Integrating Administrative Data for Policy Insights

Justine S. Hastings[1,2,3,4], Mark Howison[1,2], Ted Lawless[1], John Ucles[1], Preston White[1]

[1] Rhode Island Innovative Policy Lab, Brown University, Providence, RI, USA
[2] Watson Institute for International and Public Affairs, Brown University, Providence, RI, USA
[3] Department of Economics, Brown University, Providence, RI, USA
[4] National Bureau of Economic Research, Cambridge, MA, USA

When properly secured, anonymized, and optimized for research, administrative data can be put to work to help government programs better serve those in need.

Key Insights

- Fact-based policymaking – the practice of using data and research to guide policy decisions – is a promising solution to improving the effectiveness and efficiency of government programs.
- Administrative data can provide new facts to guide policymakers. However, understanding the quality of administrative records, and integrating, transforming, and optimizing them for policy insights presents many challenges.
- To overcome these challenges, we developed an integrated database of administrative records from multiple Rhode Island state agencies with over 800 tables and 2.7 billion records related to over 4 million anonymous individuals. These data support econometric and machine-learning research into policies with promise to deliver higher impact per dollar and better serve Rhode Island families.

Introduction

There is a growing consensus among policymakers that bringing high-quality evidence to bear on public policy decisions is essential to supporting the more effective and efficient government programs Americans want and need. At the federal level, this view is reflected in a recent congressional report by the Commission on Evidence-Based Policymaking, which recommends creating a data infrastructure that enables “a future in which rigorous evidence is created efficiently, as a routine part of government operations, and used to construct effective public policy” [2].

Administrative data can be collected from the computer systems used by government agencies to administer their programs, and transformed into databases of anonymized records suitable for insights, providing new sources of facts for policymakers to benchmark goals and measure the successes and shortcomings of existing and future programs. Often classified as “big data” due to their volume, variety, and availability [6], administrative records are also an increasingly valuable source for empirical social science research [3]. Research with administrative records can contribute new data-driven insights to inform important policy decisions (see Sidebar #1), and add objectivity and scientific rigor to measuring program impact and designing effective program changes. Moreover, scientists can inform how data from administrative systems, which are primarily designed around operational needs and often not suitable for analysis, can be transformed effectively to support research and insights.

Sidebar #1: Recent Data-Driven Insights from Administrative Records

- Records from the New York City criminal justice system show how judges often mispredict risk when making bail decisions [10]. Judges identify and release many defendants who have a low flight risk, but also release nearly half of the defendants with the highest flight risk. In simulations, replacing judges' decisions with a machine-learning prediction can reduce either crime rates (at a fixed jailing rate) or jailing rates (at a fixed crime rate), and in both cases can reduce racial disparities in outcomes.
- Transaction data from a private grocery retailer and data from the Supplemental Nutrition Assistance Program in Rhode Island show that households treat their nutrition benefits as if they were earmarked for food expenses, even when they could be substituted for cash [9]. This finding contradicts traditional economics theory which predicts that nutrition benefits should be fungible (e.g. substitutable for cash), and instead supports an alternative economics hypothesis called mental accounting. Results suggest that SNAP program impact on spending and nutrition can be influences by policies governing when and how benefits are distributed.
- Federal income tax records show there are growing inequalities in life expectancy in the U.S. across socioeconomic factors [1]. The breadth and scale of these administrative data (with over 1.4 billion person-year observations) reveal that geographic factors like government expenditure and fraction of immigrants and college graduates are positively correlated with life expectancy at the bottom of the income distribution.
- Randomized field experiments in Chicago combined school and unemployment insurance records with arrest records to evaluate the impact of a summer job support program for youth [4]. By using this integrated administrative data, the study found that the program caused declines in violent-crime arrests even though there were no significant effects on school or employment outcomes, which are the more typically studied effects of youth job programs.

As part of our work building the Rhode Island Innovative Policy Lab (RIIPL), a partnership between academic researchers and the State of Rhode Island, we have successfully constructed an anonymized database, called *RI 360*, from the administrative records of nearly every Rhode Island state agency. These integrated data support ongoing research into improving policies to alleviate poverty, and increase economic opportunity for all Rhode Island residents. The comprehensive scope of these data enable research projects across a wide range of policy challenges, such as lowering non-urgent emergency health care costs, curbing the opioid epidemic, improving worker training programs, creating tools to connect dislocated workers to benefits, helping families become more food secure, optimizing energy policy for low-income families, helping children reach proficiency on reading and math tests, and closing the college achievement gap.

In this article, we discuss our process for building a relational database from siloed agency databases, and transforming them into a usable framework for research. Although we demonstrate this approach with a state government partner, similar principles apply to federal, county, or local levels of government. The key challenges we had to overcome were:

- Deploying security controls to physically protect data and prevent data leakage.
- Developing a method for identifying individuals and joining their data across multiple tables, but while also preserving their anonymity so that researchers cannot discover their identity, even inadvertently.
- Accommodating the needs of government partners and avoiding additional development effort, by supporting data extracts in many different formats.
- Transforming the data to support research and replicable findings.
- Defining procedures for documenting and exporting research results, to maintain compliance with data sharing agreements.
- Implementing automatic documentation and lab procedures to ensure transparency and high quality research output.

