Sharing Session: Opening data, protecting privacy

International Open Data Conference, Ottawa, May 2015

These notes form a short report of the session based on our preparation notes and the materials collected during the event.

Author and contact: Javier Ruiz <javier@openrightsgroup.org>

One line session description: [Being open-by-default should not threaten the privacy and rights of citizens. How can we achieve both goals? **]**

Much of the data governments hold is about private citizens, and local communities. Whilst open data principles frequently cite the importance of respecting privacy - the details of how this should be done are rarely explored in depth.

This session will explore open data privacy principles, the risks to be aware of, and strategies to ensure the privacy risks of open data are well managed.

The session was based on very short presentations with extended group discussions in round tables.

The aims were to:

1. introduce some basic concepts around privacy and open data to participants.
2. encourage critical thinking and further discussion after the workshop within the participants' organisations.

Resources to draw upon:

- Recent Berkeley Workshop
- http://www.opengovguide.com/topics/privacy-and-data-protection/
- ORG Primer on Open Data Privacy (quite Europe focused)

Basic concepts we want to convey to participants:

- Privacy is not the same as confidentiality. This means that information that is available online can still have privacy impacts if reused. The EU case on Google's reuse of old personal data, misnamed “the right to be forgotten” is a good example. Every reuse needs consideration.
- Achieving both absolute openness, and complete privacy-protection may be

possible in theory, but really hard. In practice we will need to make some compromises.

- The focus needs to expand from the initial publication of data. What you do with data once it is open matters, reuse is as important as publication.
- We need to move beyond legal compliance, which is still important, and into wider ethical considerations.
- There are laws other than privacy that can restrict the re-use of data. Commercial agreements, intellectual property, etc.

ACTIVITY 1: why consider personal information under open data

Introduce the issue briefly for whole room discussion.

Show www.mugshots.com as example of bad open data

Open data definition says that open data is generally not personal data.

Why may we want to open personal information, then?

- research and public policy (health, social, economics, national statistics, census, etc.)
- transparency, accountability and governability (politicians, registers, public expenses, subsidies, fraud, court records, education, FOIA)
- economic development (mapping, health)
- public participation and public services (budgets, anticorruption, engagement, e-government)

Privacy risks can act against the goals we have:

- negative reaction against transparency, possibly even judicial review
- impact of Snowden leaks

ACTIVITY 2: privacy concerns and open data

Participants were asked to discuss in each table examples of open data releases or programmes and concerns or actual negative privacy impacts.

Participants received printed copies of an edited version of the Sunlight Impacts Report list of case studies. The version we gave out had our added privacy column removed to avoid

giving too many hints.

DISCLAIMER: the “potential privacy issues” listed in the spreadsheet do not amount to accusations or even serious concerns, simply point to questions we would ask the project developers if we had to do a privacy evaluation. They are not systematic and only provided in order to assist workshop participants. This list of issues would require a lot of extra work before publication could be considered.

https://docs.google.com/spreadsheets/d/1FJZhFsFQKd9E9dgXjTC3RnHwspCDhyGMOc2UwSE4QYg/edit#gid=0

The groups in the tables identified the following general privacy concerns around open data. We have grouped here in themes but the responses came in separately:

- re-identification, linking, releasing multiple datasets which link to longitudinal data and create connections, triangulation, anonymisation vs usefulness. Census and national statistics can be risky even if aggregated. Records can be identifiable within a community. Public sector collects very sensitive information.
- power of private sector v citizens, differences in who uses data and purposes, big data for the masses
- profiling, negative economic effects (even if objectively fair), discrimination, vulnerable people. Small rural communities particularly at risk. Uncertainty about the law on discrimination via identification through data. Particular risks of location data
- identity theft, fraud, spam,
- surveillance is made easier (both state and corporate), chilling effects, inhibiting political activity
- enforcement of restrictions, education of data holders
- hat happens when old sensitive data gets trumped by next wave of open data? Will protections remain?

[NOTE: the above concerns raised by participants are consistent with the discussions on open data privacy we've had elsewhere.]

