Open Data and Privacy Workshop reader

Table of Contents

1 Introduction

These brief notes aim to help frame the discussions at the Expert Workshop on Open Data and Privacy, London June 2014. Please note that these notes are not an agenda nor a comprehensive map of the issues around open data. They are designed to give an outline of some of the issues that will be discussed. Participants may add other issues as well. Participants will have different views and positions on many of the topics listed below.

There is a growing buzz about the transformative powers of open data and big data. Data is being called the new oil, without a hint of irony. The UK government is exploring a huge increase in the sharing of personal data across departments, hoping that it will revolutionise the administration.

These disparate initiatives and ideas have now been lumped together, and when they involve personal data the result is a mess. The recent scandal around the British project to disclose personal health information for research - care.data - is a very good example. We need to be a lot clearer about the terms we use for data - open, big, sharing - but also about, what is actually happening to the data.

Open data is fairly well defined, as "*data that can be freely used, reused and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike."*[1] But there is less clarity on terms like big data.

A starting point for most of those attending the meeting is that personal data can never be open data. But some people are precisely focusing on how to open their own data. Both privacy activists and businesses argue that respecting privacy is paramount, but when pressed on the detail, differences start to emerge. We expect robust exchanges, but we also think that there is a lot of convergence, and hope to build enough consensus on certain key areas.

This reader is not referenced to academic standards, but after the workshop we will be adding links and extra materials to an updated online version.

[1] http://opendatahandbook.org/en/what-is-open-data/

2 Regulatory frameworks

Some discussions about open data and personal information with innovators and technologists seem to assume that this is a completely blue-sky zone, where current regulations aren't relevant.

But this is not the case. In a way the problem with open data and privacy is the opposite: there are so many regulations that it is hard to get the full picture. And of course, like in most other areas of cutting edge technological development, reality on the ground moves faster than regulation.

These are not all the possible legal frameworks that influence this area. For example, intellectual property laws also regulate open data.

In addition there are huge differences in jurisdictions. This is mainly between the US and EU, but also with other countries.

2.1 Licensing

Licensing is central to the concept of open data. Projects like OpenStreetMap[2] are based on open licenses that allow free, non-discriminatory reuse. There are several open licensing options, but the important thing is to ensure interoperability.

Many governments, such as the UK,[3] are now open licensing public data to promote reuse and give legal certainty. But the UK Open Government License explicitly exempts personal information from reuse.

It is very unclear how the licensing of personal data actually works. The first step to licensing data is to establish that you have the right to do it[4]. But with personal data this gets complicated, even if we think we have obtained consent, for example with contributors' agreements.

[2] http://wiki.openstreetmap.org/Open_Database_License
[3] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/

There are licenses that include clauses that control some of the reuses. For example, as reported by Ton Zijlstra the Flemish government has a suite of licenses that state "The licensee commits to avoiding any unlawful re-use of the product"[5]. But it is unclear how this can be enforceable.

Another issue is how to revoke consent in licensed data.

2.2 Freedom of Information (FOI)

Freedom of Information laws place some obligations on governments. These laws operate at the national level so there is little consistency across the EU. In some cases FOI laws will have clear regulations on how to deal with privacy, maybe with explicit exemptions, but in other cases they may just refer to other laws.

It is important to note that having a right to access the information is not necessarily the same as being able to reuse it, let alone as open data. The UK is building a right to reuse data in its FOI law, but this is not the case elsewhere. In some cases there could be alleged copyright restrictions or simply direct restrictions.

In all cases FOI has to balance the rights to privacy and information. There is an excellent summary by David Banisar of how this is achieved in various jurisdictions[6].

2.3 Europe

The European Union has a highly regulated regime for personal information. Besides the legislation mentioned below, there are separate regulations for national security.

The European Convention of Human Rights protects the "Right to respect for private and family life".[7] This is broader and higher up in legal terms than data protection (DP), but DP is not a just subset of privacy. Some concepts around the

[4] http://opendefinition.org/guide/data/
[5] http://www.opendataforum.info/files/Modellicenties_ENG.pdf
[6] http://wbi.worldbank.org/wbi/Data/wbi/wbicms/files/drupal-acquia/wbi/Right%20to%20Information%20and%20Privacy.pdf

control of personal data, such as informational self-determination[8], go beyond restricting access to data for privacy considerations.

2.3.1 EU Data Protection

The European Data Protection Directive places very clear limitations on the use of personal information. Some of the critical conflicts with open data are:

1. Purpose limitation

The law says that that personal information should be used for specific purposes. But open data advocates claim that it is precisely the serendipitous use of data that allows for interesting things to happen.

