Analytical Report n3

Analytical Report 3: Open Data and Privacy

EUROPEAN
DATA PORTAL

This study has been prepared by the University of Southampton as part of the European Data Portal. The European Data Portal is an initiative of the European Commission, implemented with the support of a consortium[i] led by Capgemini Invent, including Intrasoft International, Fraunhofer Fokus, con.terra, Sogeti, 52North, Time.Lex, the Lisbon Council, and the University of Southampton. The Publications Office of the European Union is responsible for contract management of the European Data Portal.

For more information about this paper, please contact:

European Commission
Directorate General for Communications Networks, Content and Technology
Unit G.1 Data Policy and Innovation
Daniele Rizzi – Policy Officer
Email: daniele.rizzi@ec.europa.eu

European Data Portal
Gianfranco Cecconi, European Data Portal Lead
Email: gianfranco.cecconi@capgemini.com

Written by:
Elena Simperl
Email: elena.simperl@kcl.ac.uk
Kieron O'Hara
Email: kmo@ecs.soton.ac.uk

Richard Gomer
Email: r.gomer@soton.ac.uk

Last update: 15.07.2020
www: https://europeandataportal.eu/
@: info@europeandataportal.eu

DISCLAIMER
By the European Commission, Directorate-General of Communications Networks, Content and Technology. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the Commission. The Commission does not guarantee the accuracy of the data included in this study. Neither the Commission nor any person acting on the Commission's behalf may be held responsible for the use, which may be made of the information contained therein.

Luxembourg: Publications Office of the European Union, 2020

OA-BF-20-003-EN-N ISBN: 978-92-78-41895-3 ISSN: 2600-0601 doi: 10.2830/532195

[i] At the time this report was first issued the consortium consisted of: Capgemini Invent, Intrasoft International, Fraunhofer Fokus, con.terra, Sogeti, the Open Data Institute, Time.Lex, and the University of Southampton.

Executive Summary

Open Data is an important means of increasing access to data to citizens, companies and civil society, and can promote economic growth, scientific research and political and corporate accountability. However, much of the most valuable data is personal data, whose publication out in the open may threaten privacy. Indeed, in most cases, publishing personal data in the open will be illegal.

There are ways to manage the risk of publishing such valuable data about people, notably anonymising datasets to render the data within the file non-personal. However, anonymisation is a reversible operation, and so it is incumbent upon data controllers to ensure that the risk of deanonymisation is acceptably low.

A successful and sustainable Open Data programme should be based on three pillars. Morally, the data publisher should consider the privacy of data subjects. Legally, data protection law must be respected. And pragmatically, public confidence has to be maintained.

Publishers of Open Data are concerned less with the outcome of open datasets, than with **the processes of opening data**. These, which involve technical expertise, communication with stakeholders and monitoring data use, are mapped in a lifecycle of Open Data, based on interviews with practitioners. The lifecycle is descriptive, but could also be normative of best practice. The **balance of risk and utility is key**, and risk aversion is a common attitude.

To promote **the utility of data while ensuring data controllers' obligation to respect the right of data subjects to personal data protection**, this report has developed a series of 8 recommendations, as follows:

1. **Understand the data.** Consider potential use cases, the value of the data and potential risks.
2. **Consult.** Engage stakeholders about the publication programme, be mindful of additional risks that are identified.
3. **Remember the three pillars of privacy, data protection and public confidence.**
4. **Be very sure of the grounds for publishing personal data.**
5. **Anonymise well and thoroughly.** Follow guidelines for anonymising personal data.
6. **Remember utility.** There is no point publishing data which has been denuded of serious content.
7. **Don't release and forget.** Anonymisation and Open Data are not cheap options.
8. **Have a plan in place in the event of a problem.** Be not only transparent, but also transparent about your transparency.

1 Introduction: Open Data and Personal Data

The world's advanced economies are increasingly driven by data. This new imperative rests on massive and permanent changes in both the demand for and the supply of data. As more action and behaviour takes place online, events are digitised and leave symbolic traces which can be copied, searched, aggregated and stored. More effective technologies are improving our analytics, and providing them in real time. Amalgamating heterogeneous data creates so-called Big Data, which goes beyond sampling to represent nearly everyone or everything in a population. And as this happens, correlations become more significant – detected relationships between variables are far less likely to be down to sampling errors (Ayres 2007).

In this new data-driven world, data is inherently valuable. This is most obviously demonstrated by companies which are built to gather, marshal and analyse data, such as Google and Facebook. Their data describes the human world in such detail that it has revolutionised a number of industries, notably marketing (Mayer-Schönberger & Cukier 2013). In science, many deeply complex systems, such as the world's climate, biological systems and telecommunication networks, are now being described using data, which is already allowing effective interventions to be designed. Coordination within complex systems can be made more efficient. Data can, under the right circumstances, empower those capable of working with it, and can be used to hold authorities, companies and individuals to account, both for the legitimacy/legality of their actions, and their performance against set criteria (Bradshaw 2014, Margetts 2014). Transparency has often been a means to regulate companies' actions, relying on the market rather than law to constrain them (for instance, companies have to publish accounts, and release data about their environmental impact – Fung et al 2007, Gurin & Noveck 2014).

Because of this move towards Big Data to create value, it is understood that data's value resides in its use, and that the more data that can be applied to a problem, the better. The collector of data, or the holder of a database, has a number of rights to licence the use of the data, which will increase the barriers to entry using that data (cf. e.g. Reed 2007). To that end, there has been a move in the last decade or so towards the promotion of *Open Data*. Data which can be used without restriction will, all things being equal, be used more.

There are other arguments for opening data. For example, data collected by the state is legitimised by citizens electing the government, and funded by taxpayers, and so there is an argument that citizens/taxpayers should have access to non-sensitive data and services built off the back of it, if they would find them valuable. Government data is citizens' data, in an analogous sense to which government money is taxpayers' money. Furthermore, government needs to be held to account by its citizens at the ballot box, rather than by expert auditors, and so the information required to do so needs to be generally available (this is traditionally done via the mass media, but it could happen via app developers, citizen journalists or data journalists with access to Open Data). Hence, government has a specific duty that goes beyond that of the private sector to facilitate access to data. Meanwhile, in the private sector, the network effects of opening data collected by private firms will in some sectors outweigh the competitive advantage created by keeping the data closed (analogous to the ways in which opening documents, such as inventories, onto the World Wide Web is counterintuitively more valuable than keeping them private, a lesson that some firms took a long time to learn as e-commerce developed).

In this report, we will review concepts and research around Open Data and privacy, supplemented by interviews and discussions with stakeholders in the chain of data stewardship (see Acknowledgements below for details). In total, we interviewed 7 different individuals, covering data publishers, data owners and data consumers, from government and private sector, across several EU countries, although focussed on UK government publishers. A number of key themes did emerge, often common to many or all of the people we spoke to. Where appropriate, we have provided direct quotes from the people we interviewed, alongside the discussion of the issues raised.

In the next section, we define and set out concepts relevant to openness and privacy. Then we will consider the importance of perceptions and the patterns of practice uncovered in our interviews, before making recommendations and concluding.

2 Openness and Privacy: Key Concepts

Open Data is that portion of data which is open to be used without constraint, either legal or technological. To that end, it is usually taken to mean data that is:

1. Online. To be accessible, it needs to be possible to get hold of data easily. Clearly the simplest way to make that happen is to put the data online, making it available via a download from the World Wide Web.
2. Machine readable. Analytics and data fusion (i.e. aligning heterogeneous datasets into a consistent representation) need to be automated in order to be feasible on big datasets, or on a large number of datasets.
3. Under an open licence. An open licence puts minimal constraints in place for the user. Examples include Creative Commons (in particular, CC or CC-BY 4.0) or more specific ones such as the UK Open Government Licence at http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/.

