6-2018

Communicating about Data Privacy and Security

ADRF Network Working Group Participants

Follow this and additional works at: https://repository.upenn.edu/admindata_reports

Communicating about Data Privacy and Security

Description

The working group on Communicating about Data Privacy and Security identified six core stages of the administrative data research lifecycle and developed a framework for how researchers can conduct stakeholder engagement throughout the research lifecycle.

Co-Chair and Lead Author: Kelsey Finch (Future of Privacy Forum)
Co-Chair: Jules Polonetsky (Future of Privacy Forum)

Communicating about Data Privacy and Security

By Working Group Participants

ADMINISTRATIVE DATA
Research Facilities NETWORK

Acknowledgments

The following people served as Working Group participants. Together, they discussed current challenges with communications in administrative data research and developed a preliminary framework for stakeholder engagement.

Contents

Summary

This report summarizes the discussions and preliminary model stemming from the ADRF Network's working group on Communicating about Data Privacy and Security ("Communications Group"). We were motivated by recent backlash that resulted from poor communications in data privacy and use for research. These examples in the news have overshadowed the benefits of using administrative data for knowledge creation and better policymaking. Our report, intended for the broader administrative data researcher and data user community, is an initial effort aimed at improving practices in stakeholder engagement and communications. In this report, we identified the "why, when, who, and how" of communicating about data privacy and security while doing administrative data research. We developed an initial matrix tool for stakeholder engagement and applied the model to a hypothetical example. We recommend further work on this issue as the field of administrative data research continues to grow.

Background

Launched in June 2017 with funding from the Alfred P. Sloan Foundation, the ADRF Network is comprised of researchers, practitioners, and other stakeholders working to improve how administrative data are accessed and used for social science research and policy. One of the ADRF Network's early initiatives was to form three working groups around high priority issues and questions in the social science research space. The three working group topics are 1) "Data Quality and Standards", 2) "Data Sharing Governance and Management", and 3) "Communicating about Data Privacy and Security", the focus of this report.

The working groups first met in November 2017. At its first meeting, the Communications Group decided to focus on issues related communications rather than the technical aspects of data privacy and security. The group agreed that even if all the technology and security pieces are well done, research projects can still fail if research teams fail to engage and communicate with the appropriate stakeholders throughout the research process. Privacy and responsible data use can be both highly technical and highly personal subjects, with significant impacts on individuals and communities if not taken seriously. Miscommunications or silence by organizations about how personal data will be used and protected can lead to public mistrust,

internal protests, project collapse, and even legislative backlash (Bulger, Mccormick, and Pitcan, 2017; Kahn and Ingram, 2018; Shane and Wakabayashi, 2018; Strass, 2014). Engaging diverse stakeholders early and often – and providing them meaningful opportunities for input – can establish lasting trust, create legitimacy in administrative data research, and help researchers get ahead of privacy disasters.

Why engage?

Both practical and ethical considerations should drive researchers to engage early and often with those most affected by the results of their study and with those who contributed the data. We briefly illustrate four reasons below for why researchers should engage in communicating about data privacy and security.

- **Increase the value of data and its usage**. Engaging with stakeholders creates an environment of trust and legitimacy, in which people are more likely to value and use data. We think of stakeholders broadly and include people whose data are directly used for research, who are representative of the research population, who provide the data, who are conducting the research, and who will make decisions based on the outcome of the research. If stakeholders hear about administrative data and research for the first time when there is a problem, they are unlikely to want projects to succeed or to support future efforts.

- **Comply with ethical obligations**. Stakeholder engagement and outreach can serve as important transparency and accountability tools, which can support legal and ethical privacy commitments. Where traditional informed individual consent is infeasible, for example, community oversight boards or ethical review panels may be effective surrogates.

- **Preventing backlash**. Failing to engage or communicate effectively about privacy and responsible data use can spark a backlash, internally and externally. By being proactive, however, not only can researchers resolve privacy concerns, but also begin a more forward-looking conversation about the benefits of data use.

- **Improve project design and implementation**. Engaging in privacy communications can parlay stakeholders' skill and expertise into improving project design and implementation. Incorporating traditionally excluded voices throughout the project lifecycle, for example, can provide researchers with additional context for how accurate or representative administrative data are or can surface new research questions and priorities.

When to engage?

This type of engagement should occur throughout the development of a research project. For conceptual clarity, the Communications Group identified six stages of the Administrative Data Lifecycle, modeled after the UK Data Archive lifecycle chart, from project conception to post-project (Research Data Lifecycle, 2018). In the engagement matrix tool that we present in the report, we set up one axis for different research stages ("when to engage") and one axis for the level of engagement ("how to engage"), recognizing that different research stages may warrant different levels of engagement. Appendix 1 shows the template of the engagement matrix. Below, we describe each stage of the lifecycle:

1. **Project Conception.** A research project begins by defining a research problem and formulating the specific research questions. This stage may include determining the minimum viable data needed and identifying the characteristics of datasets appropriate for the study population and research questions. Researchers may also design research protocols. Finally, there might be a testing or proof-of-concept project to show the viability and value of the research question.