2. Consent

Consent should be specific and informed, so in principle people cannot consent to open ended uses of their data.

3. Export limitations

In theory, personal information should only be exported under conditions that guarantee that the data is protected. Placing open data online makes it very hard to control who can access it.

It is hard to see how a strict interpretation of open data could ever apply to personal data without breaching DP legislation.

The current Directive is about to be replaced by a Regulation. The new law should update existing legislation, although the final details are unclear. Importantly it will harmonise data protection across the EU. Right now each country implements the directive in slightly different way. Regulations are law in the EU without the need for local implementation.

[7] http://www.echr.coe.int/Documents/Convention_ENG.pdf
[8] https://www.datenschutz.de/privo/recht/grundlagen/

2.3.2 EU PSI directive

The European Commission has a large program to promote open data in Europe.[9] Central to this is the recent updating of the European Directive on the Reuse of Public Sector Information (PSI Directive).[10]

The PSI Directive promotes economic and social benefits by the reuse of public information, and the creation of a European market. This means that data should be interoperable across the EU, and also that there should not be any monopolies, whether state or privately owned.

The PSI directive sits atop of existing national laws on access to information. The directive simply mandates that if information is accessible it must be reusable by third parties. It also regulates charges for data, exclusive licensing, cross subsidies, formats, etc.

The current directive makes clear that data protection has to be respected when reusing public data. The Article 29 working party has published an important opinion on the privacy considerations around PSI reuse[11].

2.3.3 PECR

The Privacy of Electronic Communications Directive[12], aka the *Cookie Directive* regulates the use of subscriber data by communications providers for marketing or other purposes.

There is growing interest in the reuse of traffic and location data for other unrelated purposes, as part of the drive towards Big Data. But these regulations place some constraints on what companies can do with the data, which in general should be anonymised unless companies can obtain consent for other uses.

[9] http://ec.europa.eu/digital-agenda/en/open-data-0
[10] http://ec.europa.eu/digital-agenda/en/european-legislation-reuse-public-sector-information
[11] http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp207_en.pdf

2.4 US privacy laws

The US has Privacy Act (1974), but it only applies if nothing newer has overridden it, so it has lots of holes. There are specific privacy sectorial laws for telecoms, health, etc.

2.5 Market self-regulation

There is a growing advocacy, mainly in the US by people like Jaron Lanier,[13] to let markets regulate privacy. The basic premise is that the problem is not so much the pervasive data economy, but who controls it. If consumers were paid a fair amount and everyone was able to share in the profits, things should be fine.

Some critics of these positions, such as Evgeny Morozov, point out that privacy is a social good not just an individual choice. Other criticisms focus on the imbalance of power between individuals and big companies in an already imperfect market.

Europe could give a bigger role to market self-regulation. For example, instead of fines administered by public bodies, privacy violations could be punished through compensation and liability.

2.6 Anti-discrimination and uses of data

There is an opinion that the flow of data is impossible to stop. As more and more personal data finds its way to the net, it is not possible to stop people reusing it. In this view, the best we can do is strengthen anti-discrimination laws to ensure that the data is not misused.

This is happening to a point and most people will agree that discrimination should be fought. What may be problematic is to completely shift the focus away from data protection. There are many complications with applying anti-discrimination law to personal data, for example it is very hard to attribute discrimination to specific information.

---

12 https://wiki.openrightsgroup.org/wiki/Privacy_and_Electronic_Communications_Directive

2.7 The developing world

The regulation of privacy in the developing world is very patchy. But the penetration of digital technologies is huge, particularly mobile phones, so the potential threats to privacy are very high.

Government open data programs in places like Kenya[14] have little consideration of privacy, although in most cases the data is not granular enough to be an issue.

More worryingly is the situation in the private sector. For example we saw recent research to improve transport planning in Ivory Coast based on mobile phone analytics.[15] While this research is undoubtedly useful, it was unconstrained by privacy measures, even though location traces like these are almost impossible to anonymise[16].

3 Risks and benefits

The benefits of open data have been discussed ad nauseam in recent years: transparency, innovation, participation, accountability, economic growth, etc. However in economic terms it is becoming clear that the types of data with the high value tend to involve either massive investment - national maps and weather - or personal information. In addition, risks and benefits are not evenly distributed.

Many advocates of opening personal information normally follow a communalist approach, where individuals should contribute to the common good through their information. For example, there have been arguments that users of public services should contribute some form of "data tax" as quid pro quo. There is also a lot of

[13] http://www.jaronlanier.com/
[14] https://opendata.go.ke/
[15] http://researcher.watson.ibm.com/researcher/view_group_subpage.php?id=4746
[16] http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html

concern about social research being hampered by resistance to share data, a "tragedy of the data commons."[17]
The privacy risks of open data are linked to other issues in new digital environments such as cloud computing and big data analytics - but amplified by the lack of controls and potential reach of the internet.

