Hi Kelsey,

Here are 3 comments to the Open Data Risk Assessment Draft Report:

1. Page 8 Data Quality: The paragraph starts out by pointing out the multiple stakeholders and circumstances under which inaccurate, incomplete, or biased open data may have little impact and then contrasts this with other circumstances where the publication could adversely affect individuals. That is good.

I recommend deleting the last sentence on the bottom of page 8 "Because open data is used so widely and for so many diverse purposes, it is critical that any data released be accurate and unbiased." Data quality principles play an important role, but I wouldn't go so far as to state that it is "critical that any data released by accurate and unbiased." This is in contradiction to the first part of the paragraph where you have framed the discussion as being one that is contextual in nature, without getting into the weeds of defining what you mean by accuracy. The last sentence invites questions in the weeds and goes outside the scope of the document (e.g., What is accuracy in this context? Who defines it? Are we referring to accuracy of specific attributes only (names of people arrested)? Accuracy of location? Or are we referring to precision?) These concepts form a substantial area of study, debate, and standardization in a number of related domains including the one I am most familiar with (GIS). *See ISO 19157:2013 Quality Principles and Quality Evaluation Procedures.*

2. Page 9 Paragraph 3: When I read this paragraph for the first time I saw the reference to "low-quality data", immediately taking me back to my thoughts on staying out of the weeds wrt definitions. I then went to the City of Seattle Open Data Playbook to read the definitions for the 3 defined accuracy categories. I commend them for coming up with something and detect in the word choice that it was probably a lot of work to get some consensus. At first I took issue with the use of the word "accurate" in the "Good" quality category, "gaps and discrepancies" in the "Acceptable" category, and "not valuable" in the third category. To me this looked like a binary assessment, until I read: "...intended to assist with prioritization. It is not a substitute for the in-depth data quality and privacy assessments that are required prior to publication. In the template itself, you will find definitions for each assessment category as well as examples." I think this point is worth emphasizing in the section on Data Quality, perhaps by saying "as the City of Seattle acknowledges, a quick categorization into these 3 levels of quality is not a substitute for the in-depth data quality and privacy assessments that are required prior to publication..."

3. Page 20 Appendix B: Step 1**:** The category of information "*Non-Identifiable Information*" includes a reference to "GIS data." I disagree with the classification of GIS data as non-identifiable. My recommendation is to delete the reference. GIS is a database file format that by definition organizes information in a geospatial way (somehow tied to the earth or a proxy). Files released as Open Data are often stored in, extracted from (and presented in) GIS file formats. That includes files containing what you classify as "*Direct Identifiers", "Indirect Identifiers",* and "*Sensitive Attributes"* in addition to "*Non-Identifiable Information."*

Let me know if you have any questions related to the above.

Best Regards,

Kara

Kara Selke
VP of Strategic Partners and Privacy
StreetLight Data, Inc.