2. **Accessing and/or Collecting Data**. In the second stage, the researcher begins to identify the relevant data owners and data intermediaries that own the data. For restricted-access and/or private datasets, they might submit research proposals through a data intermediary or the data owner directly. Researchers and other data users from the project might then obtain legal agreements as appropriate and sign data use documents, agreeing to terms on data access and use. After the legal and administrative paperwork are complete, the researchers gain access to data relevant to the research question.

3. **Cleaning, Linking, and/or Deidentifying Data**. Upon gaining access to the data, researchers will typically prepare the research dataset by cleaning and/or linking the data according to the specifications outlined in the data use documents. This might include checking variable distribution, consulting the data documentation, and dealing with missing data. The dataset may also need to be linked with another dataset, which may or may not be restricted access. If de-identification is required, individual identifiers must be stripped and data may be aggregated. Finally, researchers or the data intermediary will manage and store the research-ready data.

4. **Analyzing Data**. In this stage, researchers analyze and interpret the results. Research outputs are produced and may require additional disclosure review to limit the risk of re-identification. If the analysis is performed by segments of the population, researchers should ensure that their results will benefit and not harm individuals and groups, particularly those from vulnerable or traditionally disadvantaged communities.

5. **Publishing and Sharing Results**. During the publishing and sharing results stage, researchers publish their research findings. In preparation for publication, researchers might cite data source and create appropriate metadata and research documentation. After the results are published, researchers will promote their findings within the academic community and for the greater public. Researchers might share the research context to help others understand the findings, including presenting definitions, explaining comparisons and trends, providing recommendations, and describing data sources and research limitations. Researchers might strive to be transparent and clear about how data was protected throughout the research processes

6. **Post-Project.** After the research is complete and results are published, researchers might work with data intermediaries to prepare the research data for long-term storage or archiving, proper destruction, or further de-identification, as appropriate. Researchers might use this stage to scrutinize their findings and review project processes, including identifying its strengths and weaknesses, engagement impacts, and lessons learned. Finally, this stage may lead researchers to developing follow-up research that will take them back to the beginning of the data lifecycle.

Who to engage?

As mentioned above, we consider a broad range of stakeholders with whom researchers should engage when communicating about data privacy and security. Stakeholder assessment should be done at various geographic levels (i.e., local, state, national, international). Box 1 shows a non-exhaustive list of stakeholders.

How to engage?

We applied the engagement framework from the GovLab's "People-Led Innovation" report to our work. From most active to most passive, the GovLab report identified four engagement modes, described below (Young et al., 2018, p. 15). Again, these four engagement modes take up the "how to engage" axis in the matrix tool that we present in this report.

Co-creating: Individuals and/or groups are asked to apply their skills and creativity to the different phases of the innovation cycle with the problem-solving team. For example, a sector-specific hackathon wherein people seek to leverage datasets to create new solutions to public problems.

Reviewing: Individuals and/or groups are asked to review approaches or initiatives in a more targeted manner – including assessing and evaluating proposals and/or interventions. For example, online or offline engagements allowing people to "upvote" or "downvote" specific proposals or ideas, or using annotation platforms to leave suggestions.

Commenting: Individuals and/or groups are given opportunities to share their opinions, priorities and preferences. For example, using a discussion platform to solicit complaints or experiences among residents to help prioritize problem areas.

Reporting: In the Reporting role, individuals and/or groups are asked to contribute data and facts to inform problem definitions, solution plans, and evaluations. For example, a

Box 1. List of Stakeholders

- Data subjects and representatives of the study population
- Data owners and contributors
- Other data users
- Administrative and political government leaders
- Academic institutions and partners
- Internal and external data experts – privacy, security, disclosure control, ethics, etc.
- Internal and external domain experts – criminal justice, education, public health, etc.
- Advocacy organizations – vulnerable and minority populations, local community and neighborhood groups, etc.
- Advocacy organizations – good government groups, evidence-based policymaking, etc.
- Business leaders (e.g., Chambers of Commerce)
- Technology service providers
- Peer networks (e.g., NNERP)
- Funders
- Media
- Regulators

crowdsourcing platform for citizens to collect incidences of local issues like graffiti or potholes for government officials to address.

We further brainstormed a broad range of engagement activities that research teams can pursue. These activities are listed in Box 2 and are categorized into less active and more active activities. All planned activities should be assessed for inclusiveness and diversity considerations (e.g., language, location, time, transportation, childcare, food, incentives, appeal, power dynamics, etc.).

