Applying Contextual Integrity to Open Data Publishing

Jane Henriksen-Bulmer
Bournemouth University
Talbot Campus
Poole
UK
jhenriksenbulmer@bournemouth.ac.uk

Shamal Faily
Bournemouth University
Talbot Campus
Poole
UK
sfaily@bournemouth.ac.uk

1. INTRODUCTION

Data is fast becoming a commodity to be traded, utilised and shared between individuals and businesses alike, so much so that it has been branded 'the new oil' (Palmer, 2006); (Van't Spijker, 2014). Open data is data which can be freely downloaded, used and shared for any purpose by anyone (Open Data Institute, 2016), while open government data is open data that originates from a government controlled entity - a public body (Open Government Data Working Group, 2016). A public body is a body or organisation that is governed by public law which exercises public functions that are woven "into the fabric of public regulation" (Burton, 2002).

Much of the market place for data is predicted to be powered by open data with an estimated global annual economic market value of up to $5 trillion (Manyika et al., 2013), making open data a very valuable commodity. Further, research has found that government open data, plays a critical role in this economic value creation (Chui et al., 2014); (Open Data Institute, 2017). For example, it is estimated that the value creating for open data originating from within the European Union (EU) was worth 55.3 billion EURO in 2016 (Carrara et al., 2015).

Most of this data is about individuals in some shape or form. While a large proportion of data may be user-generated, such as social network data, aimed at providing information and/or entertainment for friends and family, some of this data will have been generated about users by third parties including public bodies. Public data is data that is collected and used by government bodies about users for an official purpose. This would include tax, waste collection, census, education and voters roll to name but a few.

In recent years, governments have been working towards openness and increasing transparency to encourage wider participation with citizens. As part of this drive, many government bodies are publishing their data in open format (i.e. as open data). Some governments facilitate this through centralised open data portals, such as data.gov in the US and data.gov.uk in the UK. Others have adopted a more localised approach, leaving it up to individual public sector bodies or administrative areas to publish and maintain their own data (Attard et al., 2015).

Open Government seeks to promote three core values: transparency, participation and collaboration between government and its citizens (Obama, 2009). At the heart of open government is access to information and open data. Within the UK and the EU there are now various statutory provisions, such as the Re-use of Public Services Information Regulations (ROPSIR) 2005 and 2015, that place an obligation on public bodies to release data in open format.

Privacy permeates all aspects of our lives, our values, beliefs and cultural and societal norms. Much of how we conduct ourselves and negotiate our privacy values comes down to unwritten rules or norms that we abide by as part of our daily lives, often without thinking about it in any depth, it is engrained into our culture and norms. However, when it comes to information and data about us, while these rules may still apply, we, the users ("the data subjects"), may not necessarily be the ones to

apply those rules, rather, the people who collect and handle data about us will be applying the rules on our behalf. Once this data is published in open format however, users need to know that their privacy has been sufficiently safeguarded before data is shared with third parties.

Opening up government, information has been shown to strengthen citizenship, improve participation and encourage innovation (Geiger and von Lucke, 2012). Transparency can also have the opposite effect when, for example, data is mishandled or personal information is published without the full consent of the data subjects, causing privacy to be breached. This privacy threat has been demonstrated by several successful re-identification attacks relying on open data (Henriksen-Bulmer and Jeary, 2016; Ohm, 2010) and therefore, public bodies need to carefully consider the privacy implications of making data open. Consequently, while users may have an expectation that their personal information will be safeguarded, particularly when handled by a public body, this may no longer be the case with data published in open format by public bodies.

In this paper, we illustrate how privacy by design (PbD) can be incorporated into the decision making process associated with open data publishing. We describe a case study exploring the privacy decision making processes of a UK local authority (LA), and applied our interpretation of Nissenbaum's contextual integrity framework (CI) (Nissenbaum, 2010) to the LAs privacy decision making process around open data publishing using real data.

The rest of this paper is structured as follows. Section 2 describes the initial research conducted to collect background data from public bodies about their open data practices. Section 3 explains how this information informed the case study, and was used to devise a questionnaire that aligned with CI. Section 4 presents the findings, which we discuss in Section 5. We describe limitations of our study in Section 6, before presenting directions for future work in Section 7.

2. APPROACH

To gauge how proactive public bodies in the UK have been in meeting their legal obligations in publishing open data, we submitted a Freedom of Information (FOI) request to a sample group of 22 randomly selected LAs in the UK. We asked each LA whether they published information in open format and/or whether they complied with the FOI publication scheme as well as who bears responsibility for any such publication(s). We found that all published something under the FOI publication scheme, but only 37% had some form of open data platform or portal. We also found the role and/or department responsible for open data publishing vary considerably across the LA's contacted. The roles of responsible practitioners range from Information Officers through to Legal or GIS professionals.

