Case study 1

Data privacy in water sciences

Zipper, S. C. et al. (2019). Balancing open science and data privacy in the water sciences. *Water Resources Research*, *55*(7), 5202-5211. https://doi.org/10.1029/2019WR025080

"Open science practices such as publishing data and code are transforming water science by enabling synthesis and enhancing reproducibility. However, as research increasingly bridges the physical and social science domains (e.g., socio-hydrology), there is the potential for well-meaning researchers to unintentionally violate the privacy and security of individuals or communities by sharing sensitive information. [..]

High-resolution spatial data include satellite data (and derived products), outputs of hydrological models, and other geospatial data sets. Geospatial data are commonly used in the hydrologic sciences, and unmanned aerial vehicles (i.e., drones; Kelleher et al., 2018), traffic/surveillance cameras (Jiang et al., 2019; Leitão et al., 2018), and increasing access to satellite data are likely to make these data less costly to collect and more widely available. Despite not meeting traditional definitions of human subject research, this type of data could be sensitive at the individual and community levels (Rissman et al., 2017). For example, 30% of Iowa farmers surveyed felt that collecting geospatial data on private land was an invasion of privacy (Arbuckle, 2013). [..]

Potentially sensitive consumer data include household consumption of water or electricity, or other variables that are of sufficient spatial or temporal resolution to be identified with and provide information about an individual or household (McKenna et al., 2012). While these data often have a spatial component to them, they are distinct from the previous category in that they quantify resource consumption (Helveston, 2015). The potential to monetize consumer information raises issues of data ownership, along with privacy. [..]

Digital trace data include deliberate online activities (e.g., social media and Web browsing) as well as Web-enabled technologies (e.g., the "Internet of Things") and can be divided into two groups: passively and actively contributed. Passively contributed data are posted to the internet without the intent or knowledge for potential scientific use (most social media data), while actively contributed data are contributed to a specific project (most crowd-sourced citizen science research). Both types of data have been used for hydrologic research. [..]

These risks can magnify when researchers lack cultural understanding of and sensitivity toward communities to which they do not belong. In some cases, people or companies in positions of power have taken advantage of open data at the expense of the intended beneficiaries of the shared data (Donovan, 2012; Gurstein, 2011; McClean, 2011). For instance, the digitization of land records in Karnataka, India, was promoted as a tool to democratize access to information, but instead allowed wealthy landowners with more financial resources to consolidate power and capitalize on these new data (Donovan, 2012). As seen through the lens of environmental justice, these concerns are particularly acute when working with historically disadvantaged groups such as impoverished communities and indigenous peoples (Brugge & Missaghian, 2006; Christen, 2015; Radin, 2017)."

Questions for discussion:

1. How sharing the three types of data mentioned in the case description might violate the privacy and security of individuals or communities?
2. Do you agree with the authors' statement that: "Natural scientists have little guidance to deal with privacy concerns for open science, which are inherent in socio-environmental research"?
3. What should the scientists do to protect data privacy and security?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430

Case study 2

Open geospatial data in agriculture research

Prince Czarnecki, J. M., & Jones, M. A. (2022). The problem with open geospatial data for on-farm research. *Agricultural & Environmental Letters*, *7*(1), e20062. https://doi.org/10.1002/ael2.20062

"On-farm research requires collection, curation, and analysis of spatially referenced farm data (e.g., as-applied fertilizer, plant populations, yield) that is easily traced to an individual farm, and accordingly private individuals participating as collaborators (Ferris, 2017). Where geospatial data are concerned, options for de-identification of farm geospatial data are not well addressed in the literature. Acceptable methods for general geospatial data such as random perturbation and temporal cloaking are poorly matched to farm data. Shifting point locations for individual yield points and altering time stamps in fertilizer prescriptions are not a solution for obscuring field location. These methods also have potential to change statistical measures of the data and in previous work did not provide an appropriate level of privacy (Broen et al., 2021). Another common method is to remove the geospatial reference and rescale points to a spatially correct, but non-georeferenced, grid. While this preserves the spatial relationship between points, it removes the opportunity to perform contextual analyses because features with geographic concurrence (e.g., climatic data) cannot be readily identified. To be clear, there is value in reporting average yield values in tabular form with a de-identified code (e.g., Farm A) especially if released with additional contextual information (e.g., soil type), for researchers conducting metanalyses or examining general trends. However, the value is diminished as within-field spatial trends cannot be identified from such datasets (Kounadi & Leitner, 2014; Zurbarán et al., 2020). The reusability of data (and thus adherence to FAIR principles) is consequently reduced as anyone desiring geospatial data reasonably hopes to conduct these types of spatial analyses. Thus, one argument in support of the primary contention—that data-sharing requirements will make on-farm research increasingly difficult and also not achieve the stated purposes of opening data—is that de-identification is not only insufficient to protect collaborator privacy but also places limitations on future research."

Questions for discussion:

1. How sharing the geospatial data mentioned in the case description might violate the privacy and security of individuals or communities?
2. What should the scientists do to protect data privacy and security?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under GA No 101006430