology.

risk-based licensed disclosure approach could be undertaken. Here, a survey of other risk-based privacy and data protection methodologies is undertaken. From this, various principles and concepts vital for the establishment of any risk-based licensing system for anonymised data are derived and explained. Particularly, the article makes the argument that that risk assessments pertaining to the disclosure of anonymised data must consist of two questions relating to: 1) the probability of re-identification/de-anonymisation; and 2) the impact/severity of any harms that would stem from re-identification/anonymisation, were it to occur. Section seven then considers three other factors that could help inform and contribute to the development of a risk-based licensed disclosure model for use in the abovementioned context. Specifically, the section considers the possibility of incorporating additional contextual controls into the process for making FOI requests, and the possible role of metadata and privacy enhancing technologies. The article concludes with a summary of its findings and discussions.

1. Illuminating the nature of the problem

Previous contributions to the data protection law and policy literature have highlighted the nature of the problem in need of address.[2] The problem, in short, is as follows. The "release and forget" disclosure model currently utilised by the FOIA (i.e. a model of disclosure which imposes no post-release obligations on the releasers or recipients of disclosed information and data) is unfit for purpose in the context of anonymised data. Whilst personal data, as defined by the Data Protection Act 2018 (DPA) and the EU General Data Protection Regulation (GDPR),[3] are generally exempt from disclosure under the FOIA, anonymised data (i.e. personal data that have been subject to a process of anonymisation, and rendered anonymous) are not.[4] The logic for this approach being, that as anonymous data cannot be linked to any specific individuals, no harm can result from their processing. Operating under this premise, courts have started to order public authorities to disclosure data that are nominally anonymised data to the public.[5]

The "release and forget" approach to disclosure is premised on the belief that that anonymised data cannot be de-anonymised and, therefore, the processing of data of this sort cannot lead to re-

[2] *Ibid*

[3] Formally known as: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

[4] Personal data is defined by Section 3 Data Protection Act 2018 (DPA) as "any information relating to an identified or identifiable living individual". Section 40(2) FOIA 2000 then specifies that personal data will be exempt from disclosure under the FOIA if this would contravene any of the substantive data protection principles contained in Article 5 GDPR and Chapter 2 DPA 2018.

[5] See, for example: *Queen Mary University of London v (1) The Information Commissioner and (2) Alem Matthees EA/2015/0269*

identification and cannot be harmful. Ergo, the unfettered release of such data is no cause for concern, and subsequent uses of such data are undeserving of regulatory attention. This presumption, however, is false, and the prospect of anonymised data being de-anonymised and used for harmful purposes is often very real.[6] It is clearly inappropriate, therefore, for anonymised data to be disclosed on a "release and forget" basis. Put alternatively: the problem is that once data are released under the FOIA, there is nothing that can be done to limit or control harmful post-disclosure uses. Clearly, a new approach is necessary.[7] With this established, the question becomes "what should this new approach be?" Our starting point for answering this question is to examine the key rationales and purposes underpinning UK legislation pertaining to freedom of information legislation and data protection.

As set out elsewhere, the purpose of the FOIA is to heighten transparency and accountability, to allow individuals to utilise public sector data as a means of exercising economic and political rights, and to enable secondary uses of public sector data (e.g. for investigating corruption, scientific research etc.).[8] Conversely, the primary purpose of the DPA, and the GDPR, which the DPA 2018 largely concretises into UK law, is to protect individuals from harms that may stem from the processing of their personal data.[9] By considering the rationales and purposes behind these two different, but strongly intersecting, areas of law, we can discern that any reform of the FOIA must aim to give effect to, and strike a balance between, these two possibly competing positions. In other words, any reform of the FOIA's disclosure process should seek to facilitate the release of public sector data, where there is a public interest in the disclosure of that data, whilst, concurrently, ensuring high levels of data

[6] On this issue, see: L Rocher, J Hendrickx and Y de Montjoye, Y. 'Estimating the success of re-identifications in incomplete datasets using generative models' [2019] 10 Nature Communications; A Narayanan and V Shmatikov 'Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)' [2008] University of Texas at Austin 403-418; L Sweeney 'Simple Demographics Often Identify People Uniquely' [2000] Carnegie Mellon University, Data Privacy Working Paper 3; Y de Montjoye, et al. 'Unique in the Crowd: The privacy bounds of human mobility' [2013] Scientific Reports; P Ohm 'Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization' [2010] UCLA Law Review. See also: F Aldhouse 'Anonymisation of personal data – A missed opportunity for the European Commission' [2014] 30(4) Computer Law & Security Review.

[7] It should be noted that the advances in anonymisation have caused problems for the DPA and GDPR's dichotomous treatment of personal/anonymous data more generally. This topic, however, is beyond the scope of this paper.

[8] See: P Birkinshaw, Freedom of Information: The Law, the Practice, and the Ideal (Cambridge University Press 2010); J Ackerman and E Sandoval-Ballesteros 'The Global Explosion of Freedom of Information Laws' [2006] 58(1) Administrative Law Review; D Banisar 'Freedom of Information Around the World 2006: A Global Survey of Access to Government Information Laws' [2006] Privacy International.

[9] See: O Lynskey, The Foundations of EU Data Protection Law (Oxford University Press 2015); O Lynskey 'Data Protection and Freedom of Information: reconciling the irreconcilable?' [2011] Cambridge Law Journal; P Ticher, Data Protection vs. Freedom of Information (IT Governance Publishing 2008); M Turle 'Freedom of information and data protection law – A conflict or reconciliation?' [2007] 23(6) Computer Law & Security Review.

protection for any individuals who may be affected by such a disclosure. Both goals are laudable and worthy of pursuit.[10]

Accordingly, we must immediately rule out one possible solution to the abovementioned problems: the introduction of an absolute prohibition on the disclosure of anonymised data under the FOIA. Such a move would undoubtedly reduce the risk of anonymised data being de-anonymised and used for harmful purposes post-disclosure, ergo affording individuals greater levels of data protection. However, it would clearly be disproportionate, and come at the cost of debarring many FOI requests with legitimate objectives in which there may be a strong, or even overwhelming, public interest.[11] As both facilitating the release of public sector data in which there is a public interest, *and* ensuring high levels of data protection are goals that any reform of the FOIA's disclosure process should seek to achieve, the idea of adding anonymised data to the FOIA's already lengthy list of exemptions is a non-starter.[12] A more measured approach is required.

A better option would be to identify a solution that allowed public authorities to continue to disclose anonymised data, but in a way allowed them to exert a degree of control and influence over such data post-release, allowing them to manage any de-anonymisation challenges that arose over time. In other words, rather than continuing to operate under "release and forget" ethos, the FOIA should adopt a new approach to disclosing anonymised data that more closely embodied a "release and remember" ethos. This approach would entail the attachment of obligations and restrictions to anonymised data, and travelling with those data post-disclosure. This would allow the continued disclosure of anonymised data under the FOIA, and for all of the benefits associated with the

[10] Such an approach is in line with the European tradition of attempting to strike a balance between conflicting interests and legal values. See: R Wacks, Privacy and Media Freedom (Oxford University Press 2013) 258; F Cate 'Provincial Canadian Geographic Restrictions on Personal Data in the Public Sector' [2008] The Centre for Information Policy Leadership; C Kuner 'An International Legal Framework for Data Protection: Issues and Prospects' [2009] Computer Law & Security Review; K Aquilina 'Public Security versus privacy in technology law: A balancing act? [2010] 26(2) Computer Law & Security Review 130-143; C Reed, Computer Law (Oxford University Press 2011) 575-576; J Gilliam and T Monahan, Supervision: An Introduction to the Surveillance Society (University of Chicago Press 2013); A Lakhani 'Social Networking Sites and the Legal Profession: Balancing Benefits with Navigating Minefields' [2013] 29(2) Computer Law & Security Review.164-174; E Barendt, Privacy and freedom of speech in A Kenyon and M Richardson (eds) New dimensions privacy: Communications technologies, media practices and law (Cambridge University Press 2006); F Borgesius, M Eechoud and J Gray, J 'Open Data, Privacy and Fair Information Principles: Towards a Balancing Framework' [2016] Berkeley Technology Law Journal 2080-2081.

[11] For example, a request for information made for the purposes of investigating alleged corruption, or misappropriation of funds.

[12] The FOIA contains a long list of types of information that are exempt from its scope. Some critics have argued that the broad range of exemptions contained in the Act seriously undermine the purposes for which it was enacted. See: R Austin, The Freedom of Information Act 2000 – A Sheep in Wolf's Clothing? in J Jowell and D Oliver (eds), The Changing Constitution (Oxford University Press 2007); Msafiri 'Who wants freedom of information?' [2017] 33(5) Information Development; P Sikka 'Using freedom of information laws to frustrate accountability: Two case studies of UK banking frauds' [2017] 41(4) Accounting Forum.

disclosure of such data to be realised, whilst ensuring the protection of the data protection rights of any affected individuals. Such an approach, for instance, would allow for the disclosure of anonymised data, but with attached conditions that prohibited data processing activities likely to give rise to de-anonymisation and possible harms. A discussion of data protection by design is a useful starting point for considering how we might achieve this general objective.

2. Data protection by design

The term "privacy by design" (PBD) entered use around 2000, with the Workshop on Freedom and Privacy by Design at the Computers, Freedom and Privacy 2000 conference,[13] as well as being mentioned in a variety of other academic papers published around the same time.[14] Data protection by design (DPBD), an offshoot of PBD, focuses on information systems development, with the aim of ensuring that privacy and data protection-related interests are accounted for (i.e. "built in") during the lifecycle of such development.[15] The rationale behind this ethos is the belief that building data protection principles into the architecture of information systems will improve the principles' traction.[16]

In recent years, talk of DPBD has become a staple part of the data protection discourse, and recognition of its value has gradually increased. In 2010, for instance, the 32nd International Conference of Data Protection and Privacy Commissioners unanimously passed a resolution recognising PBD as an essential component of fundamental privacy protection.[17] The Article 29

[13] CFP2000, Conference on Computers, Freedom & Privacy (2000)
[14] See: M Veale, R Binns and J Ausloos 'When Data Protection by Design and Data Subject Rights Clash' [2018] *International Data Privacy Law*.
[15] As has been noted in the literature, there are many different strategies and tools that data controllers might deploy in pursuit of this objective. See: M Colesky, J Hoepman and C Hillen 'A Critical Analysis of Privacy Design Strategies' [2016] IEEE Security and Privacy Workshops 33-40; J van Rest et al, Designing Privacy-by-Design. in B Preneel and D Ikonomou (eds), Privacy Technologies and Policy. APF 2012. Lecture Notes in Computer Science (Springer 2014); D Le Métayer, Privacy by Design: A Matter of Choice. in S Gutwirth, Y Poullet, and P De Hert (eds), Data Protection in a Profiled World (Springer: 2010) 323-334; J Hoepman, Privacy by Design Strategies. In N Cuppens-Boulahia et al (eds), ICT Systems Security and Privacy Protection. SEC 2014. IFIP Advances in Information and Communication Technology (Springer 2014) 446-459. See also: D Mulligan and J King 'Bridging the Gap between Privacy and Design' [2012] 14(4) Journal of Constitutional Law 989-1034.
[16] Ibid., See also: A Cavoukian 'Privacy by Design in Law, Policy and Practice: A White Paper for Regulators, Decision-makers and Policy-makers' [2011] Information and Privacy Commissioner, Ontario, Canada; L Bygrave 'Data Protection by design and by default: Deciphering the EU's legislative requirements' [2017] 4(2) Oslo Law Review 105-120; M Hildebrandt and L Tielemans 'Data protection by design and technology neutral law' [2013] 29(5) Computer Law & Security Review 509-521.
[17] Ibid.

Working Party,[18] the US Federal Trade Commission,[19] the European Court of Human Rights,[20] and the Court of Justice of the European Union,[21] have more recently expressed similar sentiments.

Some observers have expressed doubts about the value of DPBD, [22] whilst others have gone further and suggested it may in fact be counterproductive in some instances, or may even create negative privacy impacts.[23] Others, however, have been more optimistic, and have highlighted how such initiatives have enjoyed successes in many different areas of application.[24] Irrespective of its success, or lack thereof, in other areas, it is clear that DPBD should play a significant role in addressing the abovementioned concerns associated with the FOIA's "release and forget" approach to disclosing anonymised data. There are at least three notable arguments that justify this position.