These conditions have a number of implications. The drive to lower the barriers to use Open Data favours the use of neutral formats such as CSV and RDF over proprietary ones such as Excel. The open licence will make it difficult to apply standard disclosure control methods, such as access or query controls. And the range of uses for which the data is released is not constrained. Services based on the data may be monetised or free. The idea of monetised services based (in full or in part) on Open Data does not contradict the philosophy of openness. Because access to data is unrestricted, services based on the data can only be sold on the basis of value-added. If the service adds little or no value to the data, then the availability of the data will allow other services to compete. Openness prevents rent-seeking on the part of data licensees – it is no longer possible simply to charge for access if data is open.

We will call those who make data open *publishers*, and those who download and use the data *consumers*. Clearly an organisation (or individual) could be simultaneously a publisher and a consumer.

Because of the lack of restriction, the data, like any other tool, may be used by cyber criminals as well as climate scientists and epidemiologists. Hence it is important to ensure that Open Data has a positive net social value, rather than simply exposing data that can be exploited. However, this is more easily said than done. The most valuable data is that most relevant to people's concerns, which includes data used to administer the public and private sector services they consume. Furthermore, data about institutions cannot easily be separated from data about people. For example, if we use performance data to hold institutions to account, the performance of a school, a hospital or a prison will depend at least in part on data about pupils, patients or prisoners.

Hence, broadly speaking, privacy will be an inevitable concern about Open Data. If the Open Data includes information about identifiable individuals, then their privacy will thereby be breached; the more sensitive the data, the more serious the breach. It is also possible that information ab out individuals could be generated by inference from non-personal data in an open dataset. Thirdly, more seriously, there is the possibility that information about individuals could be produced by inference from data in the open dataset augmented by other information in the intruder's hands (or even information readily available elsewhere).

> *"Something that was raised at a meeting, where we gave a presentation, were privacy concerns, and we have seen that many start-ups are not aware of that topic – it's an awareness question."*

For this reason, releases of data in the open must be scrutinised carefully to ensure that the risk of a privacy breach is as low as possible, unless there was some other reason why privacy was not a priority (for example, there may be a public interest in the publication of some personal information, such as the identities of suspected criminals).

2.1 Privacy

Privacy is a complex condition – or, more accurately, it is a condition about which individuals, groups and societies have highly context-sensitive and sometimes conflicting preferences against a background of culture-based social norms and varying legal approaches (Margulis 2011, O'Hara 2016). Privacy is a state of non-interference where a boundary suggested by the use of first person possessive language ("this is my/our business") has not been crossed by others, where a privileged status has not been contradicted, or where one is not the object of attention. On many occasions, there are good social reasons why one's privacy should be breached, and through social norms and regulations we develop expectations of where our privacy will be respected and where not.

Privacy and Data Protection

In reading this document it is vital to bear in mind the meaning of, and differences between, privacy and data protection. These terms are clearly defined in the text where they are introduced.

Privacy is the most controversial, and in many ways is an open philosophical question. It is *not* a legal concept (although there are regulations about it, and it is the subject of much case law). Its precise significance is culturally-relative and, for the purposes of this paper, it is not essential to resolve these philosophical debates. In this paper we define privacy as a state of non-interference with an autonomous being, where the boundary of the being's self is not crossed, where a privileged status has not been contradicted, or where the being is not the object of attention.

There are many types of privacy (private property, decisional privacy, ideological/religious privacy, private space), but in the open data world the key type is **informational privacy**, defined as a state where information about the person is not in the possession of others.

Data protection rights allow data subjects to ensure that data about them is accurate and proportionate to the purpose for which it was collected, and that the processing is fair and lawful. Data protection is a legal concept (defined by statute) designed to determine the rights to control information of data controllers and data subjects.

As explained in section 2.2, data protection is *not* privacy. Because it is defined by law, its application and scope are clearer. Data subjects can use data protection rights to protect their privacy (if they so wish) to a limited degree. They do *not* give subjects the right to suppress fair processing of accurate data, and there are special protections for media publication in the public interest.

Privacy has many areas of application – private property, private spaces, private decisions, for instance – but the key notion for Open Data is 'informational privacy' – privacy with respect to information about oneself. An individual has informational privacy to the extent that data about him or her is not in the possession of others. As an initial stab at a definition, we can argue that an individual X has not got privacy relative to an individual or organisation Y when Y possesses, or has access to, data about X. A stronger definition would be to say that X's privacy is lost when data about him or her is processed. In the case of Open Data, these definitions are distinct. On the first definition, merely publishing data would be a breach of privacy, because anyone with an Internet connection would have access to the data (whether or not they even knew it existed). On the second definition, privacy is threatened only when the data is downloaded and processed by a consumer. In terms of European law in the Data Protection Directive of 1995 and in the General Data Protection Regulation 2016 (GDPR) the definitions are equivalent because processing includes "dissemination or otherwise making available").

Some details about X might be private and others not relative to a data consumer – Y may possess his name and address, but not his age (which remains private as far as Y is concerned). X's address may be known by his doctor, employer and mobile phone company, but be private as far as other companies, or the police, or the general public is concerned.

2.1.1 Defining Privacy

Although control is not necessary for privacy, X's privacy is undoubtedly enhanced if he has some control over who receives it. If X has allowed his address to be published in the phone book or the electoral roll, then it is not private in this general sense. Y may not as a matter of fact know X's address, but she can find it if she likes (if she knows X's name) and there is very little X can do to stop her. The nature of Open Data means that information about X within an open dataset ceases to be private in this more general sense – even if no-one ever downloads the information about X, X has no control.

There is nothing inherently wrong or damaging in releasing information about people. X may not care whether the information is private (X may not mind his height, or the colour of his hair, being publicly known, especially as these might be gleaned easily from personal acquaintance). X may wish the information to be publicly known – for example, that X scored the winning goal in the Champions League Final may be something of which he is very proud. More to the point, people's preferences vary dramatically about what should be private. Someone's sexuality is generally not a secret, but different people will have different views about whether it may be published in a searchable database. It is hard in such circumstances to discriminate between people's different attitudes. Suppose X does not mind his sexuality being made public, but that Y wishes to keep hers private. An open dataset which includes X's sexuality and not Y's may be seen by some as a solution to this problem, but it won't do the trick in general. Users of the data may conclude (rightly or wrongly) that Y's desire to conceal her sexuality implies something about her sexuality.

This leads us to extend our notion of privacy. In this last example, Y's privacy with respect to her sexuality has been invaded not because information about her is known (inferences made about her sexuality may be wrong), but because she has been the subject of impertinent enquiry and inference, fuelled by the publication of information about X and others.

2.1.2 Harm, Sensitivity, Confidentiality, Norms and the Public Interest

Because views of privacy vary (some people appear on *Big Brother* while others are recluses), any set of general rules will be too coarse to align with people's actual preferences. Hence, when considering a privacy-relevant release of data, key factors include *harms*, *sensitivity*, *norms*, *confidentiality* and the *public interest*.

Privacy harms are often hard to quantify. It is relatively rare for people to suffer financial loss or physical harm as a result of a privacy breach (Solove 2014), but it is not unknown. Following harm such as this the privacy-breaching agent may be liable, and have to compensate the victim. More likely is reputational damage, which is also a considered distinct harm in most legal jurisdictions. But for most people the harm is a breach of their rights. Article 8 of the European Convention on Human Rights, and Article 7 of the Charter of Fundamental Rights of the European Union both enshrine a right to a private life. Where the line is drawn between private and public life is therefore important.