Developments in data mining techniques, associated with so-called Big Data, introduce new risks around the automated profiling of individuals based on statistical correlations across varied and possibly indirect attributes.

3.1 Anonymisation and re-identification

This is a critical question in this debate. Much of open data involving personal information is processed with the aim to make it impossible for individuals to be identified. There are many questions about how effective these techniques are in an open context.

If data is published openly without constraints it can be mixed with other datasets and lead to the re-identification of individuals. Clearly some identifiers such as name and address are problematic. And certain types of data such as location are intrinsically hard to anonymise. But the main issue with open data is that we cannot be sure of which piece of information will lead to re-identification[18].

This has led to calls for the acknowledgement that de-identification of open data is impossible, and to build policies around this principle.

Critics of this argument[19] point that the risks rest mainly in the implementation of such techniques, which are fundamentally sound.

The UK ICO has published a Code of Practice on Anonymisation[20], and the Article 29 WP has published an opinion on anonymisation techniques[21].

[17] http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1789749
[18] http://whimsley.typepad.com/whimsley/2011/09/data-anonymization-and-re-identification-some-basics-of-data-privacy.html
[19] http://www.ipc.on.ca/images/Resources/anonymization.pdf
[20] http://ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation

3.2 Individual privacy and collective profiling

In many cases individuals can be affected not through identification, but through inferences from open data being incorporated in collective profiles.

For example, it is well understood that mapping individual data points with personal information entails privacy risks.[22] But mapping aggregated data to the street level, say crime, could have an effect on all the people living in that street.[23] In the recent scandal in the UK about the reuse of hospital admissions data by the insurance industry, most adult males saw a rise in their insurance premiums. But this was not achieved by matching their individual medical histories, simply though using the data to adjust risk profiles for the categories to which these individuals belonged.[24]

Many will argue that there is nothing wrong in using as much data as possible to understand risks. But without a level of uncertainty and risk pooling any system for social security or private insurance is not viable.

3.3 Discriminatory services without identification

A different case involves individuals getting a different offer for services based on information about them that does not include personal identifiers. For example, in some websites Apple users were given higher prices based on information about their web browser. As individuals put more personal data online, there is an increased risk of discrimination even if personal identifiers are not available.

3.4 Social and cultural risks

In addition to the typical privacy risks associated with disclosure, we also need to consider social and cultural risks.

---

[21] http://www.cnpd.public.lu/fr/publications/groupe-art29/wp216_en.pdf
[22] http://irevolution.net/2013/01/23/perils-of-crisis-mapping/.
[23] http://www.dailymail.co.uk/news/article-1221604/Online-crime-maps-wipe-thousands-house-prices-overnight.html
[24] http://www.theguardian.com/commentisfree/2014/feb/28/care-data-is-in-chaos

Releasing data that identifies behavioural patterns of certain groups is always risky. Ethnicity and crime are classic examples. With open data these risks are amplified due to the potential for reinterpretations and new combinations of data, which may not have been expected by the original producers of the data. They can be misused by benefiting certain groups. An oft-repeated example is the use of land registry data in India by richer landowners to consolidate their position at the expense of the poor, thus "empowering the empowered".[25]

Cultural sensitivities will also make some data uses more acceptable, with the typical example being the Scandinavian transparency about individual tax records.

3.5 Sensitive data

Sensitive data has a legal status in the EU, but it is not universally defined. Each country has its own list of sensitive data types. In the UK this includes health, ethnicity, sexual life and trade union membership[26]. For example, location data is not sensitive in the UK, but it is in other countries.

The US has a different approach, with laws specific for sectors that are deemed sensitive, be it health or video rental.

Location data is sensitive in itself, but also because when combined with other datasets it can massively enrich the inferences available. Even when aggregated, geo-data still has the potential for harm. The disclosure of crime mapping in the UK, for example, has its own guidelines[27] acknowledging these specific threats.

Other areas worth discussing are health and taxes. There has been public uproar at the release and disclosure of these types of data in the UK.

[25] http://gurstein.wordpress.com/2010/09/02/open-data-empowering-the-empowered-or-effective-data-use-for-everyone/
[26] http://ico.org.uk/for_organisations/data_protection/the_guide/key_definitions
[27] http://ico.org.uk/for_organisations/sector_guides/~/media/documents/library/Data_Protection/Detailed_specialist_guides/crime_mapping.pdf

4 Organisations opening data

Organisations considering open data that contains personal information will have to be extremely careful and justify very clearly why and how they will do so.