We contacted six of these LAs by telephone and/or email to understand their data publishing practices and what barriers to publication they may have encountered. Of those, two LAs agreed to take part in more in-depth contextual interviews (Beyer and Holtzblatt, 1998). We interviewed five practitioners from across the two LAs, two senior managers and three practitioners who worked with open data. We discovered that all interviewed practitioners recognised the usefulness of open government data for public good, e.g. *"by opening up information to people you can foster growth" (P1)*. This sentiment is echoed by Government who want to promote transparency as discussed above. Further, research suggests that innovative uses of open data can be achieved when data is made publicly available (Chui et al., 2014). For example, companies like OpenStreetMap provide users with free access to maps of their local area, powered by public body open data(Open Data Institute, 2015).

With regards to publication, we found that the norm in practice appears to be not to publish until pressure dictates otherwise: *"Why do we have to do anything when we can get away with the bare minimum?" (P3)*. It also appeared that the decision to publish or not comes down to the strategy adopted by management within that public body as this will directly influence the level of resource allocated to such publication schemes: *"there is also the corporate buy in issue, you need to get that....the barriers are that what we have got to do is get buy in and the attitude is, well, what's in it for us?" (P4)*.

When asked about privacy, opinions varied about the extent privacy is preserved under current practices within public bodies, e.g.*"I am not too concerned about the privacy angle because we would never put out personalised data." (P1)*, and *"I am almost convinced that if I went back through our data that we have published over the last 4-5 years, I would find something that we'd missed" (P2)*. This difference of opinion may explain why only 37% of LAs currently have an open data portal. We therefore decided to examine the privacy protection currently available to practitioners.

From a legal perspective, users' privacy is protected by law under the Data Protection Act 1998 (DPA). The DPA controls how organisations manage, store use and share personal information by providing

strict guidelines for what the people who handle the data, the “data controllers” and “data processors”, may or may not do with the information they handle. It also enables individuals to obtain details of what personal information a particular public body or organisation holds about them upon request.

Technically, there are various ways that privacy can be safeguarded. Most of these look at protecting privacy from the perspective of the data itself. For example, data can be anonymised, which involves removing or obfuscating any personally identifiable information prior to publication (Samarati, 2001; Dwork, 2006). Alternatively public bodies can use technology to preserve privacy at system level by, for example, controlling access and/or settings within the applications that hold the data. However, with open data publishing, technical mechanisms are difficult to implement as the information is not confined to one system or organisation and therefore, there is no single environment to protect. Further, restricting access would contravene open government data principles that require the data to be freely available to all (Open Government Working Group, 2007).

Contextual Integrity (CI) looks at the privacy decision making process and whether a particular practice poses a risk to privacy. In terms of data, privacy is described as “a right to appropriate flow of information” (Nissenbaum, 2010). CI asks practitioners to consider information flows from three perspectives:

- The actors, i.e. the people who are the ’data subjects’ or who handle the data, i.e. the ’data controllers’ and ’data processors’.
- The attributes, i.e. the individual elements that make up the data.
- The transmission principles, i.e. how the data is conveyed and shared (”the data flow”).

CI then asks that a further evaluation is conducted to assess the roles in which the actors act in a particular context; how the roles interact with each other; and what defines these behaviours (norms). Finally the contextual teleology is considered, i.e. the values, purposes, goals and ends of the particular situation or setting.

Looking at these concepts, CI appears to offer an effective way of applying PbD to the open data decision making process. To evaluate this, we created a decision making questionnaire, designed to test CI in practice based on Nissenbaum’s CI framework.

3. CASE STUDY

We conducted a case study to establish whether CI can be used as an effective tool for practitioners in deciding whether or not a particular dataset is suitable for open data publication. The study was conducted in collaboration with a LA because, as a public body, a LA has a legal obligations to publish open data while at the same time acting on behalf of citizens, meaning they are likely to face more scrutiny than private organisations might.

Due to distance and time restrictions, the interviews were conducted using shared access to the questionnaire via Google Docs. Follow-up telephone conferencing was then used to discuss and capture the practitioner’s answers to these questions. Because of these restrictions, the data collection was carried out using a combination of contextual interviews (Beyer and Holtzblatt, 1998) and think aloud (Davison et al., 1997) methods. We chose this approach as the contextual interview technique was not considered sufficient because the authors could not meet with practitioners face to face. Therefore, to provide a more robust technique in the given circumstances, the think aloud element was added.