2.1. Data protection by design is proactive, not reactive

Privacy and data protection by design are characterised by their proactive and preventative, rather than reactive and remedial, approaches. As noted by Cavoukian, the "proactive not reactive, preventative not remedial" ethos is one of the foundational principles of PBD, and holds that privacy invasive events should be anticipated and prevented prior to their occurrence, rather than waiting for

---

[18] Ibid.
[19] Ibid.
[20] The ECtHR has historically made several judgments that have embraced the ideals of privacy/data protection by design. See, for instance, the decision of the Court in *I v Finland. Appl. No.20511/03, Judgment of 17 July 2008*, in which it was held that Finland was in breach of Art.8 ECHR due to a failure to implement technical/operational measures as a means of ensuring the confidentiality of patient medical data in hospitals.
[21] Though the CJEU has not ruled directly on the matter of privacy/data protection by design, in the *Google Spain* case, by rejecting Google's argument that its search engine operations were value neutral robotic applications of algorithms outside the scope of data protection law, the CJEU compelled Google (and other search engine providers) to reconfigure systemic aspects of those operations so that they would be more privacy/data protection friendly. Case C-131/12 *Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González,* EU:C:2014:317. On this issue, see: L Bygrave (n 16).
[22] B Koops and R Leenes 'Privacy regulation cannot be hardcoded. A critical comment on the 'privacy by design' provision in data protection law' [2014] 28(3) International Review of Law, Computers and Technology 159-171; S Spiekermann 'The challenges of privacy by design' [2012] 55(7) Communications of the ACM 38-40; S Shapiro 'Privacy by design: moving from art to practice' [2010] 53(10) Communications of the ACM 27-29.
[23] It has been argued, for instance, that certain confidentiality-focused data protection by design strategies used by large data controllers leave data re-identifiable by capable adversaries while heavily limited controllers' ability to provide data subject rights, such as access, erasure and objection. See: M Veale, R Binns, and J Ausloos (n 14). See also: I Brown 'Britain's Smart meter programme: A case study in privacy by design' [2014] 28(2) International Review of Law Computers & Technology 172-184.
[24] Electronic health cards, electronic ID cards, and electronic proof of earnings, for instance, have been identified as areas of application in which privacy by design has been highly beneficial. See: P Schaar 'Privacy by Design' [2010] 3(2) Identity in the Information Society 267-274. Remote healthcare technologies and big data analytics are other areas of application in which the potential for privacy by design has been mooted. See: A Cavoukian et al 'Remote home health care technologies: how to ensure privacy? Build it in: Privacy by Design' [2010] 3(2) *Identity in the Information Society* 363-378; A Cavoukian and J Jonas 'Privacy by Design in the Age of Big Data' [2012] Available at: https://jeffjonas.typepad.com/Privacy-by-Design-in-the-Era-of-Big-Data.pdf. See also: D Davies 'Why *Privacy by Design* is the next crucial step for privacy protection' [2010] Available at: http://i-comp.org/wp-content/uploads/2013/07/privacy-by-design.pdf

them to materialise and attempting to redress any resultant harms post-occurrence.[25] At the heart of privacy and data protection by design, therefore, is the notion of "prevention is better than cure".

This approach is highly salient in the context of the challenges arising in conjunction with the FOIA's "release and forget" disclosure model. As noted above,[26] a major problem with the FOIA's "release and forget" approach is the fact that data released pursuant of an FOI request will essentially become open data, and anonymised data released pursuant of an FOI request might be de-anonymised post-disclosure. When, how, and even whether, a problem of this sort might arise will often be unclear at the point of disclosure and will be difficult to predict to predict in advance. If, however, an issue is to arise, and de-anonymisation occurs, the releasing public authority can do little under the current legislative framework to manage and/or mitigate any resultant problems. The FOIA contains no provisions for dealing with any such situations. Whilst individuals may invoke data protection rights and rules, or possibly other tortious remedies, as a means of redressing harms, enforcement may prove difficult. Once such data are "out there", possibly in the hands of parties in other jurisdictions, it will be impossible to "get back" the data, and the application of post-disclosure legal rules will often be problematic and ineffective.

Given the difficulties of applying and enforcing ex-post legal rules and remedies as a means of resolving problems arising in conjunction with disclosing anonymised data on a release and forget basis, the shortcomings inherent in the FOIA's disclosure process represent a clear example of a situation in which the "prevention is better than cure" mantra rings particularly true. They are problems for which the best "solution" is the prevention of their occurrence. As privacy and data protection by design have a "prevention is better than cure" ethos at their core, there is a clear logical justification for their deployment in this context.

2.2. Data protection by design is concerned with ensuring the full functionality of data through the balancing of different, possibly competing, interests

As outlined above,[27] any prospective reform of the FOIA's disclosure model should strike a balance between the possibly competing objectives of opening up public-sector data, and the need to protect the data protection interests of affected individuals. In other words, we should not afford either of these objectives preferential treatment over the other. To this end, the inevitable dual objectives at

[25] A Cavoukian 'Privacy by Design: The 7 Foundational Principles' [2006] Available at: https://www.ipc.on.ca/wp-content/uploads/Resources/pbd-implement-7found-principles.pdf
[26] See Section 1.
[27] Ibid.

the heart of any possible reform of the FOIA's disclosure model closely align with another of privacy and data protection by design's foundational principles: the principle of full functionality.

As noted by Cavoukian:

> "*Privacy by Design seeks to accommodate all legitimate interests and objectives in a positive-sum "win-win" manner, not through a data, zero-sum approach, where unnecessary trade-offs are made. Privacy by Design avoids the pretence of false dichotomies, such as privacy vs. security, demonstrating that it is possible, and far more desirable, to have both*."[28]

In a similar vein, when discussing possible reform of the FOIA we must avoid falling into the trap of false dichotomies, such as freedom of information vs. data protection, or transparency and accountability of public authorities vs. individual rights. As demonstrated below, it is possible to have both.[29] As privacy and data protection by design are committed to reconciling and giving effect to seemingly competing objectives, so to ensure maximum functionality of data, this represents another clear logical justification for their deployment in this context.

2.3. Data protection by design is concerned with "full lifecycle protection"

Another of PBD's foundational principles that is also highly relevant in the immediate context is that of "full lifecycle protection". Cavoukian again provides an explanation:

> "*Privacy by Design, having been embedded into the system prior to the first element of information being collected, extends securely throughout the entire lifecycle of the data involved – strong security measures are essential to privacy, from start to finish. This ensures that all data are securely retained, and then securely destroyed at the end of the process, in a timely fashion. Thus, Privacy by Design ensures cradle to the grave, secure lifecycle management of information, end-to-end*." [30]

From this, we can see that a fundamental aspect of privacy and data protection by design is a desire to provide a continuous and dynamic approach to matters of privacy and data protection, rather than treating them as one-off events. As highlighted above,[31] the fact that the FOIA's release and forget model takes a one-off, static approach to disclosure is *precisely* why it is unfit for purpose in the

[28] A Cavoukian (n 25).
[29] This is a position that has also been adopted by other observers. See, for instance: F Wu 'Defining Privacy and Utility in Datasets' [2012] *University of Colorado Law Review*.
[30] A Cavoukian (n 25).
[31] See Section 1.

context of disclosing anonymised data. In other words, the biggest problem with the current release and forget approach is that it does *not* offer full lifecycle protection.

The challenge of anonymising data is contextual and dynamic (i.e. not static). As a result, whatever solution we adopt as a means of addressing the shortcomings of the FOIA's "release and forget" disclosure model, ensuring full lifecycle protection for any anonymised data disclosed will be vitally important. Any proposed reform that failed to appreciate this nuance would inevitably fall short of solving the most serious of the problems outlined previously. Accordingly, the fact that DPBD is specifically concerned with establishing long-term full lifecycle solutions to challenges stemming from the processing of data (including disclosure), the reform of the FOIA's disclosure model is a perfect setting for its utilisation.

2.4. Discussion

It is clear that privacy and data protection by design are suitable for deployment in the context of reforming the FOIA's disclosure model. Privacy and data protection by design are especially useful in situations where a problem is best solved by prevention rather than cure, where individual rights must be balanced against other competing interests, and where a problem is complex, dynamic, multifaceted and insusceptible to ex-post legal rules and remedies. The need to reform the FOIA's release and forget disclosure model is a regulatory challenge possessing all these characteristics. Accordingly, if, as argued above,[32] the goal of such reform is to allow public authorities to retain a degree of post-disclosure control over anonymised data, the possibility of "building in" post-disclosure mechanisms and controls to the FOIA's disclosure process represents a logical avenue for the achievement of this objective. The next question, therefore, is what form should these mechanisms and controls take? One notable example of a mechanism that could be "built in" to the FOIA's disclosure process as a means of shifting to a "release and remember" ethos is a system of data licensing.

3. Data Licences: mechanisms for post-disclosure control of data

The term "licence" refers to a situation in which one party gives formal authority to another, via a legal instrument, to undertake an action that would otherwise be unlawful, or would infringe rights held by the first party.[33] Common examples of this include licences that permit individuals to drive motor vehicles on public highways, sell intoxicating liquor, or enter or occupy an area of land. As

[32] See Section 1.

[33] See: A Ball 'How to License Research Data' [2014] Digital Curation Centre. See also: A Martin, Oxford Dictionary of Law (Oxford University Press 2008) 315.

alluded to above,[34] however, licences can also be used to impose conditions and/or restrictions on the uses, re-uses, and sharing of information, data, and other intangible material. According to the UK Licensing Framework:

> "*A licence is a legal document giving permission to use information*…" and is considered "…*a mechanism that gives people and organisations permission to re-use information and other material… A licence should also provide clarity as to what users and re-users are permitted to do and whether there are any restrictions on the extent of that permission*."[35]

Similarly, in relation to their possible applicability to information and data, Davies defines a licence as an instrument setting out:

> "…*explicitly what someone who accesses a dataset can do with it […] and without an explicit licence, a user does not know if they have the legal permissions to share data further, to combine it with other data, or to build a commercial service off the back of a dataset*."[36]

Licences, therefore, allow their holder to maintain and exercise control over information and data, even after agreeing to share them with another party, by way of attaching conditions and rules to those information or data. If these rules or conditions are breached, the licensor will have legal recourse against the licensee. To illustrate how licences pertaining to information, data, and other intangible material operate in practice, what follows is a survey of some notable licences of this sort, namely: The Creative Commons licensing system, the UK Open Government Licence, and the Personal Data Licence of Land Information New Zealand. These examples were chosen not only as a means of highlighting the different types of licensing obligations that can be attached to data and information, but to showcase the different types of data and information to which licences can be attached.

3.1. Creative Commons

Creative Commons (CC) is a non-profit organisation that provides licensing tools to creators who wish to exercise their copyright rights in a way that encourages sharing and creative re-uses of their artistic and intellectual works.[37] The main purpose of CC parallels that of the free software movement, which seeks to use copyright to authorise, rather than inhibit, the copying, distribution, modification and re-use of software and other works.[38] The CC licensing system is comprised of a spectrum of licences that

[34] See Sections 1 and 2.
[35] UK Government Licensing Framework (2013) 10-20.
[36] T Davies 'Ten Building Blocks of an Open Data Initiative' [2012] Open Data Impacts.
[37] Creative Commons, 'About The Licences' (*Creative Commons*) <https://creativecommons.org/licenses/> accessed 6 June 2020.
[38] S Dusollier 'The Master's Tools v. The Master's House: Creative Commons v. Copyright' [2006] 29 Columbia Journal or Law & Arts 271-293.

are attachable to an individual's intellectual works. The licences contain obligations and rules regarding the future uses and treatment of these works by other parties, and are linkable via the Web.[39] It is the most widely used licence type for open data, and one which:

> "…*provides authors with a way of formalising their legal right to offer, in effect, open access to their work*."[40]

The six licence variants that make up the CC system are as follows:

> **Attribution (CC BY):** The most open of all the CC licensing options. It states that the data to which it is attached may be used for any purpose. Individuals may add to and manipulate the data, and offer derivations of the data for commercial sale. The only obligation imposed on an individual making use of the data is that they acknowledge the source of the data.
>
> **Attribution – Share Alike (CC BY-SA):** Similar to the Attribution licence, this variant also allows users of data to do whatever they wish with data to which it is attached so long as they acknowledge the data's source. One additional obligation imposed by this variant, however, is that if the user wishes to pass on any derivations they have made or taken from the data they must do so on the same terms which they received the data.
>
> **Attribution – No Derivatives (CC BY-ND):** This variant allows for a user to redistribute data, but prevents them from modifying, manipulating, or adding to the data. The user is permitted to use the data for commercial purposes, or sell the data, so long as the source of the data is acknowledged.
>
> **Attribution – Non-Commercial (CC BY-NC):** This variant allows users to manipulate and add to the original data, and to pass the data on, but only on a non-commercial basis. When passing the data on the user is required to acknowledge the source of the data. There is no requirement that any derivative works must be licensed on the same terms.
>
> **Attribution – Non-Commercial – Share Alike (CC BY-NC SA):** Under this variant, users are entitled to do what they wish with the data, other than use the data for any commercial purposes. They may pass on the data and any derivative works, provided they acknowledge

[39] M Carroll 'Creative Commons and the New Intermediaries' [2006] Michigan State Law Review.
[40] S Chignard 'A Brief History of Open Data' [2013] Paris Tech Review 2.

the source of the data and make it clear that their work is being passed on under the same licence as the original data.

Attribution – Non-Commercial – No Derivatives (CC BY-NC-ND): The most restrictive of all the CC licences. It allows users to access the data to which the licence is attached, and use and share those data, so long as they acknowledge the original source. Users are prohibited, however, from changing the data or using the data for any commercial purposes.

