### Summary (approximately one single-spaced page)

**Provide an overview of the key elements of the case study.**

The goal is to give potential readers a quick idea if the case study describes a scenario similar to their own, what topics it covers, and how they might apply its lessons. The authors should:

* Give an overview of the data:
  + Summary of data content
  + Interesting uses of the data
  + Data provider and data host: name, location, and range of activities
  + For example: “This chapter describes a partnership between the University of [country] and the traffic and transportation department of [city] in [country], an agency with 68 employees in charge of traffic and parking regulations, road use permits, construction, and public bus transport for xx residents and yy annual visitors.”
* Briefly summarize key parameters and process of making data usable and accessible:
  + Timeline and current status
  + Hosting arrangement: data provider, researcher, or other third party?
  + Cost and resources used, staffing
  + Involved parties; e.g. data collectors and data curators (if different from data provider)
  + Access modalities
* Summarize main insights and contributions of this chapter:
  + Key conclusions and recommendations for others
  + Any new resources or systems innovations described
  + Potentially highlight elements not covered
* Reference/describe 1-3 main sources for the case study, if existing:
  + Role of the authors in the process above
  + An existing publication about the case study or a website with information about data access

### Introduction

#### Motivation and Introduction

**Explain what drove the original project to make the data available.**

Reasons could include a legal mandate, a promising research project, or a pivot towards data-driven decision making. Was the initiator a researcher or the data provider (business, non-profit, government), and what were their reasons or interests? What were the desired outcomes of the project?

*This section may name a legal mandate to make data accessible, but the details (how access is organized and who has access) should be discussed in the next section “Legal Context”.*

#### Data Use Examples

**If available, provide examples of particularly interesting, innovative, or policy-relevant uses of the data.**

Connect these examples with a description of the available data (see section “content guidelines” point 4 above). Particularly valuable are examples where

* Experiments were conducted using the data, in collaboration with the data provider or independently
* Administrative data was linked with other data sets or survey data
* The data and research were used for policy improvements.

Figures or tables can be used to support the data use examples. Do not describe prospective or speculative uses. Cite published papers as well as any policy decisions that were made based on analysis of the data.

#### Making Data Usable for Research

**Describe the steps for transforming the operational or administrative data records into usable data files (including metadata and documentation) for analysis.**

This section should discuss both the preparation of the data itself and the data documentation and metadata. Focus on information that can help others understand the general process and the time and cost involved. Steps might be:

1. Retrieving individual data records or operational data sets and standardizing formats
2. Standardizing variables and matching records across data sets
3. Cleaning, verification, reconciliation, and quality control
4. Analysis file documentation
5. Data documentation and metadata generation and publication
6. Automation of the above steps, if any

Describe the resources and time needed to carry out these steps, from staffing to software and systems. Who does this work – researchers, in-house staff, contractors? Please provide examples of or references to documentation and metadata, if useful and available.

Modifications that are needed to protect the data before researchers can access them (e.g. de-identification) or before research findings and outputs can be published should be described below under “Safe Data” and “Safe Outputs.

### Legal and Institutional Framework

#### Institutional Setup

**Describe how data access is organized from an institutional standpoint.**

Who are the parties to the data access mechanism in this case study? Is the data provider also the data custodian/data host, or has a third party (data intermediary) taken on this role? Who provides or provided legal advice to the parties?

As an example, in some cases the data provider (say, a private firm) conducts all preparation of the data for research in house, and contracts directly with researchers (“two-party model”). In other cases, the data provider or data providers (say, different ministries in a national government) contract a data custodian (data intermediary) who aggregates and prepares data and negotiates data access with researchers, representing the data provider(s).

#### Legal Context For Data Use

**Describe the legal context that permits (or mandates) if and how the data provider gives others access to the data.**

This section should focus on the legal obligations of the data provider towards those whose personal or proprietary information is contained in the data (respondents, taxpayers, firms, etc.), or those who have curated or created the data (if different). Please provide references to applicable laws (see above on legal citations).