Additional issues that were mentioned but not discussed in detail:

- copyright issues
- national security
- ethics of researchers

Open data of concern (non exhaustive):

- registers
- census data (released only after a long time)
- health (Canadian CCA report and ATIPS company information)
- courts (spelt out as legal decisions aka sentences)

After the reports back we discussed briefly the specific risks and provided our pre-prepared taxonomies:

- Need to distinguish different kinds of data e.g.:
    - Public registers
    - Big data
    - Small datasets (accidental disclosure)

- with an emphasis on thinking about how this plays out in different contexts across the world.
    1. disclosure or re-identification: discrimination, reprisals, reputation
    2. inference, open data linking, individual profiles: as above, also control and fairness, reidentification of other datasets
    3. collective profiling through big data (even if anonymised): discrimination in services or prices
    4. data value: transfer of resources, expropriation, privatisation of collective goods

ACTIVITY 3: A common language for privacy

We gave a short introduction to privacy.

Privacy is “covered” (mainly mentioned in passing) in most current open data guidelines. E.g. G8 Charter states “*14) We recognise that there is national and international legislation, in particular pertaining to intellectual property, personally-identifiable and sensitive information, which must be observed.”*

Privacy in general is culturally specific but this does not mean that people don't have privacy concerns, just that they are expressed differently (Spanish kisses vs Finnish

saunas). But in relation to information systems there is a growing convergence. Now some 101 countries have privacy laws and many constitutional rights, mostly follow the OECD privacy principles.

These principles have become the lingua franca of modern privacy and we will use them here to introduce basic privacy concepts to participants and give them some conceptual tools for discussion.

Participants received printed copies of the OECD privacy principles for discussion in tables.

http://oecdprivacy.org/

We kept the discussion format flexible and not too prescriptive e.g.:

- discuss the principles in general
- go back to the case studies or one of your own experiences and see how these fit or not with the principles.

Issues raised by participants around privacy

- consent problems with long privacy notices, jurisdiction and ToS of social media platforms
- privacy definition is not set and this is confusing, cultural specificity of privacy, including indigenous peoples' ideas of collective ownership of culture,
- context/medium/forum affect what is shared, conversely is it acceptable to share within some scope but not other?
- need to clarify what are general issues around data and those specific to open data
- illusion of control, can controllers really stop abuses?
- National regimes, what data is held outside the control of an organisation operating in a particular legal jurisdictions (local, state, federal)?
- What is anonymous data? And who determines what is private?
- Privacy can be a barrier to releasing time-sensitive data that it's immediately valuable,
- privacy as excuse to avoid publishing accountability data (raised in the context of aid donors and receiving organisations)

- incentive in protecting good social capital against reputation damage, but conversely sensitive business info has similar risks to personal data (NOTE: we didn't discuss this small privacy heresy :-)
- perceptions of authority of some public bodies may be bigger than reality, illusion of control
- what about access to your own data?
- Balancing the risks with the benefits received (NOTE: unclear but it seems to refer to same person, which is a critical aspect of balancing exercise)
- privacy laws are outdated, public servants need to fill the gaps

Specific issues raised about privacy principles:

- #3 poorly written, should include expiration of data.
- #4 authority of the law is too broad, and consent should be more specific
- #5 enforcement?
- #6 public awareness and engagement

After the feedback we explained that despited many issues the principles remain the best rough consensus on privacy and should be used as a basic guidance.

There are of course complications, such as:

- what is personal data, if scientists cannot agree on identifiability?
- who is the data controller, in complex open data reuses?
- what is consent, if you cannot read the terms and data is transferred anyway?

ACTIVITY 4: making open data respect privacy

In our final activity we distributed printed copies of the Open Data Principles of Canada (just to match local context but we could have used many others).

http://open.canada.ca/en/open-data-principles#toc95

The groups were asked to look at each individual principle and see if they were justified in a particular data release. What compromises could be made while keeping the main public interest rationale for publication.

Ideally go back to the data projects discussed earlier and see if there is anything you could do to improve the privacy situation. What trade offs - if any - would you need to make in data utility?

For example, if the concern about the data is machine harvesting for marketing, machine readability may not be appropriate. Completeness may be good, but redactions can help protect privacy while allowing for accountability. Personal data may be published but it must be clear why.

During the discussion we (organisers) raised some general issues:

How to address privacy and harms in an open data strategy?

- Anonymisation and consent are not the simple solutions some policymakers in a rush make us believe:
    1. computers are racist (will build stereotypes and profiles even without identifiers)
    2. consent as one off not compatible with open data
- Legal compliance is not enough, need broader ethical processes looking at public benefit, etc. OECD principles form the basis but are not the end.
- Open data principles must be individually justified against privacy concerns

We briefly mentioned but did not discuss in detail some examples of specific activities

1. Privacy impact assessments;
2. Redactions;
3. Privacy-by-design programmes

The groups did not feedback their discussions due to lack of time.