Certain types of information are generally more sensitive than others. Information about health, politics and finance concern more people than information about family background or employment. Sensitivity of information is linked to the harms that may be caused by a privacy breach, although it is not a direct function of them. And once more, sensitive information may be implied by the release of non-sensitive information. If the information about 99 people who have tested negative for an infectious disease is published in the open, then there is an implication (possibly false but damaging nonetheless) that the 100th, about which the data is silent, may be concealing something.

When data has been gathered under an assumption of confidentiality, then it would be unwise in the general case, not to say morally dubious, to publish in the open. Confidential information is not private *per se* (as it is held by a third party), but the third party's passing it on to someone else is implicitly or explicitly ruled out. Confidential information includes medical and legal information, but can also apply to companies (commercial confidence). For instance, two companies making a contract may need to supply information about themselves to each other, in which case that information should be treated as confidential. There are many who argue that contracts should be made open, which would sometimes result in commercially confidential information being revealed or becoming inferable. Note that revealing commercially confidential information is not strictly a privacy breach, as it may not provide any information about individuals at all.

Within societies, privacy is regulated by social norms, which the publishers of Open Data also need to take into account (Nissenbaum 2010). Norms vary across culture – for instance, in the UK financial details about income and wealth are normally seen as a private matter, while in Scandinavia, tax records may be published in the open, as this is seen as a public matter. Even if the publication of confidential information is not ruled out by law or by a contract between the parties, the norm is that it should be kept in confidence where possible. Norms may or may not fit the preferences of individuals, but they need to be taken seriously by data controllers, as they govern the expectations of data subjects.

However, there is also a public interest in the publication of information (Etzioni 1999, Jarvis 2011). The public interest in the availability of data may outweigh the private interests of the subjects of the data in keeping it under wraps. In that case, such data may be publishable in the open – but this is of course not a judgment that the data controller should take solely upon herself. The public interest is a political matter, and as such ultimately must be assessed impartially by a legitimate body. Without prior legal backing, the data controller has no legitimacy to make the decision as to whether publication is in the public interest.

2.2 Data Protection

Although privacy is a human right in the European Union, privacy itself does not necessarily loom large in domestic laws across the EU. Privacy is usually protected by rules for *data protection*. Data protection is not in itself a protection of privacy; indeed, the EU Data Protection Directive was designed in the context of the EU Single Market to ensure cross-border data flows, and therefore is certainly not designed primarily as a privacy protection. Rather, its aim is to regulate data flows so that privacy and freedom of information, and the interests of data subjects and data consumers, can be balanced. Data protection avoids many aspects of the personal and idiosyncratic nature of privacy preferences with a rule-based system embedded in law (Hildebrandt 2015). It places few domain-specific restrictions on data controllers, but furnishes data subjects (in theory at least) with strong rights to access, correct and control data that is about them. Data controllers must be able to show that their processing of personal data meets one of a number of conditions, of which perhaps the most powerful for data subjects is that the subject him- or herself has given consent for the data to be processed for a specified purpose (and if the data is to be processed for another purpose, then consent must usually be sought anew). The rules for data protection are independent of the preferences or economic incentives of the controller or the subject, which makes their application much more determinate, and forces compromise upon the various actors. On the other hand, balancing the different preferences for and against privacy, and the interests of individuals, corporations and society as a whole is a much harder problem which no-one has yet managed to encode in a straightforward set of rules. Data subjects have privacy rights, but these are hard to interpret. More recently, data protection has been proclaimed as a human right in the EU's Fundamental Charter (Article 8).

A breach of data protection in Europe can be punished with a fine from a national Data Protection Authority (DPA) of the country concerned and so the constraints on publication go beyond the social. Data protection governs the *processing* of data, and the person held responsible for processing is called the *data controller*. The data that is regulated is called *personal data*, which is defined as data from which a living data subject is identifiable by the data controller (even if the identification requires the data controller to use auxiliary data, as long as she is in possession or is likely to be in possession of that data). It follows that personal data is relative to the data controller – data is personal or not for a particular data controller, and may have a different status if it passes to another controller.

When personal data is collected, the controller must specify to the subject what purpose the data is collected for, and for how long it will be stored. The controller is not allowed to collect more data than is needed for the task, or to keep the data after the task is completed. In general, there are six conditions for the legitimate processing of personal data, at least one of which has to hold.

- The data subject consents to the processing.
- Processing is necessary for the performance of a contract to which the subject is party.
- Processing is necessary for the controller to comply with a legal obligation.
- Processing is necessary to protect the vital interests of the subject.
- Processing is necessary for an action in the public interest or for the exercise of official authority.
- The processing is in the legitimate interests of the controller.

If the controller, having completed the task for which the data was originally collected, wishes to use the data for another task, then she must ensure that at least one of these conditions holds for the further task. If the processing is to rest on the subject's consent, then she must obtain the subject's consent anew for the further task, unless another legal basis, such as legitimate interest, is applicable.

2.2.1 Publishing Personal Data in the Open

'Processing' includes "disclosure by transmission, dissemination or otherwise making available," and it therefore follows that it is illegal to publish personal data as Open Data unless one of the six conditions holds. Note also that, in the specific example already quoted of the UK's Open Government License (OGL), personal data is specifically not covered by the licence, along with other types of sensitive information such as intellectual property.

Two key principles of data protection, in this context, are purpose limitation (GDPR, Article 5.1.b, says that personal data must be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes;") and the transfer of personal data across borders (GDPR, Article 46.1, says that "a controller or processor may transfer personal data to a third country or an international organisation only if the controller or processor has provided appropriate safeguards"). These severely restrict the publication of personal data, because of the openness of the licence for consumers. The data may go anywhere and be used for anything. Therefore, a publisher publishing under a totally open licence cannot determine where the data will end up, and cannot restrict the purpose of the processing. It is hard to see a way around the second principle, although a carefully-drawn specification of the purposes of processing that leaves open the possibility of later publication may get around the first. Furthermore, licenses that place obligations on data processors to make derivative works publically available (such as CC-BY-SA), may put data consumers in an impossible legal position where republication is required under the terms of the license but prohibited by data protection legislation.

Nevertheless, it is possible to publish personal data in the open – either because it has been anonymised so that data protection obligations no longer apply (see below), or because there are legitimate grounds to publish the personal data openly. In practice, a given dataset may combine elements of both approaches, where data about some parties is anonymised, and that of others left intact. Which is most appropriate will depend on factors such as whether subject consent to open publishing was obtained at collection, the inherent sensitivity of the dataset and the broader social and organisational context in which it is to be published. For instance, there may be different concerns and legal implications around publishing a dataset detailing the meetings of a public office holder, such as a government minister, and those of a private individual. Even in such cases, rights to privacy can go surprisingly far; in a case before the Court of Justice of the EU (Commission v Bavarian Lager 2010), it was held that the names of representatives of a trade body attending a meeting with, and organised by, the European Commission, were in the private domain and should not be published in the minutes of the meeting without the representatives' consent.

It is sometimes desirable to publish Open Data sets that contain personal information. For instance, in

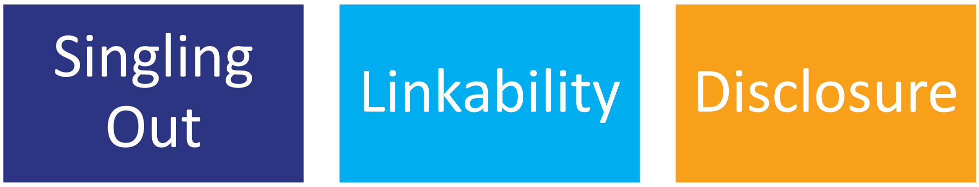

Figure 1: Three common forms of disclosure

the UK, equipment funded by the public Research Councils is published by the UNIQUIP project[1], including details of the contact person for each piece of equipment. Likewise, research grants, and the recipients of the awards themselves, are published in publicly accessible formats. The addresses of company directors are published by Companies House.[2] In other European countries, there are examples such as the publication of court records as Open Data – an important resource as it forms case law and determines precedent – containing the names of presiding judges and the lawyers involved, but not (typically) the names of the plaintiffs, witnesses or defendants.