From this overview, we can see the licences in the CC hierarchy range from allowing an artistic or intellectual work to be released publicly with no rights reserved, to prohibiting them from being used for any commercial or derivative purposes, with those in between allowing for different degrees of future uses. The stringency of the licences, therefore, operate on a sliding scale. Despite its prominence and widespread usage, however, the CC licensing system is occasionally criticised for being overly confusing to users. Many creators often do not understand which licence would be the best one to use in relation to their intellectual works, and those bound by whatever licence is chosen often do not understand what it is required for compliance.[41]

3.2. The UK Open Government Licence

The UK Open Government Licence (OGL) was developed by the National Archives to enable public sector organisations to licence the use and re-use of information they hold, under a common set of standards. The OGL is utilised by numerous government departments and public authorities, including the Ministry of Defence, HM Land Registry, and local councils. As noted elsewhere, the creation of the OGL was influenced by the enactment of the EU Directive on the re-use of public sector information,[42] and its incorporation into UK law via the Re-use of Public Sector Information Regulations 2015.[43] The licence itself encourages recipients of data to:

> "*…use and re-use the Information that is available under this licence…freely and flexibly, with only a few conditions*."[44]

To this end, the OGL grants users a worldwide, royalty-free, perpetual, non-exclusive licence to copy, publish, distribute and transmit any information or data released under its terms, adapt those data,

[41] Z Katz 'Pitfalls of Open Licensing: An Analysis of Creative Commons Licensing' [2006] 46(3) IDEA 393.
[42] Formally known as Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information OJL 345
[43] See: J Attard et al. 'A systematic review of open government data initiatives' [2015] 32(4) Government Information Quarterly 399-418.
[44] Open Government Licence for public sector information. Available at: http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

or exploit those data for commercial purposes (e.g. by combining the data with other information, or by including the data in a commercial product or application).

The OGL also imposes some requirements and restrictions on data released under its terms. Notably, any information released under the OGL is subject to copyright and database rights, and any subsequent uses of data released under it must adhere to the relevant law. Additionally, users are required to acknowledge the source of any information or data used in any product or application they develop from such data, either by including or linking to an attribution statement specified by the original information provider and, where possible, including a link to the OGL itself. If the original information provider has no attribution statement, the user of OGL data must state that their use of the data contains public sector information licensed under the OGL.[45]

The OGL does not cover certain types of information or data. Namely:

- Personal data (including identity documents such as passports);
- Information that has not been accessed by way of publication or disclosure under information access legislation (including the FOIA);
- Departmental or public sector organisation logos, crests, military insignias, and the Royal Arms;
- Third party rights over which public sector organisations do not have authorisation; and
- Other intellectual property rights (e.g. patents and trademarks etc.)[46]

As with CC, there is no requirement for individuals to register or apply to use the OGL, they must simply ensure their use of any information issued under it complies with the licence terms. If a user breaches any of the abovementioned terms and conditions, any of the rights granted under the licence automatically rescind, rendering further usage of the data to which it applies unlawful.

3.3. The LINZ Licence for Personal Data

One notable licensing scheme pertaining to matters involving personal data is that used by Land Information New Zealand (LINZ), New Zealand's government department responsible for land titles, geodetic and cadastral survey systems, topographic information, hydrographic information, and managing Crown property.[47] The LINZ Data Service (LDS) holds more than forty groupings of data on New Zealand's land and sea. Notably, the LDS holds: topographic data used to create maps and hydrographic information used to create marine charts, property from place names and addresses to

[45] Ibid.
[46] Ibid.
[47] Land Information New Zealand, 'LINZ Data Service' <https://www.data.govt.nz/use-data/showcase/linz-data-service-lds-dec-2014/> accessed 6 June 2020

boundaries and land ownership, crown pastoral land leases and geodetic systems, and aerial images which provide an accurate photographic picture of the land's surface.

Most of the data released through the LDS is available under a CC Attribution licence, so to engender their widest possible re-use. However, as acknowledged by LINZ, it is not appropriate to disclose all its data on these terms, as much of the data in its possession, relating to property ownership and similar, will contain the personal data of individuals. To strike a balance between the desire to release its data and ensuring an adequate standard of data protection for individuals whose data are contained within LINZ's records, any party seeking to obtain information containing individuals' personal data must accept a more restrictive licence: the LINZ Licence for Personal Data. This licence allows the use of personal data, subject to terms and conditions, so long as such uses do not infringe copyright rules or any other applicable laws.[48] In addition to specifying that any recipient of information containing personal data must comply with the Privacy Act 1993, New Zealand's main piece of legislation in the privacy and data protection field, the licence specifies that all such persons must:

- Take reasonable and appropriate security safeguards to ensure the data are not misused and/or provided to any other parties who have not agreed to the terms of the licence;
- Not allow the data to be indexed by any publicly accessible online search engine and ensure that any parties to whom the data are provided do the same; and
- Not use the data for any form of targeted marketing.

The licence also gives LINZ post-disclosure power to compel any recipients of data received under the licence to perform certain actions on the data. Specifically, the licence states that within five business days of LINZ making a request, any recipient of the data under the licence must:

- Amend or delete the data, or any part of the data;
- Ensure that all third parties with whom the data have been shared, regardless of whether the data were shared in an altered or modified form, do the same; and
- Provide evidence to LINZ that the above has been done.[49]

As with the OGL, breach of any of the licence terms will result in the licence being automatically rescinded, rendering further usage of the data to which it applies unlawful.

4. Data licensing: a way of incorporating data protection by design into the FOIA disclosure process?

[48] Land Information New Zealand, 'LINZ Licence for Personal Data' <https://www.linz.govt.nz/data/licensing-and-using-data/linz-licence-for-personal-data> accessed 6 June 2020
[49] Ibid.

The previous section showed how data licensing can work in practice (i.e. how licences can act as a form of post-release/disclosure control in relation to data and information), the forms such licences can take (e.g. licences that require attribution, licences that prevent certain re-uses of data, licences that allow the licensor to exert post-disclosure control over data etc.), and the different types of data and information to which licences can be applied (e.g. artistic and intellectual works, information and data compiled by public sector organisations through the performance of their tasks, and even individuals' personal data). Though issues can arise in relation to the enforcement of such licences, there is no doubting their prominence as one of the primary means by which post-disclosure uses, and re-uses, of information and data are controlled. This section examines the possibility of using data licences in relation to anonymised data disclosed under the FOIA. The argument presented below, is that not only would the adoption of a data-licensing scheme in this context be beneficial, but also that the adoption of such a system would necessarily require the development of a bespoke FOI licensing regime for anonymised data.

4.1. Why licensing the disclosure of anonymised data under the FOIA makes sense

As set out above,[50] the purpose of the FOIA is to allow individuals to access public sector data and use those data for purposes in which there is a public interest. Disclosure of information under the FOIA, however, particularly anonymised personal data, may give rise to data protection concerns. Accordingly, any possible reform to the FOIA's data disclosure process should seek to give effect to the dual objectives of maximising the benefits of opening up data held by public sector organisations and, concurrently, ensuring as high a level of data protection as possible for any individuals' whose details are contained in public sector datasets. As highlighted in the previous section, data licenses exist precisely for use in this sort of situation. Not only can licences impose restrictions regarding data usage, but they can also communicate permissions regarding how those data *can* be used. To this end, the dual functionality of data licences aligns closely with the dual objectives behind the proposed reform of the FOIA outlined above. The negative limb of a data licence (i.e. terms specifying how data "must *not*" be used) could be used, for instance, to prevent anonymised data from being processed in a way that may increase the likelihood re-identification and subsequent abuses. This would allow public authorities who had released anonymised data to enforce the terms of such licences against any recipient parties, and allow for the imposition of sanctions in the event of non-compliance.

At this point, it is worth noting that the use of criminal law may be a useful way of achieving the same result (i.e. criminalising certain uses of data). Section 171 of the Data Protection Act 2018, for instance, introduces a new criminal offence that will be committed when a defendant either knowingly or

[50] See Section 1.

recklessly re-identifies anonymised data without the consent of the data controller responsible for the anonymisation. This approach, however, is of limited utility. Whilst criminalisation communicates to users that certain data processing activities are *not* permissible, it does not communicate those that *are* permissible. Furthermore, it is far from implausible that many types of re-identification/de-anonymisation could be achieved non-intentionally or non-recklessly, and not be caught by the offence.[51] In any event, the possibility of using criminal law as a means of resolving data protection matters is, in the best of scenarios, inherently problematic. The use of criminal law, for instance, requires clear and precise legal norms that are often not available in data protection. Data protection crimes will also have to compete with "normal" crimes (e.g. theft, assault, criminal damage etc.) for enforcement resources, and it is doubtful that the police and Crown Prosecution Service would prioritise them. A lack of technological expertise on the part of law enforcement bodies and the judiciary would also likely lead to a muddled approach, which in turn would generate unsatisfactory results.[52]

Conversely, using licences as a means of protecting individuals from harms associated with contemporary data-handling practices, such as de-anonymisation, is a possibility that has been identified as promising and deserving of further investigation. It has, however, hitherto not been meaningfully explored in the FOI context.[53] As highlighted by the Article 29 Working Party, the use of licences in respect of data held by public sector authorities may have considerable potential to assist with the enforcement of various aspects of data protection law, particularly the principle of purpose limitation.[54] Such an approach would also coalesce neatly with the growing rhetoric for affording individuals greater control over their personal data, an idea the literature has considered extensively, and one that has become a prominent policy objective for European lawmakers.[55]

---

[51] On the definitions of "intention" and "reckless" for the purposes of English criminal law, see: *R v Woolin* [1999] AC 82 and *R v Cunningham* [1957] 2 QB 396.

[52] On this issue, see: R Brownsword and M Goodwin, Law and the Technologies of the Twenty-First Century (Cambridge University Press 2012)

[53] See, for example: E Verheul et al. 'Polymorphic Encryption and Pseudonymisation for Personal Healthcare: A Whitepaper [2016] Institute for Computing and Information Sciences, Radboud University Nijmegen; A Popescu et al, Increasing Transparency and Privacy for Online Social Network Users – USEMP Value Model, Scoring Framework and Legal. in B Berendt et al (eds), Privacy Technologies and Privacy. APF 2015. Lecture Notes in Computer Science (Springer 2015) 38-59; Y Joung and S Cha, Online Personal Data Licensing: Regulating Abuse of Personal Data in Cyberspace. in H Sasaki (ed), Intellectual Property Protection for Multimedia Information Technology (IGI Global 2007) 165-185; Y Joung, Y. et al. 'On personal data license design and negotiation' [2005] Computer Software and Applications Conference, 2005. COMPSAC 2005. 29th Annual International; F Borgesius, M van Eechoud and J Gray (n 10).

[54] Article 29 Data Protection Working Party (2013) Opinion 03/2013 on purpose limitation, 00569/13/En WP203, 18, 58-50.

[55] On this issue, see: C Lazaro and D Le Métayer 'Control over personal data: True remedy or fairy tale?' [2015] 12(1) SCRIPT*ed*; H Pearce 'Could the doctrine of moral rights be used as a basis for understanding the notion of control within data protection law?' [2018] 27(2) Information & Communications Technology Law 312-335.

The positive limb of a data license (i.e. terms specifying how data "*can*" be used), could also be crucial to the achievement of maximising the benefits linked to opening up public sector data (i.e. transparency, accountability and secondary uses of data). To explain this point, it is first useful to provide context. At present, one of the biggest barriers to achieving the benefits associated with opening up public sector data is a lack of clarity. Empirical research has shown that creating and maintaining an environment in which individuals can understand their rights and obligations in respect of data released by public authorities will be critical if such benefits are to be realised.[56] Nevertheless, a lack of clarity often leaves individuals unsure not only of what they are *prohibited* from doing with public sector data, but of what they are *permitted* to do.[57] Not only is the absence of clarity one of the biggest hindrances to transparency, accountability and re-using data, however, but it can often have the effect of actively discouraging individuals from seeking to obtain public sector data in the first place.[58] This lack of clarity, therefore, seriously undermines the purposes for which the FOIA, and other pieces of FOI legislation worldwide, have been enacted. The construction and development of a licensing scheme, to be deployed in the context of anonymised data disclosed under the FOIA, would help to add clarity in respect of precisely what individuals would be entitled to do with data they received pursuant of an FOI request. In so doing, it could remove some of the existing barriers that stand in the way of the achievement of the FOIA's underlying goals.