* Does the legal framework mandate data ownership or privacy protections for constituents, clients, etc. whose data are being shared, and how clear and established are these protections?
* Can third parties request purging of records or do data have to be deleted at certain time intervals (e.g. juvenile justice, deleted tweets)? Note: data purging was moved here from a separate section. If the data provider puts such rules in place without a legal mandate, please discuss in the “five safes” section.
* Does the data provider legally need to have consent from the “producers” of the data to share the data with researchers?
* Are there any legal restrictions on who can be given access to the data? Reference legal basis here, but discuss safety restrictions on research access, and how they are implemented, in detail under the “Five safes” rubric.
* Are there any legal sanctions prescribed for data intermediaries or employees of the data provider for allowing for unauthorized sharing, access, or uses of the data? Note: sanctions for data users can be referenced here but their implementation should be described in detail in the “five safes” section.

*Note: the agreement between the data provider and the users of the data (e.g. researchers) should be described below and in the “five safes” section (with reference to data sensitivity and confidentiality). The legal framework for access should be described in a way that is intelligible to lay persons and an international readership.*

#### Legal Framework for Granting Data Access

**Describe what legal agreements regulate the use of the data by others.**

This section should focus on the legal relationship between the data provider and the data users (researchers, other agencies or parts of government, the public etc.). Typically, data access is granted through or governed by some form of agreement - a non-disclosure agreement (NDA), a data use agreement (DUA), a protocol, proposal, or framework. This section should discuss such agreements, or, if no individual agreement is typically made, how else data use is governed (e.g. by applicable laws, informal agreements). If possible, the authors should provide an example or template, pruned of any identifying information.

* How much flexibility and individual negotiation goes into a typical data use agreement?
* Can the data provider impose sanctions (financial, reputational, penal) for unauthorized uses of the data? How were those chosen? Have they ever been imposed? Note: discuss details or reference back to this part in the “five safes section.
* Can the provider revoke data access (for cause, without cause, etc.)?
* Are there any payments, conditions for access, or other obligations to the researcher specified in the use agreement, for example co-authorship on research output? Note: discuss details or reference back to this part in the “five safes section where this relates to data protection.
* Does the data provider assert Intellectual Property (IP) on the original data, or on any derivative data or products created by researchers with access to the data, such as tables, research papers, source code, etc.? How is IP enforced? Note: this was previously in a separate “IP” section.
* Is there any review of outputs, e.g. does the data provider have a veto right over outputs? What are the review criteria? Note: review for disclosure risk and sensitive data should be discussed in detail in the “five safes” framework. Any review related to business or political strategy, or intellectual property rights, of the provider should be discussed here.
* Is there a specific copyright or license on the data and code generated by the project?

*Authors may perceive some overlap between this section and the “Five safes” section. This and the previous subsection should focus on the broad legal justification and framework for granting research access. For instance, if the criteria described in detail in the Five Safes section are prescribed by a law, then describe the law here, and the details of the access under the Five Safes. Describe here if the formal agreement is called and formulated as a license, a contract, a data use agreement. This section should also describe any protections for the data provider – for instance, liability limitation if a breach occurs by a researcher, intellectual property rights, non-competition clauses.*

### Protection of Sensitive and Personal Data: The “Five Safes” Framework

*This section should focus on aspects of the data access process that relate to ethics, data security, privacy, and protection of confidentiality. It is organized using the “Five Safes” framework. Please assess the cost and importance of each of the five aspects of data security below in terms of importance and cost on a scale from 1 (lowest) to 5 (highest).*

#### Safe Projects - Evaluating data analysis projects for appropriateness

**Describe the application and review process and approval criteria for data access requests.**

If available, include information on application management systems, any IRB or other review requirements, and which expertise is sought for the review of applications (e.g. legal or statistics experts). Describe how long the approval process takes and what it costs. Does the data provider require consent from the “producers” of the data to share the data with researchers?

*If there is an online portal or application process information is publicly posted, please provide a reference/citation. Note that this section should focus on protections for those in the data, e.g. is the data use appropriate, are sensitive data protected in the project, and do the benefits of using the data outweigh the risks. For other aspects of project review – in particular intellectual property rights or strategic business interests – please use the “legal framework” section above.*

#### Safe People - Evaluating researchers who seek data access

**Describe what conditions and credentials are required to be granted data access.**

* What are the requirements to apply for access and how, and by whom, are they verified - citizenship, professional standing, background security checks…?
* Do researchers need to have human subjects training or any other kind of training? Who provides this training? How frequently does it need to be renewed?
* Does the organization trust researchers for access procedures, for disclosure avoidance procedures, for data handling? Does it audit these? How are violations sanctioned? Note: refer back to the legal framework for sanctions if necessary.
* Do repeat users get fast tracked? How many users can be handled?