Determining the legality of publishing such data is non-trivial. In addition to EU-wide data protection laws, other national laws may apply in some domains; such as the UK's Rehabilitation of Offenders Act 1974, which allows offenders to suppress the mention of certain minor convictions after a period of time. From 2018 onwards, the General Data Protection Regulation (GDPR) will replace the Data Protection Directive, altering, in some key aspects, the laws surrounding the open publication of personal data.

2.2.2 Data and Inference

Despite the rule-based nature of data protection, there are still indeterminate issues, most notably with the decision as to whether data is personal. To be personal, the subject needs to be identifiable. However, much will depend on a range of factors – what auxiliary data is available, how the data is governed, whether there are access or query controls over it, and whether there are firewalls between the dataset and auxiliary data (Duncan et al 2011). In the case of Open Data, there are few or no controls on access, and so everything depends on the auxiliary data. This can be hard to pin down.

For example, a dataset containing reference to a person's height would not typically be personal data. That someone is 1.80m, or 1.44m, or 1.95m, is hardly identifying. However, some people are abnormally tall or short, and if one possesses further information about (say) who is the tallest person in a population, then it is easy to gather other information about that person from the data.

Even data that is not ostensibly about people can be identifying in certain respects. For instance, a bus timetable is not personal data, and its value as Open Data is clear. Yet it may be possible to glean important information from it – imagine a situation in which an estranged father is denied access to his children, but he knows that the children take the hourly bus to and from their school. He can easily deduce which buses the children will take, and therefore will know where they are at a part of the day when they are unsupervised. This, however, does not make a bus timetable personal data.

Furthermore, even if an Open Data release does not include personal data at the time of release, it may become personal as the amount of auxiliary data increases. For instance, it may be that an Open Data release is fine at the time of release, but a further release of Open Data in the future renders people in the original dataset identifiable. The non-personal data has now become personal.

So far, we have been discussing identification as if this involves associating an individual with the data. However, this is not the only potential pitfall. In general, we should consider the issue of disclosure, and this tends to come in one of three forms.

- Singling out. Is it possible to isolate records about a person from the database?
- Linkability. Is it possible to link records about the same person (or people within a group) from the database, or across databases?
- Disclosure. Is it possible to deduce, with a sufficiently high probability, the value of an attribute of an individual?

[1] http://equipment.data.ac.uk/.

[2] https://beta.companieshouse.gov.uk/.

2.2.3 Anonymisation, Pseudonymisation and Reidentification

One means of rendering personal data non-personal is to anonymise it. This involves creating a new dataset from the personal dataset in which the data is non-personal. To do this, the information content of the original data needs to be reduced. This can be done by a number of means: removing identifiers and quasi-identifiers such as names, dates of birth and postcodes, aggregating data (for example, storing age ranges rather than dates of birth), perturbing data (for example, adding or subtracting quantities to or from quantitative data, keeping means and deviations constant while changing potentially identifying values), or removing fields (for example, deleting gender).

> *"We linked datasets and then published the anonymised result. That linking introduces new issues, and what we published could be linked further."*

A less powerful type of anonymisation is pseudonymisation, where identifiers are replaced by randomly-generated identifiers. This can be done where the data's value depends on linking certain data points. For instance, in a medical database, it is far less helpful to know that a patient entered hospital on 12th March with chest pains, and that another patient was discharged on 14th March, than to know that *the same* patient entered with chest pains on 12th March and was discharged on 14th. A medical history, which is essential to understanding, requires that we can connect pieces of information about the same person, even if we do not know who that person is.

It should be noted that, in the opinion of the Article 29 Working Party of European Data Protection Regulators, pseudonymising data is highly unlikely to render it non-personal (Article 29 Data Protection Working Party, 2014). In their opinion, anonymisation requires demonstrating that data is not linkable, whereas the point of pseudonymisation is to support linkability of data about unidentified people. This opinion is upheld by recitals 26 and 28 of the GDPR.

A wider legal problem with anonymisation, however, is that unless the information content of the data is reduced below any reasonable level of utility – for example, by replacing all data values by 0 – it will always be possible technically to reidentify people from anonymised data with sufficient auxiliary data (Dwork 2006). Therefore the risk of deanonymisation cannot be reduced to zero. It is possible to reduce the risk of reidentification below an acceptable level (i.e. the costs of reidentification outweigh the benefits of so doing), but the context in which the data is held needs to be monitored constantly (Information Commissioner's Office 2012). In the Open Data world, this is even more essential (Rubinstein & Hartzog 2016). Testing the data before publication to see if it can be deanonymised easily (penetration testing) is a sensible idea, though not a cheap option.

Hence, anonymisation cannot be associated with a 'release and forget' mentality – we cannot just publish Open Data and forget about it. Sharing (and even more, publishing) data requires a stewardship mentality on the part of data controllers, and their managers need to ensure that the resources and institutional structures for discharging this duty of care will be available now and down the line. The UK Anonymisation Network guidance (UKAN) [3] identifies ten steps which enable the data controller to understand, measure and control the risk of anonymised data being compromised.

1. Describe the data situation.
2. Know your data.
3. Understand the use case.
4. Understand the legal and governance issues.
5. Understand consent and ethical obligations.

[3] http://ukanon.net/.

6. Identify the processes you will need to go through to assess disclosure risk.
7. Identify the relevant disclosure control processes.
8. Identify stakeholders and plan how to communicate with them in the event of a disclosure.
9. Plan what happens after the data is shared.
10. Plan how to react if things go wrong.

It is incumbent upon the data controller to understand why someone might want the data, what other data could be used to reidentify people or disclose their attributes, what consent governs the data, and what to do in the event of a data breach. This kind of responsible attitude to anonymised data will be taken into account if there is a data breach, and will affect the seriousness with which the relevant DPA will treat the breach.

> *"There are four million records, we can't guarantee that it's completely anonymous. We thought about how to test the anonymity though, so that [the organisations involved] could be happy when making statements to customers"*

It follows that releasing anonymised Open Data will be far less risky if a framework such as UKAN's is used to plan around the anonymisation and release processes.

Note also that deanonymising data turns the data into personal data. If the Open Data is anonymised sufficiently to render it non-personal, then its publication is legal. If the consumer deanonymises it, then it becomes personal data (for the consumer), and the consumer will fall under the scope of data protection law. At that point he should register with his national Data Protection Authority as a data controller (if he is in the EU), and data protection law will determine what processing of the data by the consumer is legal. The simple fact that one consumer has deanonymised the data does not render the data personal for other consumers, as long as the other consumers are not able to identify data subjects in the anonymised Open Data.

The anonymised dataset is still personal data *for that data controller* if subjects can be identified from the data in combination with other data in the controller's possession (and the GDPR will make this condition more stringent) Since publishing the data in the open is a kind of processing, it follows that the publisher of Open Data must have appropriate legal grounds (defined in article 6 of the GDPR) for the publication of anonymised data if the original personal data still exist as well.

2.3 Public Confidence

A third factor, alongside privacy and data protection, is public confidence. Ethically, an organisation ought to preserve privacy when it publishes Open Data, as a matter of respect for data subjects. Legally, it is obliged to follow data protection rules. But these may come to nothing if it is not *seen* to be doing these things. Confidence is a function of perceptions. Loss of confidence does not mean that harm has necessarily been done, or is even likely to happen. Organisations must of course work to prevent harm from happening, but a paradox of trust and confidence is that transparency about such work may serve to increase rather than decrease perceptions of risk (by raising previously undreamt-of possibilities).