3.1. Creating the Questionnaire

The CI decision making questionnaire was devised using a spreadsheet [1] consisting of 98 questions.

CI’s three key elements – *Explanation*, *Evaluation*, and *Prescription* – provided three overarching phases of the questionnaire to delineate and break down CI into manageable chunks. Within each phase, questions were then created by interpreting each of the nine CI Decision Heuristics (DHs) and devising suitable questions interpreting each of DH. For example, the first group of questions ask for information about the attributes within the dataset and the people who handle the data. Practitioners were also asked to score each answer provided using a simple traffic light marking system: red, amber, green to denote high, medium and low risk (Heiser, 2008). The intention is not for the CI questionnaire to compute or calculate the score for the practitioner as most of the evaluation will be subjective and require expert input to make a decision. Rather, the scores are intended to provide practitioners with an easily referenced focal point for making a decision in the final phase.

Phase 1 - Explanation

The first key element, ”Explanation”, seeks to establish “the governing context” (Nissenbaum,

[1] Available to download at `https://github.com/JaneHB/CIOpenData`

2010) and gather information about what we are assessing. Thus, in this phase DH1 through to DH4 were used as a basis for the questions which ask that consideration is given to: (DH1), the proposed new information flows; (DH2), the prevailing context; (DH3), who will be handling the data (information subjects, senders, and recipients) and (DH4), how the data will be shared ("the transmission principles").

Phase 2 - Risk Assessment

The second element, "Evaluation", refers to recognising the proposed change in data flow (transmission principles), and whether this change will affect privacy in light of established norms, values and goals (Nissenbaum, 2010). Thus, this phase seeks to establish the risks are associated with the proposed change in information flows (transmission principles). In light of this, we renamed this phase 'Risk Assessment'. The rationale for the change in name being that, while practitioners understand what evaluation means, in practice, they will likely conduct an evaluation in terms of assessing the risks a particular practice or system poses.

For this phase, DH5 through to DH8 were used to inform the questions. DH5 relates to evaluating the entrenched information norms, followed by an assessment of any risks associated with the proposed changed transmission principles (DH6, 7 and 8).

Phase 3 - Decision

The third phase covers the third element, "Prescription". This has been translated into 'Decision' as this is where the outcome of the risk assessment will be gathered and the suitability of publishing a particular open dataset is decided (DH9).

4. RESULTS

The case study was conducted in collaboration with a UK LA with a pre-existing open data publication scheme. We worked with two practitioners from the LA, both of whom work with publication of open data, one on a technical level, the other from a policy/process perspective.

In the case study, three datasets that had already been published were assessed by running each dataset through the questionnaire. In applying the CI questionnaire, we found it necessary to explain the reasoning behind each set of questions in greater detail than provided, to elicit fuller, more thought out responses. For example, the practitioners queried the necessity of detailing each attribute and actor separately, resulting in a detailed discussion about whether breaking the dataset down in this manner was required. However, once practitioners understood how each attribute could potentially have privacy implications in light of the informational norms, values and context, the participants really began to think about the data in context, making the rest of the assessment much more insightful for everyone.

This exercise took three hours and resulted in the identification of privacy concerns in all of the datasets assessed. One dataset contained directly identifying information, while the remaining two datasets contained data that, if linked to external data, could render the data personal or sensitive.

We found that Nissenbaum's CI framework provides an effective privacy decision making tool for open data publishing, despite the inability to define one of the roles. The inability to clearly define the end user did not appear to hamper the practitioners in the decision making process. If anything, it appeared to make heighten their caution when considering the privacy implications of releasing the data.

Our findings also show that the CI questionnaire has potential to provide a usable tool in the open data domain. We found that, despite not being able to define who the end user was, the CI questionnaire provides enough structure and guidance to elicit a balanced view of the privacy risks associated with publishing a data set in open format. Further, where risks were identified, the CI questionnaire also encourages practitioners to considered what mitigation strategies, if any, can be applied to make the dataset suitable for publication as open data.

5. DISCUSSION

Our findings show that existing processes fail to adequately address the preservation of privacy in open data publishing decision making, and highlight the need to look beyond the CI questionnaire as part of the decision making process.

To illustrate, the dataset containing personal information had been published because consent had been sought when the data was originally collected. This dataset consisted of agreements between citizens and the planning authority of any peripheral work agreed to be undertaken by the applicant as part of the planned development. It is standard practice that planning applications seek consent from the applicant in order that information relating to the proposed development (and any associated agreements) can be shared and scrutinised by interested parties who may be affected by such development, such as neighbours.