The first issue to consider is whether the organisation has the right to release such data at all.

Privacy considerations will have to take place in the wider ethical dimensions of data handling, including social and cultural considerations that may not involve privacy violations.

There will be many reasons why an organisation is looking to open up data with personal information.

Some private sector companies may want to attract external expertise. For example, Netflix released an "anonymised" dataset of movie ratings as part of competition to develop algorithms to help predict viewers' preferences. Unfortunately the data was re-identified by researchers[28].

NGOs will have their own drivers for releasing data to show impacts to funders and other stakeholders. This could involve personal information in some cases.

4.1 Legal requirements

Some organisations, mainly in the public sector, will release data that contains personal information due to legal requirements.

One important aspect to consider is that in many cases privacy is used to avoid the disclosure of data that will be very useful for accountability and other purposes. Privacy is not an absolute unqualified right, and in some cases it will be appropriate to release personal information. But disclosure is not the same as opening.

4.1.1 Public registers and records

Company directors, registered professionals such as doctors, the civil register of life events - births, deaths, marriages - etc. all involved the disclosure of personal information.

Other registers may include receivers of public subsidies, or even the tax details of the whole population, as in some Scandinavian countries.

Digitalisation brings new privacy considerations to these registers. Practical difficulties acted as barriers to abuse, but it is now possible to rebuild a whole register and combine this information with other sources. In some cases this can lead to uses far removed from the original purpose of the register. Making registers open data would remove any possible restrictions.

There are difficult questions around the balance of privacy and accountability in such registers, which are public in the first place to permit wider scrutiny.

4.1.2 Accountability of public employees

There are many transparency projects around the world looking at the expenses or performance of public employees. These projects will have to balance the privacy of these individuals and assess the benefit. For example, publishing surgery survival rates of doctors is seen by many as a valid reason.

But as this information is further processed the situation gets more complicated. For example, some doctors object to public ranking websites created directly by users of services or patients.

In any case there will be some limits to the level of accountability to respect privacy. For example, geo-tracking postal vehicles may be ok, while tracking individual postal workers may be seen as excessive.

[28] http://www.cs.utexas.edu/~shmat/netflix-faq.html

Public sector employees can expect some scrutiny, but they still have privacy rights. Besides, as private companies strive to be accountable to stakeholders, should all workers who generate a public impact be subjected to a similar regime? Conversely, as private companies deliver public services, in some cases they have less demands for openness.

4.1.3 Freedom of Information

Complying with Freedom of Information requests will involve in many cases dealing with personal information. Each case will have specific characteristics and will depend on the jurisdiction, but generally there will be some balancing exercise.

Redaction of personal information is very common, but in some cases details will have to be published.

Reusing this information should be possible in the interests of transparency, for example publication in a newspaper. But unrestricted open data reuse may not be appropriate.

4.2 Specific consent and openness

Privacy advocates normally prefer consent as the mechanism for organisations to disclose or process data. However, consent has to be specific and informed to be valid, and in the context of open data this is not straightforward.

This issue has been explored extensively in bioscience research[29], where alternatives models for consent are being developed.

There are specific issues with crowdsourced data and how to obtain permission and license it. In addition crowdsourced data introduces other questions around the biases of contributors. For example, online tools for reporting the need for street improvements can see more reports form wealthier areas.

[29] http://www.nature.com/ejhg/journal/v21/n9/full/ejhg2012282a.html

4.3 Anonymisation

In many cases organisations considering the release of open data with personal identifiers will "anonymise" this data. In principle, this would make the data non-personal and remove any legal restrictions around privacy and data protections.

However, as we saw above there are risks associated with anonymisation and organisations will need to be careful. Privacy advocates recommend transparency and peer review of de-identification techniques to allow for the spotting of vulnerabilities in re-identification.

These issues are not new. Many organisations have been dealing with statistical disclosure control[30] but most of these disclosures had an element of control. In open data recalling data if a vulnerability is discovered can be hard or impossible.

Many techniques to protect individuals' privacy are based on access or query control. This involves limiting the results that a specific user can receive. But with open data this cannot work.

Most privacy preserving techniques will involve a trade off with the usefulness of the information as they remove some detail, which can range from small modifications to individual records to full statistical aggregation.

4.4 Governance frameworks

Organisations need to consider carefully their information governance when making decisions about personal data. But with open data this is even more important as some decisions may be irreversible.