Surveys of public attitudes to data sharing (particularly government data sharing) are relatively consistent in at least some jurisdictions (Castell et al 2014, Cameron et al 2014, Deloitte 2014). In the UK, the public tends to disapprove of data sharing as a general proposition, but is relatively tolerant of particular use cases where a public benefit is clearly visible. Sharing data for commercial use is not generally supported. Meanwhile data security remains a concern; the UK public worries about what happens to data once it is shared. There are fewer studies of public attitude to Open Data, but – as Open Data is a special and limiting case of data sharing – it is clear that public confidence in Open Data may be relatively fragile.

Loss of public confidence feeds directly into reputational damage. Different stakeholders naturally have different concerns. Publishers may be concerned about reputational damage to the organisation (or a sub-unit), to the industrial sector as a whole (e.g. government, the charity sector, or the telecom sector), or even about Open Data as a concept. Reputational damage for publishers may harm relations with suppliers of data and even data subjects. Public confidence will determine whether the public is minded to share data with an organisation; if it thinks that data will be irresponsibly published in the open, then the organisation may ultimately be starved of data, as false data may be input (people may create untrue profiles to interact with the organisation – Van Kleek et al 2016), services may not be used, or consent to process may not be given.

Data consumers, on the other hand, are concerned about reputational harm (guilt by association) from the use of data that was improperly released (as well, of course, from any improper processing of it of their own). Poor anonymisation practices on the part of publishers may result in data released as non-personal being judged as personal at a later date. A backlash against Open Data more widely may hurt companies (perhaps especially small start-ups) whose business model depends on exploiting Open Data.

It follows that an organisation must be open and transparent about the Open Data it processes, clear about how the data has been anonymised (if it has been), and clear that it is a responsible steward of the data. If Open Data is about transparency, then organisations must be transparent *about* their transparency in order to be seen as responsible. Note also that an organisation may be held to account by expert and media commentary. Understanding of a problem need not be very widespread for there to be a wide perception that there is a problem of some sort.

> *"The irony is that as citizens we want to know all about everyone else, but nobody to know about us. We need to overcome people's mistrust of large organisations to use information properly."*

3 Perceptions and Practice

Open Data is a relatively clear concept – data that is open according to the definition in the previous section. But the focus of practitioners is largely on the process of *opening* data. Data that has already been collected (which in itself may require a highly complex set of processes) needs to be acted upon by data controllers to open it to a wider public. These actions create a context for perceptions of risk and a dialectic of transparency which will influence what data gets published in the open, and how.

3.1 Motivations

As noted in the introduction, there are a number of theoretical justifications for publishing data, and with respect to our interviewees these were reflected in the typical motivations for publication. There is a strong sense, certainly with governments and some officials, of the ethical (and sometimes statutory) imperatives to publish data. Such publication could enhance the welfare of data subjects, or of the population as a whole (for example, by improving healthcare or transport). Or alternatively, it might improve the reputational profile of the publisher to work *pro bono*.

Some were acting on orders, which presumably reflected the strategies of senior management. Others wanted to open the data to wider uses to get more value from the data, recognising that something inherently valuable had been created, and that publication was necessary to realise the value (even though value would at best accrue to the publishing organisation indirectly). Still others were motivated around the creation of value across sectors, so wished to open data up to other actors. Value-centric motivations put pressure on personal data, because that is where, typically, most value lies. Hence open datasets derived from personal data, with all their inherent privacy risks, are likely to be important for the foreseeable future.

3.2 The Open Data Lifecycle

There are existing guidelines on the publication of Open Data, such as the EDP Goldbook (European Data Portal, 2016), though these do not typically address the additional complexity introduced by personal data. Nonetheless, on the basis of the interviews we conducted, it became clear that a *de facto* lifecycle was emerging. Not all actors followed every single step, depending on context. But monitoring and consultation seem to be common to many.

The emerging lifecycle is captured in Figure 2. The steps are as follows.

1. Collection. Data is collected from and/or about subjects. The specified purposes, and any subject consent collected at this point, will have bearing on later publication and usage. Hence this step needs to be managed well in order to allow the data to be opened up at a future date.
2. Transfer. Data may be passed from one controller to another. Such sharing might have to involve anonymisation, depending on the original purpose of collection, the status of the consent, the contractual arrangements surrounding the data, and so on.
3. Aggregation. Data may be combined, prior to publication, with other datasets, either by the original controller or a third party. The resulting dataset may then be heterogeneous in terms of data source, original collection purposes and associated subject consent.

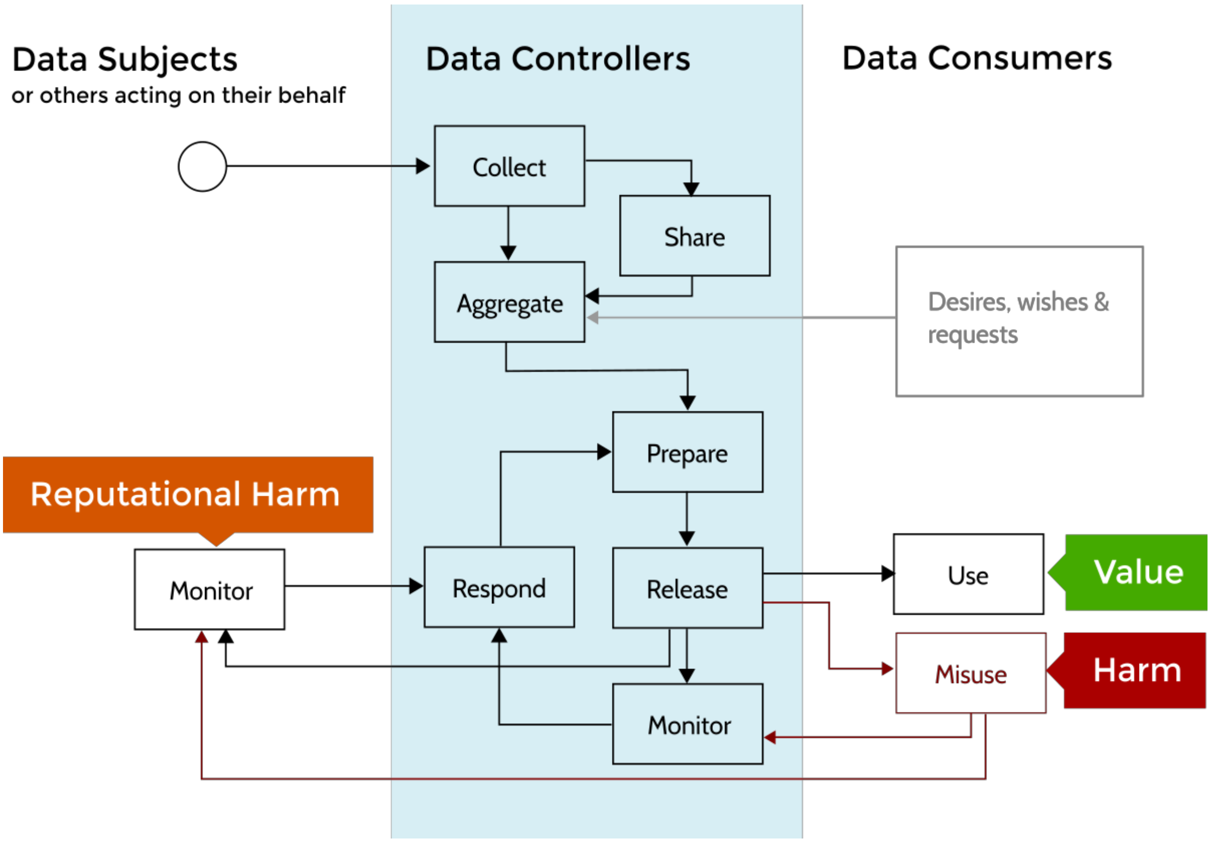

Figure 2: The emerging Open Data lifecycle

> *"None of the data was owned by [our organisation], all the [source] organisations had to be happy, and confirm the legality of releasing it."*