However, these applications will potentially include detailed plans and layout of the property as well as dates for when work will be carried out, leaving the applicant(s) very vulnerable indeed. Thus, while an applicant will appreciate that details of their proposed development will be shared with their neighbour and other interested parties, they may not be aware that this information will also be freely available for anyone to download online.

Upon speaking to the participants taking part in the case study, it transpired that they had little background knowledge around consent and informed consent. Further, although they, as publishers, quality checked datasets prior to publication, they relied heavily on the originating department (i.e. the department who collected and maintained the raw data) to conduct their own quality assurance and ensure that the data submitted was compliant with DPA. Arguably however, in this instance, those internal assurances were insufficient.

This finding shows that privacy considerations have design implications that go beyond the overall decision making process which need addressing. For example, it highlights a need to look at how PbD principles can be incorporated into organisational process and the systems that house the data during the early phases of collecting and collating the data, i.e. the data gathering, storing and handling stages in the originating departments. Addressing this aspect will require a review of existing practices with a view to determining how PbD principles can be better integrated into processes and system design. This will require consideration to be given to how PbD can be facilitated within systems and processes in a manner that is compatible with the understanding that the data may, in future, be published in open format.

Public bodies looking to publish data in open format will have to consider data from multiple departments on a variety of subjects. There is no way that one person, such as the information officer responsible for the publication process, can understand or account for all the legal, policy and contextual nuances of each dataset. Perhaps therefore, there is a need for the privacy risk decision to be split between multiple stakeholders to ensure sufficient expertise is applied to the decision making process, particularly in this sphere. Consequently, one design implication of these findings is a real need for identifying and developing practical methods of preserving privacy not only on a technical level, but also, on more strategic level to provide a more holistic assessment overall. This case study has shown that the CI questionnaire has potential to be an effective way of incorporating PbD into the open data privacy decision making process before publication occurs.

The findings also show that practitioners remain unclear how best to preserve privacy. As a result, they may either publish data that contains personally identifiable information as highlighted in our study, or chose not to publish at all as found in a previous study where one Senior Manager interviewed stated; “the easiest thing is to not make the data available. You’re not going to make any mistakes if you don’t make the data available” (Barry and Bannister, 2014). However, in the current climate of openness and transparency, public bodies will increasingly find it difficult, if not impossible, to not publish any open data if they are to meet public expectations, and indeed, their legal obligations under ROPSIR and similar legislation.

Another implication for design is that because privacy sensitive data is already being published by public bodies resulting in privacy breaches, there is an urgent need for more structured and robust methods of assessing privacy when making decisions relating to publishing open data to be developed. Thus, we, the HCI community, need to look at ways we can incorporate privacy holistically into our designs. For open data, privacy needs to be considered, not just as part of technical implementation or design, but also on a more strategic level before publication occurs.

6. LIMITATIONS

Because this study only considered three datasets from one public body, the next step will be to conduct a wider, more detailed study to further validate these findings. Further, while the CI questionnaire was largely effective, some of the questions required modification and further explanation and some questions were found to be redundant. For example, following the initial discussions around the need for all attributes and actors to be considered mentioned earlier, it transpired that, as part of answering the initial questions on attributes, many of the questions that followed were answered as part of those initial responses. Consequently, these questions require revision in future adaptations of the questionnaire.

Once the case study had been conducted, practitioners were asked whether they felt the fact that the end user could not be defined had prevented them from considering how the data might be perceived in light of informational norms, or in the context of potentially conflicting values or morals. The practitioners felt that, rather than acting as an obstacle, this served as a reminder that extra care and time needs to be taken when considering privacy implications of

publishing open data. This could of course, have the opposite effect and cause less rather than more data to be approved for publication. However, to prove this or otherwise, a larger sample group would have to be tested, something we intend to evaluate in future work.

7. CONCLUSION

Public bodies face a number of barriers in meeting their obligations in open data publishing. These include fear of adverse consequences, litigation, a lack of adequate processes and resources, technical constraints, and an insufficient understanding of how to deal with privacy implications and/or compliance (Barry and Bannister, 2014). Thus, the practitioners at public bodies themselves face a real problem in overcoming these obstacles.

This paper has looked at one of these barriers – privacy – and found that existing process fail to sufficiently preserve the privacy of individuals. Running these datasets through the CI questionnaire identified a gap in existing processes which, if not addressed, could result in public bodies continuing to publish privacy sensitive information without full consultation with the data subjects. Further, this has also shown that there is also a need for PbD to be integrated into organisational processes and system design from the start.