As highlighted by the above survey of data licences, data licensing systems are already used in the context of other public authority data, particularly in relation to data released as part of open data initiatives. As data licences are already widely accepted as being appropriate for use in the context of open data initiatives, there is no obvious reason for why they would not also be suitable for use in the context of anonymised data released under the FOIA. Moreover, as has been remarked elsewhere, the development of a data licensing system in an FOI context should help with the construction of a

[56] M Khayyat and F Bannister 'Open Data Licensing: More than meets the eye' [2015] 20(4) Information Polity 231-252. See also: N Korn and C Oppenheim 'Licensing Open Data: A Practical Guide' [2011] JISC.
[57] Ibid.
[58] See: P Miller, R Styles and T Heath 'Open data commons, a license for open data' [2008] Proceedings of the 1st Workshop about Linked Data on the Web (LDOW2008); I Ermilov and T Pellegrini 'Data licensing on the cloud: empirical insights and implications for linked data' [2015] Proceedings of the 11th International Conference on Semantic Systems 153-156; T Pellegrini 'Integrating Linked Data into the Content Value Chain – A Review of News-related Standards, Methodologies and Licensing Requirements' [2012] Proceedings of the 8th International Conference on Semantic Systems 94-102; R Hosking and M Gahegan, The Effects of Licensing on Open Data: Computing a Measure of Health for Our Scholarly Record. In: H Alani et al. (eds), The Semantic Web – ISWC 2013. ISWC 2013. Lecture Notes in Computer Science (Springer 2013); S Villata. and F Gandon 'Licences Compatibility and Composition in the Web of Data' [2012] Third International Workshop on Consuming Linked Data (COLD2012); Q Groom et al. 'The importance of open data for invasive alien species research, policy and management' [2015] 6(2) *Management of Biological Invasions* 119-125; S Leucci 'Preliminary Notes on Open Data Licensing [2014] *Journal of Open Access to Law*.

more stable data ecosystem, which would be beneficial to both private and public sectors, as well as individuals.[59]

4.2. The limitations of existing licensing models

As set out in the preceding sections, there are clear indications that the deployment of a system of data licensing for anonymised data released under the FOIA could be beneficial. However, for several reasons it is doubtful that current licensing models, such as the prominent types outlined above, are suitable for use in this context. Primarily, the subject matter at the heart of a licensing scheme pertaining to anonymised data disclosed under the FOIA (i.e. personal data that have been subject to a process of anonymisation) would necessarily be significantly and quantitatively different from that at the heart of many prominent data licensing models. As explained above, for instance, the CC licensing system allows individuals to exercise intellectual property rights in relation to their artistic or intellectual works. The UK OGL does similarly, and grants users a range of rights in relation to information and data in which the releasing party has either copyright or database rights. To this end, it is important to note that the theoretical and conceptual justifications for allowing individuals to exert rights in respect of their intellectual and artistic works are rooted in completely different ground to those corresponding to an individual's personal data. As remarked elsewhere, for instance, copyright, database rights, and other forms of intellectual property rights, are legal mechanisms designed primarily to protect an individual's economic or pecuniary interests.[60] These rights only arise following the creation of a "work" capable of constituting a type of intangible property. CC and the OGL both operate on this premise. Personal data, and by extension anonymised data, however, cannot be described in this way. Such data are not "created", and the law does not formally recognise them as a form of intangible "property".[61] Instead, an individual's data protection rights are justified by way of reference to fundamental rights and the notion of human dignity, as opposed to economic or pecuniary interests.[62] These conceptual differences alone raise a number of difficult questions that

[59] See: A Williams, J Willbanks and S Ekins 'Why Open Drug Discovery Needs Four Simple Rules for Licensing Data and Models' [2012] 8(9) PLoS Computational Biology.
[60] See: L Bently and B Sherman, Intellectual Property Law (Oxford University Press 2009) 1.
[61] It is worth noting, however, that whilst the orthodox view is that the law does not, and should not, treat personal data as property, there is an ongoing debate in respect of this position. For an overview of this debate, see: H Pearce 'Personality, property and other provocations: exploring the conceptual muddle of data protection rights under EU law' [2018] 4(2) European Data Protection Law Review 190-208.
[62] See: H Pearce 'Could the doctrine of moral rights be used as a basis for understanding the notion of control within data protection law?' [2018] 27(2) Information & Communications Technology Law 133-165. It has been noted, however, that the system of damages outlined in the GDPR does imply that harms stemming from the errant processing of personal data may potentially be pecuniary in nature, and that redress for any such harms can be measured in economic terms. See: J Victor 'The EU General Data Protection Regulation: Toward a Property Regime for Protecting Data Privacy' [2013] Yale Law Journal.

would require address were attempts made to utilise licences in the CC or OGL mould for anonymised data disclosed under the FOIA.

Conceptual issues aside, however, there would also be various practical limitations to the use of these approaches in this context. Both the terms of the various CC licences and the OGL are limited to constraining information and data in respect of certain economic activities or uses. As highlighted previously, however, the de-anonymisation of anonymised data, and any subsequent harms, could quite conceivably stem from data processing activities with non-economic purposes. Releasing anonymised data under the most restrictive CC licence or the OGL, therefore, would not sufficiently protect individuals from the full catalogue of harms that may stem from de-anonymisation, thereby further highlighting how these licence types would be ill-suited for use in this context.

The LINZ Licence for Personal Data suffers from none of the abovementioned conceptual difficulties. It exists specifically to ensure adequate levels of data protection for individuals, and provides a tangible example of how licences can be constructed for, and applied to, individuals' personal data. Once again, however, there are reasons to doubt it, or any similarly constructed licence, would be suitable for deployment in the context of anonymised data disclosed under the FOIA. Specifically, whilst the LINZ licence outlines a number of conditions attached to the processing of individuals' personal data, it operates on a one-size-fits-all basis. The terms of the licence apply to all personal data in the same way, regardless of their sensitivity, nature, content, or character. For this reason, such a licence would be too inflexible to be suitable for use in the context of the disclosure of anonymised data. As mentioned above, and demonstrated elsewhere, for instance, anonymisation and risks of re-identification are highly context-dependant.[63] Accordingly, any licensing scheme designed to counteract the problems associated with de-anonymisation risk must constructed in such a way that encompasses this important nuance. To this end, whilst the LINZ licence, or another licence in a similar vein, would not be extensive nor multifaceted enough to provide a solution to the problems that any reform to the FOIA's disclosure process would need to solve.

4.3. A bespoke FOI licensing scheme for disclosing anonymised data under the FOIA?

As argued above,[64] a system of data licensing represents a potentially promising way of correcting problems inherent in the FOIA's disclosure. However, existing data licensing models and approaches are evidently unsuitable for deployment in this context. Ergo, the construction of a novel and bespoke

[63] See: M Elliot et al. 'The Anonymisation Decision-making Framework' [2016] UKAN; M Elliot et al. 'Functional anonymisation: Personal data and the data environment' [2018] 34(2) Computer Law & Security Review; S Stalla-Bourdillon and A Knight 'Anonymous Data v. Personal Data – A False Debate: An EU Perspective on Anonymization, Pseudonymization and Personal Data' [2017] Wisconsin International Law Journal.
[64] See Sections 1 and 2.

data licensing system, designed specifically for use in the context of disclosing anonymised data under the FOIA, would be required if any data licensing solution were to have any realistic prospect of success. As also mentioned above, though some observers have discussed the possibility of developing data licensing schemes in the context of both public-sector data, and personal data more generally, to date the possibility of "building in" a licensing system for anonymised data into FOIA's disclosure process has not been explicitly considered in the literature. Nevertheless, by drawing on the existing data licensing schemes considered above, it is not difficult to see how we might use some of their constituent elements to inform the development of such a system.

The LINZ Licence for Personal Data, for instance, provides a helpful example of how licensing systems can be constructed for, and applied to, individuals' personal data. Similarly, the UK OGL demonstrates how licences can be attached to data held and released by public authorities. Likewise, the CC licensing system demonstrates how licences can operate on a spectrum, whereby terms of varying stringency are applied to data depending on the characteristics and contextual peculiarities of those data. Aspects of all of these approaches could conceptually serve as useful templates for the development of a system governing the disclosure of anonymised data that could be built in to the FOIA's disclosure process. For example, under a licensing spectrum for anonymised data, different licensing conditions and provisions could be attached to anonymised data disclosed under the FOIA. These conditions could be commensurate to the level of risk associated with their disclosure. Under such an approach, the more "high risk" the disclosure of specific anonymised data was, the more stringent the licence attached to those data would be, in terms of restrictions imposed in relation to future re-uses and sharing. Alternatively, the more "low risk" a disclosure was, the more relaxed the relevant licensing obligations would be. Such an approach would allow for the reconciliation of the law with the practical realities of anonymisation, as it would allow public authorities to undertake disclosures of anonymised data in a way that was commensurate with the "data situation" of those data, and for any emergent risks to be suitably managed and mitigated. If, for instance, the risk of re-identification/de-anonymisation of an anonymised dataset requested under the FOIA was considered "high risk", and not suitable for disclosure were it to be released on a "release and forget" basis. The disclosure of the same data could, however, plausibly be considered "low risk", ergo suitable for disclosure, if they were released with certain contextual controls and re-use restrictions attached. As highlighted elsewhere, a risk-based approach to the concepts of personal and anonymised data in this vein would also seemingly be compliant with the substantive terms of the GDPR and the jurisprudence of the CJEU.[65] Not only this, but it would also be broadly consistent with the emerging consensus that policy choices

[65] R Hu et al. Bridging Policy, Regulation, and Practice? A Techno-Legal Analysis of Three Types of Data in the GDPR. in R Leenes et al. (eds), Data Protection and Privacy: The Age of Intelligent Machines (Hart 2017).

regarding how best to regulate issues relating to the disclosure of anonymised data should be built around the notion of risk management.

For the reasons set out above, it is clear that development of a new model of regulation, whereby licences with post-disclosure obligations are assigned to anonymised data based on perceived risks associated with their disclosure under the FOIA, is an idea worthy of investigation. The development and application of an approach in this vein, however, would necessarily require public authorities to possess an understanding of the notion of risk, and how the level of risk associated with an act of disclosure under the FOIA could be determined. Without this, it would be impossible for such a model to operate effectively or consistently. The important next step, therefore, is to consider these issues. To this end, the subsequent section examines the nature of risk and risk-based regulation, how risks relating to the disclosure of data can be assessed and, pursuant of this, how a risk-based approach to disclosure could be built into a licensing scheme for anonymised data released under the FOIA.

5. Risk-based regulation and its applicability to matters of data protection

Risk-based regulation involves the targeting of enforcement and resources on the assessment of the risks that a particular activity poses to a regulator's aims.[66] The key components of these assessments will be the evaluation of the risks of noncompliance and the impact that said noncompliance will have on the ability of regulatory bodies to achieve their objectives. In its idealised form, therefore, risk-based regulation offers an evidence-based means of targeting the use of resources, and of prioritising attention to the highest risks, in accordance with a transparent, systematic and defensible framework.[67] Over the last decade or so there has been an observable rise in regulators adopting risk-based frameworks of supervision worldwide.[68] Due to recent technological developments, including the exposure of the limits of anonymisation techniques, we are now seeing the emergence of proposals for the adoption of risk-based regulatory strategies as a means of repairing aspects of data protection law that have come under strain. In particular, calls for the adoption of risk-based approaches to the concept of personal data have become prominent.[69] Whilst some of their

[66] J Black, Risk-based regulation: choices, practices and lessons learnt" in Risk and Regulatory Policy: Improving the Governance of Risk (OECD 2010) 185-224.
[67] J Black and R Baldwin 'Really Responsive Risk-Based Regulation' [2010] 32(2) Law & Policy.
[68] This is particularly true in the context of regulatory issues pertaining environmental protection, food safety, occupational health, and financial services. See: J Black (n 66).
[69] On this issue, see: The Role of Risk Management in Data Protection [2014] Centre for Information Policy Leadership. Available at: https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/white_paper_2-the_role_of_risk_management_in_data_protection-c.pdf; Risk, High Risk, Risk Assessments and Data Protection Impact Assessments under the GDPR: CIPL GDPR Interpretation and Implementation Project [2016] Centre for Information Policy Leadership. Available at: https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_gdpr_project_risk_white_paper_

constituent elements may differ, these proposals have one key underlying premise: the idea that data should only be considered "personal" if there is a significant risk of those data being used to identify an individual and/or those data being used in a way that causes harm to the individual to whom they relate.[70] As also noted previously, this is an approach to issues regarding personal data and the anonymisation thereof, that appears to be compliant with the substantive terms of the General Data Protection Regulation and Data Protection Act 2018.[71] The possibility of incorporating a risk-based approach to data protection as part of data protection by design strategies has also been discussed elsewhere.[72]

However, there are several challenges inherent in the adoption of risk-based approaches to data protection that are often overlooked.[73] In short, though proposals for the adoption of risk-based approaches to aspects of data protection law are often prima facie appealing, a detailed explanation relating to how such strategies could be designed, enforced and applied is often lacking.[74] Accordingly, were a risk-based licensing framework for anonymised data to be built into the FOIA's disclosure process, these factors relating to design, enforcement, and application would all need to be considered. To this end, the remainder of this section is dedicated to addressing the following questions:

1) How can public authorities assess the level of risk associated with the disclosure of anonymised data?

2) What issues and factors would they need to take into account when undertaking such assessments?

What follows below is a survey and summary of some already-existing prominent methodologies for assessing risk in the context of data processing activities. This has two purposes: to illustrate how

21_december_2016.pdf; Article 29 Data Protection Working Party [2014] Statement on the role of a risk-based approach in data protection legal frameworks, 14/EN WP218.

[70] Ibid.

[71] See: R Hu et al. (n 65).

[72] See, for example: S Spiekermann and M Oetzel 'Privacy-by-Design Through Systematic Privacy Impact Assessment – a Design Science Approach' [2012] ECIS – Conference Proceedings, 2012.

[73] H Pearce, 'Big data and the reform of the European data protection framework: an overview of potential concerns associated with proposals for risk management-based approaches to the concept of personal data' [2017] 26(3) Information & Communications Technology Law 312-335.

[74] Ibid. On the difficulties relating to designing, enforcing and applying risk-based approaches to regulation generally, see: O Renn, 'Three Decades of Risk Research: Accomplishments and New Challenges' [1998] 1(1) Journal of Risk Research 49-71; R Brownsword, 'Nanoethics: Old Wine, New Bottles?' [2009] 32(4) Journal of Consumer Policy 355-379; M Siegrist et al. 'Perception of Risk: The Influence of General Trust, and General Confidence' [2005] 8(2) Journal of Risk Research 145-156; N Nicholson et al. 'Personality and Domain-Specific Risk Taking' [2005] 8(2) Journal of Risk Research 157-176; R Baldwin and M Cave, Understanding Regulation: Theory, Strategy and Practice (Oxford University Press 1999).

other already-existing risk-based approaches to data governance and disclosure operate in practice, and to highlight examples of best practice for the purposes of informing the development of a new bespoke risk-based disclosure model for anonymised data released under the FOIA. The methodologies examined represent prominent examples of approaches for assessing the risks associated with data processing activities, published by notable regulatory and non-regulatory bodies from a range of different jurisdictions.