4. Preparation. Data needs treatment prior to publication. Often this involves 'cleaning up' the data, removing egregious errors or inconsistencies, and generally improving quality. Preparation is sometimes needed to reduce the risk of the publication, by removing more sensitive aspects (from a privacy or commercial perspective) or by anonymising it such that it no longer constitutes personal data.
5. Release. At some point, the dataset is made available. In some cases, this involves opening the whole dataset, in others only portions of it may be released in a fully open fashion.
6. Usage. The data is obtained and processed by third parties. If the released data still constitutes personal data, or becomes personal data because of other data held by those parties, there will be obligations on those third parties from EU data protection laws. Ultimately, usage is where the value of the data is realised. However, *misuse* of the data is a potential source of harm to data subjects. Linking datasets together is an important aspect of post-publication processing, with the potential to increase the sensitivity of, or even deanonymise, those datasets, and so represents a particular risk to Open Data derived from personal data.
7. Monitoring. Following publication, the responsible body may (and we argue should) monitor the usage and context of the released data. This could be done actively (although this is problematic for truly Open Data where no registration is required) or by listening carefully to concerns raised by interested parties. Data publishers should be particularly mindful of the risks of deanonymisation, but other data protection concerns with accuracy, relevance, and the right to be forgotten may also arise. Crucially, other parties with an interest in the data may make attempts to monitor the usage of the data or the risks associated with it. This monitoring, while reducing the risk of harm to data subjects, could pose a source of reputational harm to the data publisher. *Ideally, monitoring would be a collaborative activity undertaken between data publishers and other interested parties* (O'Hara 2012).

> *"There's a general degree of monitoring on usage because we work under the auspices of [the Statistics and Registration Act 2007] and we try to understand that we listen to our user community and responding to their wants and needs."*

8. Responding. In the event that issues arise following the publication of an Open Dataset, publishers should be able to act to mitigate them. This could include altering, or even withdrawing, the published dataset itself, notifying third parties that are using that data (where possible) and potentially notifying data subjects and other stakeholders directly.

This is a descriptive list of steps, but clearly has potential for normative guidance, especially in combination with the UKAN framework for anonymisation discussed above. For example, it follows that publishers should provide and maintain forums for responding and mechanisms for collecting responses.

3.3 Risk

Many discussions within publishing organisations revolve around risk management, which is clearly a concern for data controllers. However, there is little sense from the interviews that we conducted of consensus about best practice risk management. Indeed, uncertainty remains about the nature of the risks (and potential harms) that Open Data produces.

"Until 18 months ago, we wouldn't touch personal data with a barge pole! We thought that was safer."

Naturally enough publishers would like risk to be absorbed by consumers, and vice versa. To some extent, this differs from non-personal Open Data, where there is typically little risk to the data consumer. The residual risks in Open Data, such as deanonymisation (where it has been derived from personal data), or failing to ensure a legitimate interest to process the data at all (where it includes personal data), bring some of the risks to consumers. As a result, additional information published alongside the datasets to help consumers better understand any residual risks or obligations would be welcome.

"I don't think that we as an organisation can or should do [anonymization or redaction] better than [government organisations] that spent millions of Euros in that domain. How are we supposed to do that better than them? It would be kind of unfair if the problems were passed on to re-users of the data. It should be dealt with at source."

Ultimately, both publishers and consumers are concerned about the potential to harm data subjects, but harm (via deliberate or accidental misuse) is only one way in which negative consequences may arise. For publishers, and to lesser extent consumers, it is sufficient for the reasonable possibility of harm to be discovered or, crucially, perceived. Reputational harm to a data publisher or consumer, in particular, does not necessarily require actual harm to occur; it would cause reputational damage if a dataset that was believed to be anonymous was deanonymised, for instance, independently of the harms done to data subjects.

"[Failing to protect personal data] would be pretty bad for us! Our reputation depends on being trusted."

"It would be reputationally harmful for us if data was reidentified, even though the data itself is not that sensitive."

"Some of the systems that we have access to are very restricted, so we must be very precise in our use of data. We mustn't jeopardise access to internal or cross-departmental data sources."

"It's mostly the PR angle, we try to be nice to people. If someone comes to us and says 'you're doing something I don't like' you can either apologise and then not do it again, or you can say 'the law says we can do this' - but that's not a good way to go."

For consumers, the main concerns are reputational damage and an increase in costly obligations. Reputationally, the consumer is concerned with the consequences of involvement with datasets where the publisher has failed, in some way, to publish the data responsibly. And because, as noted, the status of data as personal or non-personal depends upon the resources of a data controller, downloading non-personal Open Data may render that data, or other datasets, personal. If an individual is not identifiable from dataset A or from dataset B, but is identifiable from A+B, then a consumer already in possession of B who downloads A from an Open Data publisher finds himself in possession of personal data whether or not he wishes to be so, or even whether or not he knows that he is. Because personal data brings with it a set of obligations with respect to regulators, there is a cost to possessing it, which implies a responsibility on data consumers to assess the relative risks and value of processing it.

However, publishers do not have unlimited resources or time. There are not only questions about the efficacy of techniques such as anonymisation, but identifying such techniques requires resourcing. Several interviewees mentioned that the value of the published data may decrease if there is a delay in publishing it, and thus the time taken to undertake risk management itself reduces the potential reward from publishing. Equally, the anonymisation process itself may reduce the utility of the data to such an extent that publication is not really warranted.

As with other activities that pose legal or reputational risks, organisations must allocate resources based on their own analysis of risk, benefit and individual circumstances. There is still a great deal of uncertainty over how the GDPR will be interpreted, and public opinion changes rapidly regarding personal data and privacy. As its interpretation becomes clearer, so will the balance of responsibility between the publisher of open data (whose role is to anticipate the risk) and the consumer (whose activities may constitute reckless or negligent processing). Most risky resource transfers build up a body of case law to determine where responsibility lies (for instance, selling a gun is subject to regulation).

For the publishers of open data, the risk assessment revolves around proportionality – what harms are possible with the data, how sensitive is it, is it available elsewhere, how expensive would it be to reidentify data subjects in anonymised data? For the consumers of open data, the prime objective must be, of course, to use the data in a responsible and safe way.

3.4 Uncertainties

Publishing Open Data derived from personal data also requires publishers to confront unknowns on several fronts. Those commonly reported include:

- The possibility of anonymisation (or other de-risking techniques such as removal of sensitive fields) being reversed by future techniques or in combination with other datasets.
- Uncertainty over the risks to data subjects that are posed if the data were to be deanonymised, or misused in some way.
- Uncertainty over how the law should be interpreted. Some of our interviewees were surprised at the subjectivity of their legal advisors' answers, and how they differed between people.
- Uncertainty over future regulation and technology. The ultimate effects of the GDPR, the evolving relationship between the EU and the US (not to mention the rest of the world), the evolution of data handling technologies such as the cloud, etc., are not easily predictable.

> *"Legal Advice is not as objective as I thought it was!"*

4 Approaches

Some approaches to various concerns associated with the risks to publication emerged from our discussions with interview participants and other relevant groups. The first, stakeholder dialogue, emphasises the importance of engaging with data subjects and consumers throughout the Open Data lifecycle as a way to identify and manage risk. The second, data subject consent, focuses on the rights of data subjects to a say in how their personal data is used from both a legal and ethical perspective.