Practitioners expressed concern over the lack of guidance in dealing with privacy in practice. This has proven to be a valid concern as the study highlighted how a lack of guidance, coupled with minimal structured processes currently being in place, resulted in identifying information being made public as part of existing open data already published. Further, while consent was not within scope of this study, our findings also highlight that a wider discussion needs to take place around consent and what information should be made available in open format.

We believe the CI questionnaire provides public bodies with the means of assessing the balance between the privacy of the data subject and the needs of the LA, thereby providing an important first step towards that goal. We believe the questionnaire is generalisable to any situation that requires privacy implications to be considered and could therefore, be adaptable to any requests for information received by a public body such as a FOI request.

Future work will further evaluate whether the inability to define the end user will result in less data being published in open format. It will also consider whether more than one stakeholder needs to be involved in the decision making process. For example, if the proposed framework is broken into sections, practitioners within each specialism, e.g. political, legal, data management etc., can be asked to complete sections, thereby arriving at a better, more informed decision for each dataset prior to publication.

This study has highlighted the potential for open data utilisation. This is an understudied area that the public bodies are keen to promote, to help their case in obtaining buy-in to extend and expand open data projects. However, to do this they need to better understand what the end users expect and want from open data. During the study, a participant asked: *"is your research looking at the actual evidence-based end user stage of use for open data as well? Because that is what we are struggling with" (P5)*. Consequently, additional work by the HCI community and the end users in exploring this space, both from a privacy and design perspective would be welcomed by government bodies.

REFERENCES

Attard, J., Orlandi, F., Scerri, S., and Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly*, 32(4):399 – 418.

Barry, E. and Bannister, F. (2014). Barriers to open data release: A view from the top. *Information Polity: The International Journal of Government & Democracy in the Information Age*, 19(1/2):129 – 152.

Beyer, H. and Holtzblatt, K. (1998). *Contextual design : defining customer-centered systems.* San Francisco, Calif. : Morgan Kaufmann Publishers, c1998.

Burton, J. (2002). Mind the gap. *The New Law Journal*, 152(152 NLJ 1933).

Carrara, W., Chan, W. S., Fisher, S., and van Steenbergen, E. (2015). Creating value through open data: Study on the impact of re-use of public data resources. Technical report, Capgemini Consulting on behalf of the European Commission (EC): European Data Portal.

Chui, M., Farrell, D., and Jackson, K. (2014). How government can promote open data. Technical report, McKinsey & Company.

Davison, G. C., Vogel, R. S., and Coffman, S. G. (1997). Think-aloud approaches to cognitive assessment and the articulated thoughts in simulated situations paradigm. *Journal Of Consulting And Clinical Psychology*, 65(6):950 – 958.

Dwork, C. (2006). *Differential Privacy*, pages 1–12. Springer Berlin Heidelberg, Berlin, Heidelberg.

Geiger, C. P. and von Lucke, J. (2012). Open government and (linked) (open) (government) (data). *eJournal of eDemocracy & Open Government*, 4(2):265.

Heiser, J. (2008). A simple method for expressing information criticality and classification.

Henriksen-Bulmer, J. and Jeary, S. (2016). Re-identification attacks: A systematic literature review. *International Journal of Information Management*, 36:1184–1192.

Manyika, J., Chui, M., Farrell, D., Van Kuiken, S., Groves, P., and Almasi Doshi, E. (2013). Open data: Unlocking innovation and performance with liquid information. Technical report, McKinsey & Company.

Nissenbaum, H. F. (2010). *Privacy in Context: Technology, Policy, and the Integrity of Social Life.* Stanford Law Books, Stanford: California.

Obama, B. (2009). Transparency and open government: Memorandum for the heads of executive departments and agencies.

Ohm, P. (2010). Broken promises of privacy: responding to the surprising failure of anonymization. *UCLA Law Review*, 57(6):1701 – 1777.

Open Data Institute (2015). Open data means business: Uk innovation across sectors and regions.

Open Data Institute (2016). What is open data?

Open Data Institute (2017). The value of open data.

Open Government Data Working Group (2016). Open government data.

Open Government Working Group (2007). Open government data principles. In Re: Open Government Working Group Meeting in Sebastopol, CA, 2007.

Palmer, M. (2006). Data is the new oil.

Samarati, P. (2001). Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027.

Van't Spijker, A. (2014). *The New Oil: Using Innovative Business Models to Turn Data Into Profit.* Technics Publications, Basking Ridge: NJ, e-book edition.