5.1. CNIL Methodology for privacy risk management

The *Methodology for privacy risk management* is the English translation of guidance published by the French Data Protection Authority, the *Commission nationale de l'informatique* et des *libertés (*CNIL*)*.[75] It describes a method for managing the risks to individuals to which the processing of personal data can give rise. To date, it is one of the most comprehensive and detailed methodologies for approaching privacy and data protection risks associated with the processing of personal data published worldwide.

The methodology comprises of two substantive chapters. The first explains the theoretical and conceptual aspects of risk management, whereas the second has a more practical focus and explains a method for managing data protection and privacy risks. The approach outlined in the second chapter is based on the French information security risk management methodology EBIOS.[76] Unlike EBIOS, however, the CNIL methodology is not concerned with information security in a general sense, and instead focuses only on issues relating to privacy and data protection. Because of this, the CNIL methodology is sometimes described as a customised version of EBIOS.[77]

The methodology uses an analytical approach to identify and treat "privacy risks". These are broadly defined as any events or negative impacts that may be experienced by individuals because of their personal data being processed. This approach consists of five individual steps that should be applied to all instances of personal data processing suspected of being potentially harmful:

1) **A background study**: In this stage the data controller will be expected to consider what data are being processed, the purposes of the processing, the identity of the data processor, and the conditions under which the processing is to take place/how the data will be stored.

[75] 'Methodology for privacy risk management' [2012] CNIL.
[76] For an overview of EBIOS, see: W Abbass, A Baina and M Bellafkih, 'Using EBIOS for risk management in critical information infrastructure' [2015] 5th World Congress on Information and Communication Technologies (WICT).
[77] 'Privacy impact assessment and risk management: Report for the Information Commissioner's Office' [2015] Trilateral Research & Consulting 141. Available at: https://ico.org.uk/media/1042196/trilateral-full-report.pdf

These questions are not limited to those pertaining specifically to the data in question, but also the circumstances surrounding them. Notably, the methodology suggests data controllers examine what kind of hardware and software will be used for the processing of the data, and what (if any) kinds of computer communication networks the personal data will traverse.[78]

2) **A "feared events" study**: Here, the data controller is required to list of all the feared possible events that may stem from the processing of the data, and their severity. The methodology specifically lists the following as possible feared events:
   - Legal processes becoming unavailable to the individual(s) to whom the data relates;
   - Personal data being processed for purposes other than for which they were initially collected;
   - Personal data being illegitimately accessed by other parties;
   - Personal data being illegitimately modified or used for illegitimate purposes; and
   - Personal data being deleted or lost.[79]

   The methodology suggests that events such as these may lead to undesirable consequences, such as fraud, identity theft, unfair treatment, and privacy violations.[80]

   Having devised a list of all foreseeable feared events, the data controller is then required to estimate the prejudicial effect each event would have on affected individuals were it to occur. According to the methodology, there are four possible levels of severity:

   > **Negligible**: data subjects will either not be affected, or may encounter minor inconveniences that can be overcome without difficulty (e.g. time spent re-entering or correcting information, annoyances, irritations etc.)
   >
   > **Limited**: data subjects may encounter significant inconveniences that they will be able to overcome, albeit with some possible difficulties (e.g. denial of services, financial costs, stress etc.)
   >
   > **Significant**: data subjects may encounter significant consequences that they should be able to overcome, albeit with major difficulties (e.g. misappropriation of funds, property damage, loss of employment, non-permanent/life-changing health issues etc.)

[78] CNIL (n 75) 10.
[79] Ibid. 6.
[80] Ibid.

Maximum: data subjects may encounter significant or irreversible consequences that potentially may not be overcome (e.g. financial distress, inability to work, long-term physical problems, permanent injury, death etc.)[81]

3) **A threats study:** In this stage the data controller is required to develop a list of possible threats that may cause the abovementioned "feared events" to occur, and their likelihood. In other words, the data controller must ask "What factors may cause harmful consequences to arise, and what is the *probability* of their occurrence?" To determine probability the data controller must consider the possible risk sources. These include, people (including motivated intruders and employees of the data controller) and their motivations, computer viruses and malicious software that may compromise the data or the processing thereof, faulty hardware/software, acts of god, and even rodents. Once done, the probability of harm can be denoted as either negligible, limited, significant, or maximum.[82]

4) **A risk study:** The fourth stage in the process requires the data controller to map the results of the feared events and threats studies to determine the overall level of risk associated with the data processing operation. Specifically, the methodology advises that by taking the scores assigned to the data processing operation during stages 2 and 3 (i.e. negligible, limited, significant, maximum), and cross-referencing the two, it will be possible to plot the overall level of risk associated with the processing operation on a risk matrix.[83]

5) **A measures study:** Having determined the level of risk associated with the data processing activity, the final stage in the process requires the data controller to decide how best to mitigate any risks identified.

In summation, according to the methodology, the best way to assess the level of risk associated with an act of data processing is to start by considering the contextual peculiarities of that processing, and analysing the possible *severity* or *impact* of harms that may stem from it. An analysis of the *likelihood* of such harms occurring, regardless of their severity or impact should then be performed, and the scores derived from the both analyses used to determine an overall level of risk. This score should then be used to determine how best to proceed with the treatment of that risk. Of note is the way in which the methodology emphasises that its five steps should be considered cyclical and recurring, rather than a one-off exercise. The methodology invites data controllers to evaluate the risks

[81] Ibid.
[82] Ibid.
[83] Ibid. 18.

associated with their data processing operations continually. The methodology, therefore, bears some similarities to anonymisation-specific guidance relating to assessing de-anonymisation risks espoused by the United Kingdom Anonymisation Network's (UKAN) Anonymisation Decision-making Framework.[84]

5.2. NIST SP 800-122, Guide to Protecting the Confidentiality of Personally Identifiable Information

The National Institute of Standards and Technology (NIST) is a non-regulatory agency of the United States Department of Commerce responsible for developing standards and guidelines relating to information security. Its *Guide to Protecting the Confidentiality of Personally Identifiable Information*, published in 2010,[85] is designed to assist Federal agencies in protecting the confidentiality of personally identifiable information (PII) and, in so doing, protect the individuals to whom such information relates from harms that might stem from its processing.[86] Whilst the Guide focuses primarily on providing guidance to the employees of Federal agencies, its text implies its guidance will also be relevant to other areas of application where the use of information may give rise to privacy and data protection risks.[87]

The Guide is comprised of an executive summary, an introduction to PII, and chapters on PII confidentiality impact levels, PII confidentiality safeguards, and incident responses to breaches involving PII, as well as various appendices. Holistically, it provides a methodology for assessing whether, and to what extent, PII is at stake as part of a data processing activity (i.e. the level of risk associated with said data processing activity) and, with examples, indicates how such incidents should be handled, and how emergent risks should be mitigated. Specifically, the Guide suggests that NIST's Risk Management Framework,[88] another NIST publication focusing on risk management more generally, should be used to assess the "impact levels" of data processing activities. The term "impact levels" in this context denotes possible harms:

> "*…any adverse effects that would be experienced by an individual whose PII was the subject of a loss of confidentiality, as well as any adverse effects experienced…"*[89]

---

[84] The UKAN Anonymisation Decision-making Framework (ADF) represents the culmination of a three-year cross-sector collaboration process between multiple disciplines. Though its 156 pages are not legally binding, its guidance is widely considered to be authoritative. See: M Elliot et al (n 63).
[85] 'Guide to Protecting the Confidentiality of Personally Identifiable Information' [2015] NIST
[86] It should be noted, however, that PII, as generally defined, is a narrower concept than the concept of personal data, as defined in the GDPR.
[87] NIST (n 85) 11.
[88] Ibid.
[89] Ibid.

The possible levels of impact, according to the Guide are low, moderate or high. The impact of any data processing will be "low" if it has the potential to have a "limited adverse effect", such as causing individuals to experience a minor financial loss, or other minor inconvenience. Conversely, data processing will have a "moderate" impact if it may lead to a "serious adverse effect", such as individuals experiencing significant financial losses, or other significant, but non-fatal, harm. "High" impact data processing activities are those that may have a "severe or catastrophic adverse effect". The Guide lists major financial losses and other ruinous events, such as the loss of life or incurrence of serious life-threatening injuries, as examples of such events.[90] According to the Guide, the greater the level of potential impact, the more extensive efforts should be to apply safeguards and ensure the security and integrity of the processing of those data, or to even cease the processing activity entirely if need be. In this regard, the Guide suggests that operational safeguards (e.g. the creation of data governance policies and procedures, raising internal awareness regarding the importance of data security, personnel training, and education), privacy-specific safeguards (e.g. minimising the use, collection and retention of PII, conducting privacy impact assessments, anonymising data) and security controls (e.g. access restrictions) will all help mitigate the level of risk associated with data processing.

The Guide advises that determinations regarding which impact/risk category a data processing activity falls into (i.e. low, moderate, or high) should be calculated by way of an examination of the following factors:

- The level of identifiability possible from the data;
- The quantity and scope of the data;
- The sensitivity of the data;
- The contextual details regarding how the data will be used;
- Whether there is any obligation to protect confidentiality; and
- Which parties have access to the data, and where/the conditions under which the data are stored.

As has been remarked elsewhere, however, is not initially clear whether the factors mentioned above represent *impact* factors or *likelihood* factors, or a combination of the two.[91] In other words, it is not clear from the Guide's wording whether the above factors should be accounted for when calculating the *level* of harm likely to arise from a data processing activity, the *probability* of harm, or both. Clarification, however, is derivable from other more recent NIST publications. For instance, a draft guide for Privacy Risk Management for Federal Information Systems, released in 2015, suggested that

[90] Ibid. 18.
[91] Trilateral Research & Consulting (n 77).

risks associated with data processing activities should be expressed as a function of the likelihood of an adverse outcome occurring *because* of a data processing activity, multiplied by the magnitude of the adverse outcome *should* it occur.[92] This ethos was retained and incorporated into the final version of the guidance, published in January 2017.[93] This message is also stressed in two other significant NIST publications, the NIST Risk Management Framework for Information Systems and Organisations,[94] and the NIST Guide for Conducting Risk Assessments.[95] When read in conjunction, it is clear that, whilst the NIST Guide for Protecting the Confidentiality of PII does not explicitly say the factors mentioned above should be viewed through a combination of both impact and probability lenses, this is something that is strongly implied other NIST publications. Of interest, however, is the way in which the 2015 draft guide, despite stressing the importance of data protection risk assessments encompassing both a consideration of the probability and likely impact of harms, acknowledges and advises that quantifications regarding the impact of possible harms may be difficult to make with any consistency. This is mainly due to the divergence in the way individuals experience problems, particularly those relating to embarrassment and other psychological issues.[96]

5.3. ICO PIA Code of Conduct

In 2013, the UK Information Commissioners' Office (ICO) published an extensive report on privacy impact assessments and risk management, prepared by Trilateral Research & Consulting.[97] The report aimed to bridge the gap between privacy impact assessments (PIAs) and risk management standards and methodologies. In so doing, it aimed to stimulate the development of best practice regarding assessing privacy and data protection risks inherent in data processing activities, and the mitigation of such risks. Following the report's publication, the ICO subsequently published a PIA Code of Conduct in February 2014.[98]

The Code provides systematic guidance on how to conduct PIAs, with particular emphasis the consideration of privacy and data protection risks. It consists of nine chapters, the majority of which

[92] NIST, 'Privacy Risk Management for Federal Information Systems, NISTIR 8062' (Draft) [2015] 22. Available at: https://csrc.nist.gov/csrc/media/publications/nistir/8062/draft/documents/nistir_8062_draft.pdf
[93] NIST, 'An Introduction to Privacy Engineering and Risk Management in Federal Systems, NISTIR 8062' [2017] Available at: https://nvlpubs.nist.gov/nistpubs/ir/2017/NIST.IR.8062.pdf
[94] NIST, 'Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy, NIST Special Publication 800-37' (R2) [2018]. Available at: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-37r2.pdf
[95] NIST, 'Guide for Conducting Risk Assessments, Special Publication (NIST SP) – 800-30 Rev 1' [2012] Available at: https://www.nist.gov/publications/guide-conducting-risk-assessments
[96] NIST (n 85).
[97] Trilateral Research & Consulting (n 77).
[98] 'Conducting privacy impact assessments code of practice' [2014] ICO. Available at: https://www.pdpjournals.com/docs/88317.pdf

explain the nature of PIAs and when they are necessary. The first chapter explains PIAs as processes that assist organisations in identifying and minimising the privacy and data protection risks of new projects and policies. The following chapters then explain the PIA process, how data controllers should start by considering which of their data processing operations will require a PIA, and how consultation with relevant stakeholders (i.e. those potentially affected) will be important for any PIA. Of particular interest are chapters 6 and 7, which deal with identifying and managing privacy and data protection related risks. Chapter 6 of the Code specifies that possible risks relating to privacy and data protection include:

- Personal data being shared inappropriately;
- The context in which personal data are processed changing over time, leading to data being used for purposes other than for which they were originally collected;
- Intrusive measures being taken against individuals as a result of the collection of their personal data;
- The sharing or merging of datasets resulting in the collection of a much wider set of information than individuals might expect;
- The collection of direct and indirect identifiers having the effect of preventing individuals from using a service anonymously;
- Vulnerable persons being identified via the disclosure of information;
- Anonymised data becoming de-anonymised and individuals being re-identified from those data;
- Keeping personal data for longer than necessary.[99]

Having outlined these possible risks, the Code advises that the most thorough way to assess the overall level of risk associated with any data processing activity is to consider both the probability and impact of such risks.[100] Chapter 7 provides a number of "privacy solutions" that can be used as a means of eliminating or reducing any identified risks, and can be applied in a way that is commensurate to the levels of those risks. Here, the Code suggests all data processing activities involving personal data will be capable of having some impact on individuals to whom those data relate. Accordingly, there will always be some risk when processing personal data. The important thing, the Code advises, is to ensure that any such impacts are proportionate to the objective of the processing, and that the processing has a basis in law.[101] Where the level of risk associated with a data processing operation is high, the Code implies that it should not ordinarily go ahead, or that steps should be taken to reduce

[99] Ibid. 22-23.
[100] Ibid. 24.
[101] Ibid. 26.

the level of risk to a more tolerable level. To this end, the Code advises that the following measures will likely be particularly useful in this context:

- Implementing appropriate technological security measures;
- Anonymising data where possible;
- Establishing procedures for data subject access requests; and
- Selecting data processors which will provide high levels of data security, and ensuring data sharing agreements are in place to protect the integrity of any processed information.[102]

Once the above risk analyses have been undertaken, a decision can then be made regarding how to proceed. This will normally result in a decision that the risks associated with the data processing activity have been reduced to a satisfactory level, and proceeding, or deciding the risk level remains too high, and that the processing should not go ahead.[103] Subsequent, more recent, guidance published by the ICO on the subject of DPIAs broadly endorses the approach suggested by the Code. ICO guidance on accountability and governance, and GDPR compliance, both published in 2018, for instance, suggest that to determine the level of risk associated with any data processing activity data controllers should consider both the probability of a risk materialising, and the likely impact its materialisation would have on any affected individuals.[104]

5.4. Discussion

The methodologies reviewed above vary in terms of structure and substance. However, they all share a key similarity. Either explicitly or implicitly, they all endorse the idea that calculating the level of risk inherent in any personal data processing activity will require an analysis of both the *probability* of an individual experiencing harm because of their personal data being processed and, were that harm to manifest, the *severity* of the harm that would be experienced. This is consistent with the global risk management standard ISO 31000, maintained by the International Organization for Standardization, which defines "level of risk" as the "magnitude of a risk or combination of risks" and "their likelihood".[105] In addition to the methodologies considered above, other notable organisations, such as the Canadian Institute for Health Information, and Ireland's Health Information Quality Authority,

[102] Ibid.
[103] Ibid. Chapters 8 and 9.
[104] See: 'Accountability and governance: Data Protection Impact Assessments (DPIAS)' [2018] ICO. Available at: https://ico.org.uk/media/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/data-protection-impact-assessments-dpias-1-0.pdf; 'Guide to the: General Data Protection Regulation (GDPR)' [2018] ICO. Available at: https://ico.org.uk/media/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr-1-0.pdf
[105] 'ISO 31000:2009 Risk management – Principles and guidelines' [2009] International Organization for Standardization.

have also published guidance and methodologies that state that the determination of risk associated with any data processing activity can only be calculated by considering both the likelihood and impact of possible harms.[106] Various other methodologies proposed in conjunction with managing risks associated with the processing of personal data have also endorsed this approach.[107]All the methodologies and approaches surveyed above also, either explicitly or implicitly, acknowledge the benefit of applying contextual controls (e.g. information security practices, data access restrictions etc.) in a way that is commensurate to the level of risk in order to mitigate said risks. Based on these findings, we can clearly derive a number of examples of good practice that should be incorporated into the risk-assessment stage of any risk-based anonymised data licensed disclosure scheme. Specifically, any risk-assessment exercise undertaken in relation to the disclosure of anonymised data should consider both:

- The *probability* of the data being de-anonymised/re-identified (and how this probability to be mitigated by way of contextual controls); and

- The level of *impact* that de-anonymisation/re-identification would likely have on any affected individuals

Essentially, therefore, in order for public authorities to calculate the risks associated with the disclosure of anonymised data, and assign a licence to those data based on the level of said risk, they would be required to consider the following questions:

1) What is the probability of the anonymised data being de-anonymised and used to identify individuals contained in the data (and how could that probability be reduced/mitigated)?

2) In the event de-anonymisation occurs, what level of harm would likely be experienced by any identified individuals?

---

[106] See: 'Privacy and Security Risk Management Framework' [2015] Canadian Institute for Health Information. Available at: https://www.cihi.ca/en/psrm_framework_enweb.pdf; 'Privacy Impact Assessment toolkit for health and social care' [2017] Health Information and Quality Authority. Available at: https://www.hiqa.ie/sites/default/files/2017-10/Privacy-Impact-Assessment-toolkit-A5.pdf

[107] See, for instance: S Joyee De and D Le *Métayer,* 'PRIAM: A Privacy Risk Analysis Methodology' [2016] Available at: https://hal.inria.fr/hal-01302541/document*;* J Friginal, J Gulochet and M Killijian (2014) 'AMORES L1.2 – a Privacy Risk Assessment Methodology for Location-Based Systems' [2014] Rapport LAAS 16048, LAAS-CNRS. Available at: https://hal.archives-ouvertes.fr/hal-01282191/document*);* K Greenaway, S Zabolotniuk and A Levin,'Privacy as a risk management challenge for corporate practice' [2012] Ted Rogers School of Management, Ryerson University, Privacy and Cyber Crime Institute, Available at: https://www.ryerson.ca/content/dam/tedrogersschool/privacy/privacy_as_a_risk_management_challenge.pdf; 'A Risk-based Approach to Privacy: Improving Effectiveness in Practice' [2014] Centre for Information Policy Leadership. Available at: https://www.huntonak.com/files/upload/Post-Paris_Risk_Paper_June_2014.pdf

5.4.1. Assessing the probability of de-anonymisation/re-identification

To determine the probability of an individual being re-identified from anonymised data (i.e. anonymised data being de-anonymised) it will be important for public authorities to identify and consider the different types of threat capable of undermining the techniques or approaches originally used to anonymise the data in question. As noted by the Article 29 Working Party, the main types of de-anonymisation threats can be delineated into three general categories: "singling out","linkability" (with linkability itself being split into three sub-categories of local, domain, and global linkability), and "inferences".[108]

"Singling out" refers to the possibility of isolating either some, or possibly all, records within a dataset which can be used to identify an individual. "Linkability" refers to the possibility of an individual being identified through the linking of two records concerning the same data subject or a group of data subjects. "Local linkability" refers to the possibility of linking records from within the same dataset. "Domain linkability" refers to the possibility of linking records regarding a data subject from two or more datasets held by the same data controller. "Global linkability" refers to the possibility of linking records regarding a data subject from two or more datasets that are not necessarily in the possession of the data controller, but in the possession of other parties. "Inference" refers to the possibility to deduce the identity, or the value of an attribute, of a data subject through analysing the values and attributes of other data items.[109]

Assessing these threats in relation to particular data will require an assessment of many contextual issues. These include the existence of any data security measures, the existence of any motivated intruders, the existence of data access and usage restrictions, the identity of data processors, the purposes for which data are to be processed, and the existence of other data.[110] Such assessments will also require an analysis of which other data may exist and be in the possession of other parties who may come into contact with the data that is to be released, and whether any additional data may become available in the future that may also affect this situation (i.e. the "data environment").[111] This is likely to be extremely challenging in some circumstances. As suggested elsewhere, for instance, requiring data controllers to consider what data may be "out there" in the future is in some ways akin

[108] See: Article 29 Data Protection Working Party (2014) Opinion 05/2014 on Anonymisation Techniques, WP 216, 8-12.
[109] Ibid.
[110] M Elliot et al (n 63).
[111] Ibid.

to asking them to predict the future and, as is generally accepted, the future is inherently unpredictable.[112] Because of this inevitability, it will be important for public authorities to obtain as much relevant contextual information as possible regarding any anonymised data they aim to disclose, so to ensure their probabilistic assessments of de-anonymisation are as informed and accurate as possible. The possibility of using additional contextual controls and procedures, as a means of acquiring such information, is an idea discussed below.

5.4.2. Assessing the possible impact of de-anonymisation/re-identification

There are multiple factors public authorities would be required to take into account when assessing the possible impact of de-anonymisation. Notably, issues such as whether de-anonymisation may lead to negative social, economic, financial, or physical impacts, and the severity of any such impacts, will be highly pertinent. However, as hinted at above, it will not always be straightforward to assess the possible impact de-anonymisation might have on affected individuals. Whilst a public authority might go to great lengths to ascertain all conceivable possible impacts de-anonymisation could have on an individual, there will always remain a possibility that it might overlook an obscure, previously unthought-of, impact. As noted by Rubinstein and Hartzog, the notion of impact can be difficult to quantify in this context. Some impacts may not be immediate and could occur years after data disclosure. Other impacts may occur quickly but never be detected.[113]

Where there are doubts regarding the impact de-anonymisation may have on affected individuals, the best option will be for public authorities to consider the nature of the data, particularly the sensitivity of those data. The more sensitive personal data are, the more likely it is that they will be of interest to nefarious actors and other third parties, and the more severe the consequences (i.e. the impact) would be for affected individuals, were such data used for improper purposes.[114] This position is implicitly endorsed by the GDPR, which specifies that the processing of "special category" personal data (i.e. data pertaining to an individual's ethnic origins, political opinions and affiliations, religious beliefs, trade union membership genetic makeup, sex life or sexual orientation, or health – known as "sensitive personal data" under the now defunct Data Protection Directive) is subject to stricter conditions than non-special category personal data.[115]

However, whilst the GDPR gives a list of data types classified as "sensitive", there are additional nuances that must also be teased out if we are to be able to fully conceptualise sensitivity in the

[112] H Pearce (n 73).
[113] E Rubinstein and W Hartzog, 'Anonymization and Risk' [2015] 703 Washington Law Review 730.
[114] M Elliot et al (n 63).
[115] See: Article 9 GDPR.

immediate context. As suggested elsewhere, for instance, the true sensitivity of data will depend on a range of different factors including the social utility of sharing those data, and the possibility of using those data to obtain or create new data.[116] In this regard, recent research can help determine the sensitivity of certain types of data, and thus assist with quantifying the levels of risk associated with their disclosure. Of particular note is one recent study from the USA, which attempted to map the degree and types of risk individuals perceive when sharing their personal data.[117] The study noted that, generally speaking, it was possible to identify six general clusters of data types: financial information (e.g. credit card information, bank account information, credit scores, handwriting samples, driving licence numbers), secure identifiers (e.g. health insurance data, passport information, medical history, DNA, finger prints), basic demographics (e.g. occupation details, gender, race, marital status, height, date of birth), contact information (e.g. IP addresses, work addresses, phone numbers), personal preferences (e.g. sexual orientation, email addresses, political affiliations) and community interactions (e.g. information pertaining to law enforcement, social networking profiles, friends' contact details).[118]Alongside this, it was discovered that the risks associated with sharing personal data can be broken down into four categories: monetary risk (i.e. the risk of potential financial loss), social risk (i.e. the risk associated with reputational loss and threats to self-esteem), physical risk (i.e. the risk of bodily harm) and psychological risk (i.e. the risk of emotional and psychiatric damage).[119] An important implication of this is that individuals' perceptions of risks relating to the disclosure of their personal data are multidimensional. Not only are there different types of data that carry with them different levels of risk, but there are at least four distinct types of risk of between which individuals differentiate.

Perhaps unsurprisingly, the overall greatest level of perceived risk across all the different risk types was associated with data falling within the "secure identifier" cluster, with the lowest level of risk being associated with data falling within the "basic demographics cluster". Data falling within the "financial information" and "community interaction" clusters were ranked similarly in terms of perceived risk, but were adjudged to be of a considerably lower risk than data within the secure identifier cluster. The fact that these two categories were ranked similarly casts doubt on the commonly held belief that an individual's financial information is of a more sensitive nature than information pertaining to their community interactions. This finding may reflect current practical realities, where individuals are increasingly concerned with the maintenance of their public image in

[116] S Joyee De and D Le Métayer, D. Privacy Risk Analysis (Morgan & Claypool 2016)
[117] G Milne et al. 'Information Sensitivity Typology: Mapping the Degree and Type of Risk Consumers Perceive in Personal Data Sharing' [2017] 51(1) The Journal of Consumer Affairs 133-161.
[118] Ibid. 144.
[119] Ibid. 139.

a world of pervasive, visually based social media.[120] This sort of information will be highly pertinent in the context of attempting to discern the possible harms that may stem from an individual being re-identified from anonymised data disclosed under the FOIA. No directly comparable study on attitudes to perceived risks associated with different data types appears to have been undertaken in the EU. This research, however, may still serve as a useful point of reference, and its results appear to correlate with other recent research undertaken in the EU that has also suggested that individuals are more concerned with risks associated with disclosing certain types of data than they are with others.[121]

6. Other noteworthy factors/issues

6.1. Using contextual controls to assist with risk assessments

The the utilisation of contextual controls will be key to managing and mitigating the level of risk associated with the disclosure of anonymised data, particularly in relation to assessing the probability of de-anonymisation. If a spectrum-based system of data licensing for anonymised data released under the FOIA were to become a reality, many contextual controls could be embedded in the terms of licences themselves. For instance, the licence terms could impose obligations on recipients of disclosed data to avoid certain uses of the data that could cause the level of risk associated with those data to rise. A licence might, for instance, prohibit the sharing of the data with specific named parties, or even with any other parties; prohibit certain data processing activities (e.g. merging of datasets); or mandate that certain data security measures were put in place. However, it is not only in the licences themselves that contextual controls could have a role to play.