4.1 Stakeholder Dialogue

Broadly, the purpose of engaging with stakeholders – primarily data subjects and consumers – is to identify (and balance) potential risks and value (O'Hara 2011). Many of our interviewees pointed to the importance of engaging with data consumers as a means of understanding what aspects of the data are most valuable, and to ensure that, where possible, these aspects are preserved when data is anonymised or otherwise de-risked. Such dialogue should, in the case of personal data, assist the publisher in identifying not only whether the data is suitable for the purposes proposed by the potential consumers but also whether data protection obligations associated with the data are an impediment to its legal use by those consumers.

Most importantly, ongoing dialogue with data subjects (or their representatives, in the form of focus groups or lobbying organisation) can potentially identify contextual risks to data subjects – perhaps arising from social or cultural concerns – of which data controllers and/or publishers may be unaware (O'Hara 2011). Furthermore, dialogue offers a means through which publishers can educate data subjects about the risks and associated mitigation strategies employed, and reassure them prior to publication. Data subject dialogue therefore represents a means through which concerns that would be damaging to the publisher's reputation can be variously identified, avoided and allayed. Correspondingly, there is significant scope for data subject dialogue, in the form of consultation or even more involved co-design techniques, to be codified alongside existing formal mechanisms such as privacy impact reviews.

> *"I don't know [what the repercussions are]. That's the problem! Every year someone asks us to release timetable data, but in the past there have been incidences of stalking, so we haven't done this. I didn't even know that, when we started, so it concerns me that there are these 'unknown unknowns'."*

However, dialogue can go beyond initially scoping the publication of a dataset. Given that monitoring the published data for potential misuse or emergent threats such as deanonymisation can itself involve multiple stakeholders, maintaining mechanisms for dialogue throughout the life of the dataset may be important in terms of mitigating harms to data subjects, and avoiding reputational harm to data publishers and consumers.

A forum for dialogue, therefore, needs to contain not only data publishers and data consumers, but also subjects' representatives, domain experts who understand the value of data, and technical experts who can assess the vulnerabilities of the data (O'Hara 2011). Key stakeholders might also include partner organisations, the media, funders and special interest groups, the government and

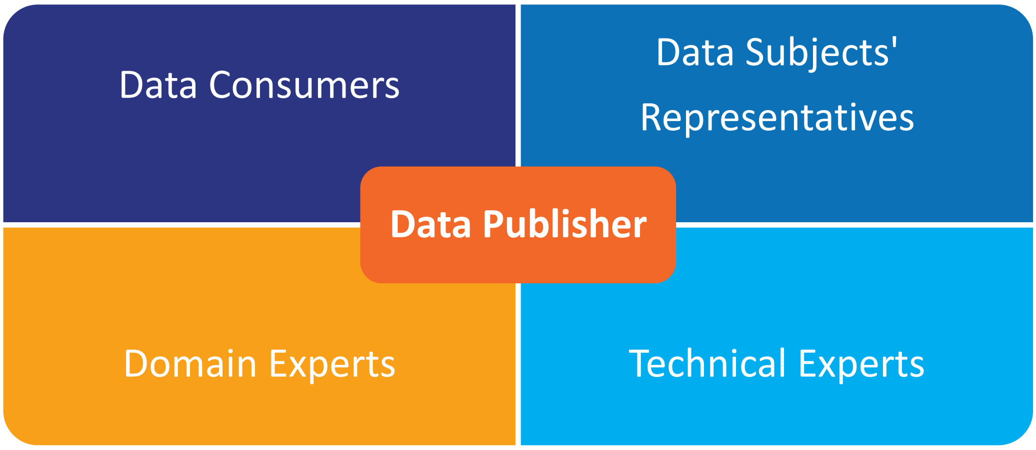

Figure 3: A five-party forum for dialogue

Privacy
Data Protection
Public Confidence

Figure 4: Three pillars to the safe release of Open Data

possibly even the general public. All the organisations interviewed for this report had consultation programmes in place.

What these stakeholders want to know will naturally differ. Some will want to know about privacy or confidentiality, some may wish to be reassured about how data have been anonymised, others may be concerned that the data have utility for them. Understanding their points of view will help, by establishing the legitimacy of the publishing programme (or adjusting it in the face of constructive criticism), and also by understanding stakeholders' information needs when communication is needed.

Because stakeholders are diverse, diverse communication methods will be required.

- Press releases can reach wide audiences with small outlay.
- Social media allow wide-ranging conversations with key stakeholders.
- An actively-maintained website supports accessible messages consistently over time.
- Web surveys, perhaps tailored to different stakeholder groups, can allow more targeted, detailed and confidential interaction.
- Communication events, such as focus groups or public meetings, can help put a human face on particular messages.
- Research is also helpful. What Freedom of Information requests have been received in this space?

Key messages to get over might include case studies of successful reuse of data, or testimonials from users. Privacy Impact Assessments are important, and can be made public.

4.2 Data Subject Consent

In principle, data subject consent (in the abstract) should be the goal of publisher-subject discourse. Consent, which involves the data subject in decisions about how their personal data will be used, provides a level of choice and control, and hence empowerment, to individuals, while respecting their autonomy and their assessments of their own interests. More practically, and as mentioned previously, the consent of data subjects is one grounds upon which data may be processed by a data controller, in compliance with EU data protection laws. In practice, there are several reasons that publishers may have to rely on subject consent in order to publish open datasets that are not fully anonymised.

1. It may be difficult to argue legitimate interest in publishing data without specific use cases in mind.
2. Publishing the data renders it available for transfer outside of the EU. Since the Open Data model (which does not typically require registration) provides little opportunity to enforce the 'model clauses' or to ensure that the data consumer is bound by other regulations such as the EU-US Privacy Shield, explicit subject consent to the transfer may be the only legal justification for such a transfer.

Even where a publisher is legally justified in publishing personal Open Data, there is no automatic justification for a data consumer to process it. Data consumers must have their own legitimate interest in the data, or obtain consent from the data subjects. At present, the latter is likely to be unworkable, but the fact that the data is freely available in the public domain can be taken into account, to some extent, when balancing any legitimate interest of the consumer against the rights of the data subjects, potentially lowering the barriers to processing.

Recital 42 of the GDPR requires that "*for consent to be informed the data subject should be aware at least of the identity of the controller and the purposes of the processing for which the personal data are intended*" and thus appears to rule out the possibility of general consent, for all purposes and all controllers, being sought by the publisher. However, recital 33 does provide a partial exemption for scientific research, where "data subjects should be allowed to give their consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research." In the case where such consent is obtained by the publisher (or another controller involved in the chain of custody) then this should be made clear, alongside the published dataset, in enough detail for eligible data consumers to evaluate whether it is sufficient for their needs.

In the future, efforts towards 'consent management' systems, which allow consent to be sought from data subjects dynamically throughout the lifetime of a dataset, may provide a workable means for obtaining consent to processing by new data consumers, or for new purposes. Such systems have received academic attention (eg Kaye 2014), and some commercial systems are available[4].

5 Conclusion

To conclude, there are three pillars to the safe release of Open Data:

- Privacy.
- Data protection.
- Public confidence.

The only legal requirement is the observance of data protection rules. This can be an unsatisfactory box-ticking exercise, but is essential to avoid fines that will become larger when the new Regulation is in place. However, the sustainability of an Open Data programme will depend on retaining public confidence in the publication process, which in turn means ensuring (a) that privacy is preserved, as far as is feasible and publicly desirable, and (b) that privacy is seen to be preserved. The public needs to remain confident that its privacy and dignity is taken seriously by any organisation opening data up to unlimited (and possibly unaccountable) outside use.