There are many models for decision-making and accountability. A senior person can be responsible, as in the Caldicott Guardians[31] in the UK health system. An ethics committee or an advisory board may approve decisions to disclose data.

[30] http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/statistical-disclosure-control/index.html
[31] http://www.connectingforhealth.nhs.uk/systemsandservices/infogov/caldicott

Currently there is a preference for broader participation in risk management in sectors of cutting edge technical development where there is little past experience. In some areas, such as nanotechnology and synthetic biology, this has been developed in so called Responsible Innovation Frameworks[32].

Code of conduct can also be helpful, as are external validation systems. The Open Data Institute has a certificate that includes compliance with privacy considerations[33].

5 Organisations reusing personal data

Organisations looking at reusing information available online, including open data have to consider the privacy implications, even if the data is already public available.

Information in public websites is subjected to copyright and any reuse may be limited. Open licensing aims to enable reuse by providing clarity in this aspect.

If personal data is involved, it will be important to consider the origin of the data. Self-disclosure by the individuals themselves normally means that it is legitimate to reuse it, while if it's data published by an organisation this is more complicated.

But there is a difference between being allowed to reuse data and the need to do it fairly. This is a critical distinction in data protection.

5.1 Google ECJ case

The Google case in the ECJ showed that re-users of publicly available data could be made responsible under data protection. Google was found to create a form of profile when presenting results of searches under the name of individuals. This meant that Google was now a new "data controller" with new responsibilities different from those of the original publishers of the information.

[32] http://www.ambafrance-uk.org/IMG/pdf/Richard_OWEN.pdf
[33] https://certificates.theodi.org/overview

5.2 US aggregators

Organisations in the US have a very different regime. There are many sources of publicly available personal information, such as the criminal justice system and education.

There are growing concerns around the implications of the digitisation of these public records. Repurposing these records as open data brings new privacy challenges.

6 Sharing my data

Growing numbers of individuals wish to share their own personal information. Patients release their records hoping to find answers. The quantified self movement aims to track everyday activities as part of a philosophy of self-improvement. Consumers share their bills to find a better deal.

6.1

Volunteering data for social good

Some people may wish to volunteer data for the social good, including health and social research. Most polling organisations have long relied on the goodwill of individuals giving away their views to researchers.

But in an open data context this information could be used for purposes far removed from the original.

6.2 Monetisation of personal data

If data is the new oil, some believe that it is fair that the profits are shared with the originators of the value. There are many complex issues around the value of

personal information. The OECD[34] and many organisations are trying to find some agreed metrics.

But there is no universal agreement that the issue is simply fair sharing of the profits. Authors such as Morozov[35] argue that monetising our intimate life details - in the form of data - is selling our "autonomy", a form of voluntary digital slavery.

For Morozov and others self-disclosure is not just an individual act. It changes the balance of privacy for everyone else around, who will now feel the pressure to share as well.

6.3 Obtaining my data

Anyone wishing to get hold of his or her own data may find that this is not straightforward. There are legal issues, but also many practical obstacles. Some data may be available for download, but rarely in more useful formats, such as APIs.

6.3.1 Subject Access

In Europe, data protection laws allow people to ask for a copy of any information held on them by organisations, for a small fee. But in the increasingly complex chains of data and corporate control this can be complicated. For example, until the recent ruling, Google has refused access to personal data on the grounds that the data is held by its US branch Google Inc., which is based in the USA.

6.3.2 Sensors and trackers

The trend to embed sensors and connectivity in everyday objects, of course including smartphones, means that troves of data are now available. But not in all cases can users access this data directly.

---

34 http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/ICCP/IE/REG(2010)4&doclanguage=en

35 http://www.newrepublic.com/article/117703/selling-personal-data-big-techs-war-meaning-life

6.4 Licensing

Licensing your own data as an individual is not that easy. If a company has generated the data, there will be issues around intellectual property and possibly other restrictions. There is a difference between having privacy rights over data and having the right to license it.

In addition, self-disclosure brings additional complications in Europe by weakening anfuture control that the individual wishes to exert. Any additional benefits form licensing have to be weighted against the potential losses under data protection.

Underlying these discussions are fundamental questions around ownership and control of data. These are not the same.

6.5 Tools to control our data

In addition to legal issues, there are practical and technical problems for individuals wishing to obtain and share their data.
There are now many projects aiming to solve these problems. Personal Data Stores, such as Mydex[36], Vendor Relationship Management tools[37], and the Respect Network of[38] personal cloud tools are some of the main examples.

[36] https://mydex.org/
[37] https://en.wikipedia.org/wiki/Vendor_Relationship_Management
[38] https://www.respectnetwork.com/