---

[120] Ibid. 150.

[121] See, for example: 'Summary Report of Qualitative Research Into Public Attitudes to Personal Data and Linking Personal Data' [2013] The Wellcome Trust. Available at: https://mesh.tghn.org/site_media/media/articles/Qualitative_Research_into_Public_Attitudes_to_Personal_Data_and_Linking_Persona_FLr04DM.pdf; 'Privacy and personal data' [2014] Ipsos MORI. Available at: https://www.ipsos.com/sites/default/files/migrations/en-uk/files/Assets/Docs/Polls/jrrt-privacy-topline-nhs-2014.pdf; 'Annual Track 2014' [2014] ICO. Available at: https://ico.org.uk/media/about-the-ico/documents/1043485/annual-track-september-2014-individuals.pdf; 'Data Protection' [2015] Eurobarometer. Available at: http://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_431_en.pdf; 'Data protection rights: What the public want and what the public want from Data Protection Authorities' [2015] ICO. Available at: https://ico.org.uk/media/about-the-ico/documents/1431717/data-protection-rights-what-the-public-want-and-what-the-public-want-from-data-protection-authorities.pdf; 'Data privacy: what the consumer really thinks' [2015] DMA. Available at: https://dma.org.uk/uploads/ckeditor/Data-privacy-2015-what-consumers-really-thinks_final.pdf; 'ODI/YouGov Poll Results – Attitudes towards data sharing' [2018] Open Data Institute. Available at: https://theodi.org/blog/odi-survey-reveals-british-consumer-attitudes-to-sharing-personal-data-online. See also: J Addae et al. 'Measuring attitude towards personal data for adaptive cybersecurity' [2017] Information & Computer Security.

As explained above, the crux of the notion of data protection by design is the incorporation, or "building in" of data protection ideals to entire data processing processes and activities. In accordance with this general ethos, there is no reason to confine contextual controls to the post-disclosure stage of the FOIA disclosure process (i.e. through the application of licences). Instead, they could have a role to play much earlier, at the stage FOI requests are initially made. To this end, one type of contextual control that could be embedded in the FOIA's disclosure process is a step that required the requester to state the reasons behind their request, and the purposes for which they wished to acquire the data requested. The inclusion of such a mechanism would have at least two discernible benefits. First, if requesters of anonymised data under the FOIA were required to state their objectives and intended uses for the data (e.g. to conduct scientific research on a dataset in order to develop treatments to a particular condition, to investigate corruption, or monitor public spending), this would help public authorities to identify precisely which data were at the centre of the FOI request. Knowledge of the intended purpose would help with the identification of precisely which data would be necessary to respond to the request in full. In so doing, this would both expedite the FOIA disclosure process and make it more effective. Second, knowledge of the motivations and intended purposes behind FOI requests relating to anonymised data would greatly assist with the abovementioned risk-assessments relating to the probability of de-anonymisation. As noted above, the probability of de-anonymisation will always be highly context dependant. Ergo, awareness of the contextual peculiarities regarding the purposes of an FOI request would help with risk analyses in respect of the release of those data. For example:

> *Person A makes a FOI request in relation to an anonymised dataset held by Public Authority B. When making their request Person A specifies that they wish to use the dataset for purpose X. With this information in mind, Public Authority B undertakes a risk assessment, and determines that, if the dataset was to be accessed by Person A only, and used for purpose X only, the probability of de-anonymisation would be low.*
>
> *If, however, the dataset was to be used for purpose Y, or shared with Person C, the probability of de-anonymisation would be significantly higher. Pursuant of this, Public Authority B determines that the open and unconditional release of the dataset, or the release of the data under an unrestrictive licence, would not be appropriate. Doing so would give rise to unacceptably high levels of risk.*
>
> *Releasing the dataset to Person A under a licence which prohibited the dataset being used for anything other than purpose X (or at least prohibited the data from being processed for any purposes incompatible with purpose X) and prohibited the dataset from being shared with Person*

B (or possibly prohibited any subsequent sharing at all), on the other hand, would conceivably be appropriate. Disclosure on this basis would not give rise to the same high level of risk.

Requiring individuals making requests to access anonymised data under the FOIA to state their reasons and intended objectives, however, would not amount to a panacea to difficulties and challenges that may arise in conjunction with the assessment of risk. Even if the intended processing activity of an applicant was known at the time a FOI request was made, for instance, some types of data processing activities may be inherently insusceptible to risk analyses.[122]

6.2. The possible importance of metadata

In addition to the above discussions relating to data licensing, risk assessment, and contextual controls, this section focuses on the notion of metadata, and examines how it may also have a role to play in a new model for disclosing anonymised data under the FOIA based on the aforementioned elements.

Metadata is commonly defined as "information about information", or "data about data".[123] However, the "data about data" tagline is arguably superficial, misleading, and may be too broad to be of any meaningful practical value.[124] A more complete definition devised following the results of one 2006 study, suggests that it would be more appropriate to define metadata as:

> "*…structured, encoded data that describe characteristics of information bearing entities to aid in the identification, discovery, assessment, and management of the described entities*"[125]

Regardless of its definition, however, there is no doubt as to the key constituent aspects of the concept. It is widely accepted, for instance, the crux of metadata is that they provide descriptive information about the producer, content, quality, condition, and other characteristics of data.[126] These descriptions endeavour to provide a robust list of common terminology and definitions for data elements. Typically, these elements provide information about one of the following four topics:

---

[122] On this issue, see: J Black and R Baldwin (n 67); T Aven and O Renn, 'On risk defined as an event where the outcome is uncertain' [2009] 12(1) Journal of Risk Research 1-11.

[123] See, for instance: K Jeffery, 'Metadata: an overview and some issues' [1998] ECRIM News. Available at: https://www.ercim.eu/publication/ws-proceedings/11th-EDRG/jefferey.pdf; O Lassila, 'Web metadata: a matter of semantics' [1998] 2(4) IEEE Internet Computing 30-37; B Newell, 'The Massive Metadata Machine: Liberty, Power and Secret Mass Surveillance in the U.S. and Europe' [2014] I/S: A Journal of Law and Policy for the Information Society 481-522; K Huner et al. 'Collaborative management of business metadata' [2011] International Journal of Information Management.

[124] P Nadkarni and M Prakash, Metadata-driven Software Systems in Biomedicine (Springer 2011) 1-16.

[125] J Ma, 'Managing metadata for digital projects' [2006] 30 Library Collections, Acquisitions and Technical Services 3-17.

[126] S Guptill, Metadata and data catalogues. in P Longley et al (eds), Geographical Information Systems: Management Issues and Applications (Wiley 1999) 678.

Availability: data needed to determine the sets of data that exist in relation to a specific subject or object.

Fitness for use: data needed to determine whether a dataset meets a specific need.

Access: data needed to acquire an identified dataset.

Transfer: data needed to process and use a dataset.[127]

Metadata, thus, allow individuals or organisations to ascertain the meaning of data within a dataset allows them to search for relevant data within a dataset, to determine whether the data contained within the dataset is relevant or useful to them, and possibly retrieve copies of the data contained within the dataset. Metadata can, therefore, reveal which datasets will be of interest to a particular individual or organisation and, once, identified, provide the information required to access or retrieve the dataset.[128]

Metadata are pervasive throughout all forms of electronic communications and interactions. To use an illustrative example, when person A sends an email to person B, the content of that email can be considered data, as can the subject of the email and any of that email's attachments. The email addresses of person A and person B, their IP addresses, and the data and time that the email was sent, will all be examples of the email's metadata.[129] Metadata, however, are not just associated with electronic communications, they also serve to document various properties of other events, behaviours, documents or processes. A notable example of this are vehicle licence plate recognition systems, which create metadata about the location of vehicles at specific points in time. Similarly, digital cameras will often create metadata about the location in which the camera was used, as well as data in respect of its aperture, focal length and shutter settings. Word processing programmes typically save metadata regarding the documents they produce, such as the name of the author, the time and date on which the document was typed, the dates of any subsequent changes to the document, and the total number of words and pages in the document.[130]

In summation, the key factor which distinguishes metadata from raw data is the fact that raw data are generally considered elements of information that model or represent real-world phenomena,

---

[127] Ibid.

[128] R Ianella and A Waugh, 'Metadata: enabling the Internet' [1997] DSTC Pty Ltd. Available at: http://archive.ifla.org/documents/libraries/cataloging/metadata/ianr1.pdf

[129] A Gray, 'Cloud" Atlas – a Map to Amending Metadata Privacy Law in the Modern Era' [2016] 52(2) Gonzaga Law Review.

[130] B Newell (n 123).

whereas metadata is a term used to refer to information regarding the storage, cataloguing, sharing, accessing, and altering of those data.

6.2.1. The value of metadata and metadata catalogues

Data have become increasingly valuable and ubiquitous in commerce and industry worldwide. Data, however, regardless of their quantity and quality will only be of use if they can be identified and located. For instance, data cannot be used by anybody for any purpose whatsoever if their existence is unknown or they cannot be located or accessed.[131] Identifying and locating data, however, is not always straightforward, and is becoming increasingly challenging in the emerging big data environment. Studies have suggested, for example, that as little as twenty percent of the data held by many organisations can be located easily, and that on average up to thirty percent of a member of that organisations' working day will be wasted attempting to locate the data that they need.[132] It is against this background that metadata themselves have become extremely valuable.

As noted above, one key purpose of metadata is that they can help provide structured and searchable information that allow individuals and organisations to locate and identify datasets that may be of use to them. The analysis of metadata itself, however, may also be very valuable and allow for the drawing of novel conclusions. It is for this reason that suggestions are increasingly being made that metadata are, in many contexts as, if not more, useful and valuable than the raw data they can help identify.[133] In conjunction with this, services known as metadata catalogues have also become increasingly relevant.

Metadata catalogues are mechanisms that allow for the storing and accessing of descriptive metadata. They allow their user to search for data based on desired attributes and parameters. Records within these catalogues will contain the metadata for a dataset. The dataset to which the metadata refers is

---

[131] This was a point acknowledged and discussed in detail during a discussion session at the Open Data Camp conference held in Belfast in October 2017. See: 'Catalogues and metadata' (Open Data Camp) <http://odcamp.org.uk/catalogues-and-meta-data/> accessed June 2020.

[132] 'Various Survey Statistics: Workers Spend Too Much Time Searching for Information' (Cottrill Research) <https://www.cottrillresearch.com/various-survey-statistics-workers-spend-too-much-time-searching-for-information/> accessed June 2020.

[133] See, for instance: J Pomerantz, Metadata (MIT Press 2015); M Chessell, 'The case for open metadata' [2016] IBM Big Data and Analytics Hub; M Lytras and M Sicilia, (2007) 'Where is the value in metadata?' [2007] 2(4) International Journal of Metadata, Semantics and Ontologies; W Wescott II, (2007) 'The Increasing Importance of Metadata in Electronic Discovery' [2007] 14(3) Richmond Journal of Law & Technology; A Zuiderwijk et al. 'The potential of metadata for linked open data and its value for users and publishers' [2012] 4(2) JeDEM 222-244; E Mitchell, 'Assessing the Value of Metadata in Information Services' {2013] 30(2) Technical Services Quarterly; R Parekh et al. 'The importance of metadata to assess information content in digital reconstructions of neuronal morphology' [2015] 360(1) Cell and Tissue Research 121-127; A Santana et al. 'The importance of metrological metadata in the environmental monitoring' [2016] 733(1) Journal of Physics: Conference Series; Hare, J. 'What is metadata and why is it as important as the data itself?' [2016] OpenDataSoft.

not part of the catalogue. By allowing users to search their records, catalogue services support the ability of individuals to publish and search for data, services, and other tangential information.[134] In other words, metadata that are searchable through catalogues represent resource characteristics that can be queried and presented for evaluation, and further processing, by both humans and software.[135] The purpose of metadata catalogues, therefore, is to support the discovery, invocation and retrieval of other information resources.