*If the data provider uses a “circle of trust” model, please describe it.*

#### Safe Settings - How can the data be accesssed?

**Describe the technical and physical infrastructure for data access and the handling of security breaches.**

* Please describe software, systems, and IT resources and weigh their advantages and disadvantages and their cost.
* How is access restricted, e.g. is it possible only at a specific physical location, a secure computer, remote submission, or via secure remote interactive access (sometimes called a virtual data enclave), etc.? Are there multiple access settings?
* Are the various safe settings only used for researchers, or are they also used internally (by staff of the data provider or the data intermediary)?
* How is researcher-provided software and code handled?
* Security: What are considered to be breaches of the safe settings? How do you handle them, who is notified about suspected or actual breaches? What are the penalties? *Refer back to “legal” section if needed.*

*Refer to the handbook chapter on “Physical security” if necessary. Be concrete and provide references if possible to help others who want to implement the same solutions.*

#### Safe Data - How is disclosure risk the data verified and mitigated?

**Provide concrete information on the data processing steps involved in modifying the analysis data files for different levels of (researcher) access.**

* Describe the resources and staff capacity invested in creating researcher-accessible data files. Are these files created as part of your data processing pipeline, or for researchers only; either on-demand, or on a schedule?
* Are synthetic or fake datasets used, combined with remote processing or validation?
* What methods are used to make the data “safe”? Are data elements stripped or coarsened? To what degree are the researcher-accessible data altered from the original data, and what are the consequences for the usability of the data?
* Discuss in particular if it is possible – and what steps are involved – to link the administrative data with other external data sets, such as survey data, that have personal identifiers. Is it possible to conduct field experiments with the data (e.g. by linking treatment information with the administrative data set, or by using the administrative data as a sample frame)? *Note: added question.*

*Note: these are processing steps carried out after the data was made usable for research in order to remove PII and/or obscure indirect identifiers. Discuss other data processing for analysis in the “making data usable” section.*

#### Safe Outputs- How is disclosure risk in statistical analysis results and tabulations verified and mitigated?

**Describe processes, requirements, and tools used to verify the disclosure risk from published analysis results.**

* Are safe-output rules incorporated into data use agreements?
* How (and how often) are safe outputs verified?

Discuss the practical implementation, including e.g. how long the review takes and who conducts it. *Note: refer back to the legal section above regarding output review as needed. The present section should focus on how review for data protection is implemented.*

### Data Life-Cycle and Replicability

#### Preservation and Repoducibility of Researcher-Accessible Files

**Describe to what degree and over which time period researcher-accessible files and master files are preserved.**

* Is there active curation, or only preservation of current bitstreams?
* Does the data provider perform the archival/curation functions itself, is another entity within the organization responsible for this activity, or is a third-party (national, regional, university archive) responsible for this?
* Are persistent identifiers generated for these files?
* Can researcher-accessible files be consistently regenerated, or are they snapshots of a dynamic database or data pipeline? Is the query mechanism reproducible (can similar files be recreated, can older files be retrieved), or manual (with risk of inconsistency, variation across personnel)?

#### Preservation and Reproducibility of Researcher-Generated Files

**Are researcher-generated intermediate and final data files and tables, and processing code and programs preserved?**

* Can other researchers access these files, with or without the permission of the original researchers?
* Does the data provider generate persistent identifiers for any of these?

### Sustainability and Continued Success

#### Outreach

**Provide examples of successful outreach activities that have made “safe people” aware of the access point and data offerings.**

Discuss outreach by data intermediaries/researchers to data providers as part of “Metrics of Success”.

#### Revenue

**Discuss what makes the data access mechanism financially sustainable.**

How are incremental costs covered, is there full cost recovery? What is the pricing model?

*Where possible, precise numbers should be used.*

#### Metrics of Success

**Describe how the success of the data access project is evaluated.**

What metrics for success are used, and who is the audience for metrics - funders, data provider, legislature?

*Provide sample statistics where possible. Note: section was renamed.*

### About the Authors

*Note: Authors may expand here on who they are, providing a short bio, and including additional affiliations, support, and relevant related activities.*

### Disclaimer

The views expressed in this paper are those of the authors and not those of any sponsors.