6 Recommendations

Given the issues raised in this paper and the interviews with practitioners, there are a number of recommendations that the authors would make for a data controller or an organisation that was considering making datasets that have been derived from personal data open. These apply even more if personal data is directly to be opened.

1. Understand the data. Consider potential use cases, and the value of the data. Even though this is Open Data (and so the full range of use cases cannot be anticipated), it is likely that the data, if it is to be useful, will be the subject of immediate demand. Identify stakeholders. Consider the effects on stakeholders of making it public. How sensitive is it?

[4] See, for example, "Consentua"; http://www.consentua.com/

2. Consult. Engage stakeholders about the publication programme. Key stakeholders include data subjects, potential consumers, domain experts, internal data controllers, and technical advisors. If consent for data releases can be obtained, then try to obtain it.
3. Remember the three pillars of privacy, data protection and public confidence. By law, a data controller has to obey data protection legislation. But morally, the controller should consider subjects' privacy, and retaining public confidence is essential for the sustainability of an Open Data programme.
4. Be very sure of the grounds for publishing personal data. A decision to publish personal data should be made only on the basis of a strong case in law.
5. Anonymise well and thoroughly. Follow guidelines for anonymising personal data. Consult the ICO (Information Commissioner's Office) code of practice, within the UK for example, and follow the UKAN guidelines.
6. Remember utility. There is no point publishing data which has been denuded of serious content. Ideally, anonymisation will go far enough to reduce the risk of identification while retaining sufficient utility to meet the needs of data consumers. Complete risk aversion means there is no point publishing Open Data at all.
7. Don't release and forget. Anonymisation and Open Data are not cheap options. They demand responsible data stewardship, including monitoring the context in which the data is likely to be used. Preparation needs to be thorough. Consider testing the anonymised data.
8. Have a plan in place in the event of a problem. Be transparent *about* your transparency (as discussed in Section 2.3). Can you contact key stakeholders? Is it possible to contact data subjects? Are you able to withdraw the data, and contact consumers?

Acknowledgements

We are grateful for the involvement of interview participants from a number of organisations, whose input was invaluable in the compilation of this report. They included representatives from:

Project Provenance Ltd. – www.provenance.org

The Department for Environment, Food and Rural Affairs (UK)

Open Laws – www.openlaws.eu

The Cabinet Office (UK)

The University of Southampton Open Data Service

The Department for Energy and Climate Change (UK)

Capgemini Consulting, for their review efforts

References

Article 29 Data Protection Working Party (2014). *Opinion 05/2014 on Anonymization Techniques*, http://www.cnpd.public.lu/fr/publications/groupe-art29/wp216_en.pdf.

European Data Portal (2016). Open Data Goldbook for Data Managers and Data Holders, http://www.europeandataportal.eu/sites/default/files/goldbook.pdf

Ian Ayres (2007). *Super Crunchers: How Anything Can Be Predicted*, London: John Murray.

Paul Bradshaw (2014). 'The transparency opportunity: holding power to account – or making power accountable?' in Nigel Bowles, James T. Hamilton & David A.L. Levy (eds.), *Transparency in Politics and the Media: Accountability and Open Government*, London: I.B. Tauris, 141-165.

Daniel Cameron, Sarah Pope & Michael Clemence (2014). *Dialogue on Data: Exploring the Public's Views on Using Administrative Data for Research Purposes*, IPSOS MORI, http://www.ipsos-mori.com/researchpublications/publications/1652/Dialogue-on-Data.aspx.

Sarah Castell, Anne Charlton, Michael Clemence, Nick Pettigrew, Sarah Pope, Anna Quigley, Jayesh Navin Shah & Tim Silman (2014). *Public Attitudes to Science 2014*, IPSOS MORI, http://www.ipsos-mori.com/Assets/Docs/Polls/pas-2014-main-report.pdf.Deloitte (2014). *Making Digital Default: Understanding Public Attitudes*, Deloitte, http://www.deloitte.com/view/en_GB/uk/industries/government-public-sector/making-digital-default/index.htm.

George T. Duncan, Mark Elliot & Juan-José Salazar-González (2011). *Statistical Confidentiality: Principles and Practice*, New York: Springer.

Cynthia Dwork (2006). 'Differential privacy', in *Proceedings of 3rd International Colloquium on Automata, Languages and Programming (ICALP)*, Berlin: Springer, 1-12.

Amitai Etzioni (1999). *The Limits of Privacy*, New York: Basic Books.

Archon Fung, Mary Graham & David Weil (2007). *Full Disclosure: The Perils and Promise of Transparency*, New York: Cambridge University Press.

Joel Gurin & Beth Simone Noveck (2014). 'Corporations and transparency: improving consumer markets and increasing public accountability', in Nigel Bowles, James T. Hamilton & David A.L. Levy (eds.), *Transparency in Politics and the Media: Accountability and Open Government*, London: I.B. Tauris, 179-196.

Mireille Hildebrandt (2015). *Smart Technologies and the End(s) of Law*, Cheltenham: Edward Elgar.

Information Commissioner's Office (2012). *Anonymisation: Managing Data Protection Risk Code of Practice*, Wilmslow: Information Commissioner's Office, https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf.

Jeff Jarvis (2011). *Public Parts: How Sharing in the Digital Age Improves the Way We Work and Live*, New York: Simon & Schuster.

Kaye, J., Whitley, E. A., Lund, D., Morrison, M., Teare, H., and Melham, K. (2014). Dynamic Consent - A Patient Interface for 21st Century Research Networks, European Journal of Human Genetics 23(2), 141-146.

Helen Margetts (2014). 'Data data everywhere: Open Data versus big data in the quest for transparency', in Nigel Bowles, James T. Hamilton & David A.L. Levy (eds.), *Transparency in Politics and the Media: Accountability and Open Government*, London: I.B. Tauris, 167-178.

Stephen T. Margulis (2011). 'Three theories of privacy', in Sabine Trepte & Leonard Reinecke (eds.), *Privacy Online: Perspectives on Privacy and Self-Disclosure in the Social Web*, Berlin: Springer-Verlag, 9-17.

Viktor Mayer-Schönberger & Kenneth Cukier (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*, London: John Murray.

Helen Nissenbaum (2010). *Privacy in Context: Technology, Policy and the Integrity of Social Life*, Palo Alto: Stanford University Press.

Kieron O'Hara (2011). *Transparent Government Not Transparent Citizens: A Report on Privacy and Transparency for the Cabinet Office*, London: Cabinet Office, http://eprints.soton.ac.uk/272769/.

Kieron O'Hara (2012). 'Transparency, Open Data and trust in government: shaping the infosphere', *Proceedings of ACM Web Science 2012*, http://eprints.soton.ac.uk/337558/.

Kieron O'Hara (2016). 'The seven veils of privacy', *IEEE Internet Computing*, 20(2), 86-91.

Chris Reed (2007). 'Database protection', in Chris Reed & John Angel (eds.), *Computer Law: The Law and Regulation of Information Technology*, 6th edition, Oxford: Oxford University Press, 397-427.

Ira R. Rubinstein & Woodrow Hartzog (2016). 'Anonymization and risk', *Washington Law Review*, 91(2).

Daniel J. Solove (2014). *Privacy and Data Security Violations: What's the Harm?* www.linkedin.com/pulse/20140625045136-2259773-privacy-and-data-security-violations-what-s-the-harm.

Max Van Kleek, Dave Murray-Rust, Amy Guy, Kieron O'Hara & Nigel Shadbolt (2016). 'Computationally mediated pro-social deception', *Proceedings of CHI 2016*.