Sample Engagement Matrix: Promise Scholarship Pilot (Hypothetical Example)

Having described the two axis of the engagement matrix, we now present an example of how the tool can be applied to a research project. Matrix 1 is an example that illustrates many possible ways that a hypothetical school district and team of researchers could actively engage a diverse set of stakeholders throughout a multi-year promise scholarship pilot program. While this sample matrix is meant to emphasize the variety of options a project team could consider to engage stakeholders, **in real world situations the project team would use the combination of engagement modes and activities that best fits their needs based on the scale, scope, and context of their data-driven activities, as well as available time and resource considerations.** Not every stage in the Administrative Data Lifecycle is equally suited to particular modes of engagement, and less active engagement at one phase can be supplemented by more active engagement at another (e.g., a team may offer co-creation opportunities to the public during project conception, but only commenting opportunities while cleaning and linking the data).

Box 2. List of Engagement Activities

More active:

- Public meetings
- Stakeholder briefings
- Hackathons
- Focus groups
- Workshops and working sessions
- Citizen advisory committees
- Expert panels
- Participatory decision-making
- Open houses

Less active:

- Public surveys, interviews, questionnaires
- Notice and comment periods
- Social media engagement
- Public voting/ballots
- Discussion or annotation platforms
- Infographics
- Reports
- Fact sheets
- Websites
- Newsletters (online/offline)
- News articles and op-eds
- Exhibits/displays in public areas

This hypothetical example is from the ADRF Network Working Group on Data Quality and Standards:

> *Consider the needs of a large metropolitan school district and how a well-documented integrated data system (IDS) can support policy planning, monitoring, and evaluation of student educational outcomes. College matriculation rates are low in the district. [...] Upon gaining an understanding of the problem, say the school district posits that a scholarship program could be an effective way of raising college matriculation rates.* ***The school district approaches a private funder interested in supporting and evaluating a promise scholarship program. This pilot program would offer high school freshman a college scholarship if they met certain academic requirements.*** *The funder and the school district want to monitor how well such a program is being implemented, or even evaluate whether this scholarship offer improves students' outcomes such as academic performance while in high school, future income, public assistance participation, and criminal justice involvement after high school graduation, as well as their families' income and public assistance program participation while the students are in high school.*

Recommendations for Next Steps

We believe that stakeholder engagement and communicating about data privacy and security are crucial to the future success of administrative data research. While strong privacy safeguards are the foundation of any administrative data research, learning to effectively communicate about how and why administrative data are being used and protected and providing stakeholders with meaningfully input in the research process is essential to maintaining public trust. Without public trust, administrative data research will struggle to find support and legitimacy. Through the working group, we proposed a preliminary model to help researchers and research teams communicate with relevant stakeholders and community members with varying levels of engagement.

Future work should refine the model and best practices for different types of administrative data research. For example, researchers using de-identified, secondary data to conduct research at the national level may choose different engagement strategies than researchers carrying out a randomized controlled trial in a specific community. The DC study on police body-worn cameras is an example of the latter that incorporated public input and community engagement throughout the research process (Yokum, Ravishankar, and Coppock, 2017). Once the engagement model is refined, we recommend a pilot to apply the model to a

real-life administrative data project. These steps are important for building trust and public buy-in with administrative data research while the field is still relatively young.

References

Bulger, M., Mccormick, P., & Pitcan, M. (2017). *The Legacy of inBloom.* Retrieved from Data & Society Research Institute website: https://datasociety.net/output/the-legacy-of-inbloom/

Kahn, C., & Ingram, D. (2018). Americans less likely to trust Facebook than rivals on personal data: Reuters/Ipsos poll. *Reuters.* Retrieved from https://www.reuters.com/article/us-usa-facebook-poll/americans-less-likely-to-trust-facebook-than-rivals-on-personal-data-reuters-ipsos-poll-idUSKBN1H10K3

Research Data Lifecycle. (2012). In *UK Data Service.* Retrieved from: https://www.ukdataservice.ac.uk/manage-data/lifecycle

Shane, S., & Wakabayashi, D. (2018). 'The Business of War': Google Employees Protest Work for the Pentagon. *The New York Times.* Retrieved from https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html

Strauss, V. (2014). $100 million Gates-funded student data project ends in failure. *The Washington Post.* Retrieved from https://www.washingtonpost.com/news/answer-sheet/wp/2014/04/21/100-million-gates-funded-student-data-project-ends-in-failure/?noredirect=on&utm_term=.e57ea8945bff

Yokum, D., Ravishankar, A., & Coppock, A. (2017). *Evaluating the Effects of Police Body-Worn Cameras: A Randomized Controlled Trial.* Retrieved from The Lab @ DC Website: http://bwc.thelab.dc.gov/TheLabDC_MPD_BWC_Working_Paper_10.20.17.pdf