As a case in point, the US-based computer software and machine-learning firm Tamr suggests that capturing and harnessing metadata in a robust, easily accessible catalogue can potentially create dramatic new opportunities for those involved in the fields of data science and analysis.[136] Using metadata catalogues will, for instance, allow individuals and organisations to more quickly and confidently gather data that are necessary for their intended analyses.[137] Others have gone further, and suggested that not only are data catalogues becoming increasingly important, but that they will soon become a necessary prerequisite if maximum value is ever to be derived from the data held by public sector organisations worldwide.[138] In correlation with this general ethos, an increasing number of government bodies across the globe are making data in their possession available online in the form of data catalogues. One notable example of this is data.gov, a website dedicated to the open data initiative of the US government, which contains details of a list of catalogues regarding datasets, held at both regional and local levels.[139]

6.2.2. Metadata and its possible contribution to a new FOIA disclosure framework

The utilisation of the metadata of public sector datasets can contribute to the development of a new framework for disclosing anonymised data under the FOIA, as alluded to above, in numerous ways. Mainly, the utilisation of public authority metadata would help to streamline and make the process for making FOI requests more efficient. As explained above, datasets including raw personal data will ordinarily be exempt from disclosure under FOIA. Releasing the metadata linked to raw personal data,

---

[134] N Santos and B Koblitz, 'Metadata services on the Grid' [2006] 559(1) Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 53-56; N Santos and B Koblitz, 'Security in distributed metadata catalogues' [2008] 20(17) Concurrency and Computation: Practice and Experience 1995-2007.
[134] S Guptill (n 126).
[135] M Dekkers and M Craglia, 'Temporal Metadata for Discovery' [2008] JRC Scientific and Technical Reports.
[136] 'Harnessing Meta Data in a Robust Catalog Opens Dramatic Opportunities for Organisations' (Tamr) <https://www.tamr.com/four-steps-for-managing-your-metadata/> accessed June 2020.
[137] Ibid.
[138] D Woods, 'Why You Can't Be Data-driven Without a Data Catalog' (Forbes, 2015) <https://www.forbes.com/sites/danwoods/2015/09/25/why-you-cant-be-data-driven-without-a-data-catalog/#55da2cec4510> accessed June 2020.
[139] 'The home of the U.S. Government's open data' (Data.gov) <https://www.data.gov/> accessed June 2020)

however, assuming the metadata themselves were not "personal", would not fall within the scope of this exemption.[140] The release of metadata relating to a public authority's datasets would allow interested parties to search for, enquire about, and ask questions in relation to, specific items of data contained within such datasets. Once the seeking party had identified a dataset, or an aspect thereof, held by a public authority that was of interest to them for a specified purpose, a request for access could then be made in relation to intended purpose (e.g. medical research). This would have two major benefits.

First, the ability to query the metadata would allow for individuals considering making an FOI request to more easily discern whether a public authority was in possession of the information they sought. This would lead to a reduction in time and resources required for public authorities to respond to FOI requests. With a better idea of the exact information at a public authority's disposal, for instance, an individual's FOI request would likely be more specific. Consequently, the time needed to process the request, identify the relevant information, and determine whether it was appropriate to disclose the information, would be reduced. This could also reduce the number of vexatious or speculative FOI requests received by public authorities, as individuals would have a greater idea of precisely which data public authorities held, and, this greater specificity of requests would, as outlined above, make it easier for public authorities to undertake risk assessments in relation to the disclosure of anonymised data.

Second, given the high value of metadata itself, the release of the metadata of public sector datasets will be beneficial in itself. As noted in the computer science literature, for instance, metadata analysis has the potential to transform the way in which we fight diseases and gain insight into human

[140] As has been noted elsewhere, however, metadata that can be used to identify an individual, or individuals, will still constitute personal data for the purposes of EU data protection law. See: S Gnesi, et al. My Data, Your Data, Our Data: Managing Privacy Preferences in Multiple Subjects Personal Data in B Prencel and D Ikonomu (eds), Lecture Notes in Computer Science (Springer 2014) 154-171; C Millard and W Kuan Hon 'Defining 'Personal Data' in e-Social Science' [2012] Information Communication and Society. This view is supported by the jurisprudence of the CJEU and has also been endorsed by the data protection authorities of EU Member States as well as the Article 29 Working Party. See: Case C-101/01 *Bodil Lindqvist*, EU:C:2003:596, Case C-70/10 *Scarlet v SABAM*, EU:C:2011:711, Case C-342/12 *Worten*, EU:C:2013:355, Case C-201/14 *Bara*, EU:C:2015:368 and Case C-582/14 *Patrick Breyer v Bundesrepublik Deutschland*, EU:C:2016:779; 'Opinion 4/2007 on the Concept of Personal Data, WP136' [2007] Article 29 Data Protection Working Party; 'Determining what is personal data' [2012] ICO. Available at: https://ico.org.uk/media/for-organisations/documents/1554/determining-what-is-personal-data.pdf. See also: 'What is personal data? – A quick reference guide' [2012] ICO. Available at: https://ico.org.uk/media/for-organisations/documents/1549/determining_what_is_personal_data_quick_reference_guide.pdf; 'Determining what information is 'data' for the purposes of the DPA' [2012] ICO. Available at: https://ico.org.uk/media/for-organisations/documents/1609/what_is_data_for_the_purposes_of_the_dpa.pdf

behaviours.[141] Another notable example of metadata's latent value is the BioSample Database, a database at the European Bioinformatics Institute, containing approximately two million records regarding the biological records of more than eighteen thousand species, utilised for the purposes of advancing research in the field of DNA sequencing.[142] The database is reliant on the structured use of metadata to give context to the sample data stored in its records.[143] Metadata is also very valuable in the context of multi-label image annotation initiatives,[144] developing new medical treatments,[145] developing new approaches to architectural design,[146] and research electronic data capture.[147] These examples highlight precisely how metadata themselves can be a valuable resource, and how considerable benefits and insights can be derived from their analysis, without needing to analyse the raw data to which they pertain.

If, therefore, it was not possible for a public authority to release an anonymised dataset, for instance due to high risks associated with its disclosure, some, if not all, of the aims, objectives and motivations of an FOI request for that dataset may still be achievable from an analysis of the metadata of that dataset alone. Ergo, the disclosure of metadata could still lead to many of the purported benefits of the FOIA being achieved (i.e. allowing for public sector data to be used for purposes in which there is a public interest), even if the raw data could not, for whatever reason, be disclosed.

6.3. Privacy enhancing technologies (PETs)

[141] See, for example: Y de Montjoye, Y. et al. 'Unique in the shopping mall: On the reidentifiability of credit card metadata' [2015] 347 Science 536-539. See also: J Mayer, P Mutchler and C Mitchell, 'Evaluating the privacy properties of telephone metadata' [2016] 113(20) *PNAS* 5536-5541.

[142] 'BioSample' (NCBI) <https://www.ncbi.nlm.nih.gov/biosample/> accessed June 2020.

[143] See: M Courtot et al. 'BioSamples database: an updated sample metadata hub' [2019] 47(1) Nucleic Acids *Research* 1172-1178.

[144] J Johnson, F Ballan, and L Fei-Fei, 'Love Thy Neighbors: Image Annotation by Exploiting Image Metadata' [2015] The IEEE International Conference on Computer Vision (ICCV) 4624-4632.

[145] M Pereañez, M. et al. 'Patient Metadata-Constrained Shape Models for Cardiac Image Segmentation' [2015] 9534 Revised Selected Papers of the 6th International Workshop on Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges 98-107; S Hill, 'How do we know what we know? Discovering neuroscience data sets through minimal metadata' [2016] 18 *Nature Reviews: Neuroscience* 735-736; L Nieroda et al. 'iRODS metadata management for a cancer genome analysis workflow' [2019] BMC Bioinformatics 20-29; K Jordan et al. 'Astrocyte-Mediated Neuromodulatory Regulation in Preclinical ALS: A Metadata Analysis' [2018] 12(491) Frontiers in Cellular Neuroscience.

[146] A Bhattacharya et al. 'Automated Metadata Construction to Support Portable Building Applications' [2015] Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments 3-12; B Balaji et al. 'Brick: Towards a Unified Metadata Schema For Buildings' [2016] Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments 41-50.

[147] P Harris et al. 'Research electronic data capture (REDCap)–A metadata-driven methodology and workflow process for providing translational research informatics support' [2009] 42(2) Journal of Biomedical Informatics 377-381.

The term "privacy enhancing technologies" (PETs) refers to the use of technology to help achieve compliance with rules and laws relating to privacy and data protection.[148] These technologies come in a variety of forms, and typically aim to allow individuals to protect themselves from harms associated with the processing of their personal data. Some are concerned with providing individuals with anonymity. Others are concerned with preventing network invasions, ensuring the integrity and security of communications and other data processing activities, and providing individuals with a means of identity management.[149] One specific area of deployment where PETs may be particularly useful is in the context of risk-assessments relating to the disclosure of anonymised data under the FOIA. Specifically, some PETs, if combined with a metadata catalogue or other publication scheme, as explained previously, could facilitate query-based access to anonymised data held by public authorities.

The term "query-based access" refers to a situation in which individuals have the opportunity to interact with data within a dataset by posing queries, typically over a secure internet connection.[150] By asking individuals making an FOI request in respect of anonymised data what type of query they wished to pose (for example, either an aggregate level query or an individual level query), a different PET could be applied to the data as a means of mitigating disclosure risk. One example of a tool that could be particularly useful in this regard is differential privacy. Differential privacy is a set of techniques whereby the results of queries made by an individual in respect of a dataset are altered, often by the addition of noise, so that any released information does not reveal the identity of any individual contained within the dataset with certainty. By using differential privacy in the context of a query-based access framework, therefore, the results of queries submitted to anonymised data held by public authorities would be altered so that the information released to the individuals making such

---

[148] 'Protecting privacy in practice: The current use, development and limits of Privacy Enhancing Technologies in data analysis' [2019] The Royal Society. Available at: https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/privacy-enhancing-technologies-report.pdf; S Kenny, 'An Introduction to Privacy Enhancing Technologies' [2008] IAPP. Available at: https://iapp.org/news/a/2008-05-introduction-to-privacy-enhancing-technologies/; See also: J Argyrakis, S Gritzalis and C Kioulafas, Privacy Enhancing Technologies: A Review. in R Traunmüller (ed), Electronic Government. EGOC 2003. Lecture Notes in Computer Science. Volume 2739 (Springer 2003) 282-287; V Senicar, et al. 'Privacy-Enhancing Technologies – approaches and development' [2003] 25(2) Computer Standards & Interfaces 147-158; H Burkett, Privacy-Enhancing Technologies: Typology, Critique, Vision. In P Agre and M Rotenberg (eds) Technology and Privacy: The New Landscape (MIT Press 1998)

[149] Y Shen and S Pearson, 'Privacy Enhancing Technologies: A Review' [2011] HP Labs. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.377.2136&rep=rep1&type=pdf. See also: H Pearce, 'Systems thinking, big data, and data protection law: Using Ackoff's Interactive Planning to respond to emergent policy challenges' [2016] 18(4) European Journal of Law Reform 478-504; H Pearce, 'Online Data Transactions, Consent and Big Data: Technological Solutions to Technological Problems?' [2015] 21(6) Computer and Telecommunications Law Review 149-153.

[150] S Kinney et al. 'Data Confidentiality: The Next Five Years Summary and Guide to Papers' [2009] 125 Journal of Privacy & Confidentiality.

queries would not be capable of revealing an individual's data with certainty. Crucially, the use of query-based access methods in the immediate context is that they would allow individuals to make, and receive the results from, statistical queries in respect of anonymised data held by public authorities without needing direct access to the underlying datasets. In so doing, this would circumvent many of the threats of de-anonymisation discussed above. The possibility of incorporating PET-supported query-based access controls as a means of managing a mitigating the risks associated with the disclosure of anonymised public sector data is something that has been specifically acknowledged by a recent report on PETs published by the Royal Society.[151] It has also been discussed in the scholarly literature, and identified as an area in need of further research.[152]

7. Conclusion

As noted at its outset, the purpose of this article was to build on the pre-existing literature in the field that has argued that the FOIA's approach to disclosing anonymised data is unfit for purpose and in need of reform. Specifically, the article aimed to begin to explore what shape this reform might take. To this end, the article started by arguing that one attractive mode of reform would be for the FOIA to shift from its current "release and forget" ethos, to an approach that more closely embodied a "release and remember", ethos. In this regard, it was argued that a data protection by design approach should be utilised to "build in" mechanisms that allowed public authorities to exercise post-disclosure control over anonymised data released under the FOIA.

The notion of data licensing as a form of post-disclosure control was then discussed. A series of different data licensing models was considered, and it was argued that a system of data licensing, through which licensing provisions and restrictions were applied on a sliding scale, commensurate with the risk-level associated with the disclosure of anonymised data, represented a promising model for a new "release and remember" approach. Specifically, it was highlighted how there is an observable appetite for the development of licensing schemes in relation to matters pertaining to both public sector data and personal data in the academic and scholarly literature, and how such an approach would broadly be consistent with existing legislative frameworks and policy objectives. At this stage, it was also highlighted that such an approach would necessarily require public authorities to engage with the notion of risk, and how risk-levels associated with the disclosure of anonymised data under the FOIA could be calculated.

To outline how such calculations could be performed, the article's fifth section examined the notion of risk-based approaches to regulation, both in a general sense, and with a particular focus on risk-

[151] The Royal Society (n 148).
[152] I Rubinstein and W Hartzog (n 113).

based approaches to data protection. From this, it was identified that determining the level of risk of disclosing anonymised data under the FOIA will require a consideration of both the probability of an individual being re-identified from those data, and the level of harm that would be experienced by that individual were re-identification to occur. How such assessments could be undertaken was then explained. The final section identified some other issues and factors that could play an important role in the development of a new model for the disclosure of anonymised datasets under the FOIA. Specifically, the incorporation of contextual controls in the preliminary stages of the FOIA disclosure procedure, the significance of metadata, and the role of privacy enhancing technologies. Having elucidated how a risk-based data licensing approach to disclosing anonymised data under the FOIA could help address some of the deficiencies inherent in the current "release and forget" approach. There is now ample scope for future research to consider how the abovementioned elements and concepts could come together to form a new FOIA disclosure process for anonymised data.