Young, A., Brown, J., Pierce, H., & Verhulst, S. (2018). *People-Led Innovation: Toward a Methodology for Solving Urban Problems in the 21st Century.* Retrieved from The Gov Lab website: http://www.thegovlab.org/static/files/publications/people-led.pdf

Matrix 1. Applying the Engagement Matrix to the Promise Scholarship Pilot Example

	Active engagement ← – – – – – – – – → Passive engagement			
	Co-creation: *Individuals and/or groups are asked to apply their skills and creativity to the different phases of the innovation cycle with the problem-solving team.*	**Reviewing:** *Individuals and/or groups are asked to review approaches or initiatives in a more targeted manner – including assessing and evaluating proposals and/or interventions.*	**Commenting:** *Individuals and/or groups are given opportunities to share their opinions, priorities and preferences.*	**Reporting:** *Individuals and/or groups are asked to contribute data and facts to inform problem definitions, solution plans, and evaluations.*
Project conception	Past and present students, teachers, parents, and others are invited to workshops to identify potential barriers to academic success within their community and prioritize them based on their impact and feasibility of being solved.	IDS peers, academic partners, domain experts, and data experts privately evaluate the proposed minimum viable data needed to evaluate the scholarship's impact on high school academic performance.	Funders, school leaders, and IDS leaders host accessible public meetings within the metropolitan school district, presenting the proposed pilot program and research study and giving community members a forum to share their comments, concerns, and questions.	Funders, school leaders, and IDS leaders publish or present the premises and evidence on which the pilot program is based, and then invite teachers, students, parents, domain advocates, and the public to provide comment and fact-checks.
Accessing and/or collecting data	A community advisory committee (representative of metropolitan, IDS, school, parent and student leaders) creates data use agreements for the pilot.	A group of internal data experts from across local agencies provide a risk assessment reviewing the proposed pilot datasets.	Researchers within IDS peer networks or other academic partners are asked to share their experiences working with similar datasets in a working session or focus group.	Education, public services, and criminal justice domain experts brainstorm potentially relevant data sources for use in the pilot.

Cleaning, linking, and de-identifying data	A disclosure review board of internal and external data experts, IDS leaders, and data users determines whether and how particular datasets can be linked and used for the pilot.	A multidisciplinary group of internal domain and data experts reviews and votes on specific policies and procedures for researchers to better check, validate, and clean data to be used in the pilot.	External experts and other interested groups are asked to provide comments and suggestions on a public report documenting the pilot program's proposed de-identification policies and safeguards.	Local high school and college students research data quality and sanitization methods and present to IDS/school project leaders or community advisory board (at appropriate levels of sophistication).
Analyzing data	An ethical panel with inclusive community representation conducts a fairness review to ensure that benefits of analysis/disaggregation of data outweigh risks to individuals and groups.	A working group of internal data and domain experts is asked to review and share questions and suggestions on preliminary analysis or interpretations, such as on an internal annotation platform.	Trusted advocates and external domain experts come together in working sessions to discuss the strengths and limitations of the proposed analytical methodology.	Interview data subjects and participating institutional representatives (such as students, parents, high school teachers, college admissions officials, police, public assistance case workers, etc.) about their experiences throughout the pilot.
Publishing and sharing results	Local political, education, and public services leaders and funders host community events where stakeholders help crystalize and scale the pilot's policy implications, such as by helping prioritize future scholarship program sites	IDS leaders give advanced briefings on the final analysis and results of the scholarship pilot to relevant data and domain advocacy organizations, asking participants to evaluate the impacts of the findings on their	IDS leaders promote their findings and conduct Q&As on social media, soliciting public comments on the findings and suggestions on next steps for the pilot and the research.	Local institutions (businesses, universities, other funding organizations, etc.) are asked to identify ways in which the pilot's findings might inform or impact their operations, which can be highlighted in

	or making recommendations for actions that can be taken based on the findings.	communities or identify areas for additional study.		the IDS team's newsletter or findings.
Post-project	Metropolitan and IDS leaders host idea workshops or policy jams with community stakeholders and individuals to generate new research proposals or programs for improving students' academic	IDS peers, academic partners, and other data researchers gather in a working session to reflect on the strengths and weaknesses or lessons learned from the pilot process.	Advocacy organizations, external domain experts, funders, or other interested parties are invited to complete surveys indicating which data elements or findings were the most/least useful to them, to inform future research.	Set up reporting mechanisms for data subjects or participating institutions to provide additional feedback about their experiences or post-pilot lives, if they desire.

Appendix 1. Template of Engagement Matrix

	Active engagement ⟵ ⟶ *Passive engagement*			
	Co-creation	**Reviewing**	**Commenting**	**Reporting**
Project conception				
Accessing and/or collecting data				
Cleaning, linking, and de-identifying data				
Analyzing data				
Publishing and sharing results				
Post-project treatment				