Securing the data

We start by isolating the research database in a secure computing environment called a *data enclave* [11]. The enclave's key features are that it is physically secure and isolated from the Internet, data transfers in and out are restricted and subject to a documented approval process, all access is comprehensively audited, and access is granted to only a limited group of approved researchers. These security controls protect against both unauthorized access and unapproved use of data, and ensure that researchers access the data in compliance with the legal agreements governing their use.

With the physical and network security in place, we can begin transferring data from multiple agencies into the enclave to achieve two competing goals: joining records about the same individual across multiple data sources, and anonymizing records so that researchers cannot determine the identity of an individual. Figure 1 diagrams the process we use to achieve these goals, and we will refer to it throughout the article.

The original data are securely transferred from each agency into an encrypted tank in the enclave (Figure 1a). Access to the encrypted tank is protected by a two-party password, known only by senior leadership. A two-party password means two people each know a different half of the password, and both of the senior parties have to be present and consent to access the encrypted tank. This ensures that no individual researcher can access data that may reveal personally identifiable information (PII).

Once the original data have been successfully transferred into the encrypted tank, we run an automated pipeline to separate the PII (Figure 1b). Sensitive identification numbers – such as social security numbers (SSNs) or other agency identifiers deemed sensitive by the agency – are flagged by the agencies ahead of time and are automatically replaced with irreversible hashes, a technique that is widely used for protecting passwords [7]. Following this separation, the remaining data contain no PII and are de-identified (Figure 1c).

Anonymizing the data

One of the biggest challenges with using administrative data in research is identifying the same individual across data sets [3] while preserving the highest level of confidentiality and anonymity possible and ensuring that information about an individual cannot be viewed alongside an individual's identity. Agencies may use multiple sources of information to uniquely identify individuals in their records, and may not consistently identify individuals across records. Without a universal identifier, the task of identifying unique individuals is difficult, and joining individual-level data across agencies becomes more complex as the number of agencies and records increases. This creates several problems for research and analysis: researchers could end up spending more time identifying and joining data than actually performing analysis; that effort could be duplicated across projects; inconsistencies may arise if different projects take different approaches to joining data; and individual identities may be seen by researchers alongside data on those individuals during a matching process thus lowering the degree of anonymity provided during the database construction process.

To address these issues, we construct a global identifier called the RIIPL ID (Figure 1d). Using the RIIPL ID, researchers can join information about an individual without knowing any PII for that individual. Assigning RIIPL IDs to records occurs after separating PII (Figure 1d). An automated script concatenates all hashed SSN, first name, last name, and date of birth (DOB) records into a single table, while maintaining an encrypted link to the source table and row for the records[1]. If a record contains information on multiple individuals (such as a birth record that describes both the child and the parents), it is expanded into one row per individual. All names are cleaned to remove non-letter characters, and first names are converted to Soundex values [13].

A RIIPL ID is assigned to every valid hashed SSN, and to every distinct combination of first name Soundex, last name, and DOB that cannot be matched to a single valid hashed SSN. For example, if multiple records match on first name Soundex, last name, and DOB, but only one record has a valid hashed SSN, then all of those records will inherit the RIIPL ID corresponding to the valid hashed SSN. However, if those records instead match to several valid hashed SSNs, then a distinct RIIPL ID is assigned to each valid hashed SSN as well as to the remaining

[1] Prior to hashing, SSNs are validated using rules published by the Social Security Administration. A flag is stored in the database to indicate whether an SSN is valid, questionable, or invalid.

unmatched combinations of first name Soundex, last name, and DOB. Finally, records that are missing a valid hashed SSN *and* are also missing one of first name, last name, or DOB, are considered too ambiguous and are not assigned a RIIPL ID. This occurs for only 3.9% of records.

The RIIPL ID is deterministic, and is designed to minimize false matches (incorrectly identifying two different individuals as being the same person) at the expense of having more missed-matches (in which two records of the same person are not matched). As an alternative to this deterministic approach, RIIPL IDs could be constructed with probabilistic record-linkage methods that would likely have fewer missed-matches, but would also carry higher costs for computation and manual curation, as well as a higher likelihood of false-matches [8]. Evaluating probabilistic approaches and including time-varying PII such as residential addresses are areas of future research.

Integrating the data

We receive data extracts from administrative systems in various formats, but most typically as comma-separated text (with varying delimiters and quoting conventions), fixed-width text, XML, or Excel files. A key principle of our approach has been to meet government data partners where they are, and to accommodate data extracts in the format they can most easily produce. Most agencies have perpetual operational demands on their administrative systems, and they are not resourced to support additional development for data warehousing or analytics. In the best case scenario, we can identify an existing extract that the agency runs for an operational purpose, and arrange to receive a copy of that extract to avoid additional development and validation work for the agency.

Since there is no universal format or data dictionary across agencies, we normalize the data into a consistent format and typing structure with a lightweight integration tool developed in Python at RIIPL. Although we initially investigated enterprise tools in the business intelligence space for data modeling and integration, the up-front cost of those solutions led us to instead develop a custom integration framework, called Secure Infrastructure for Research with Administrative Data (SIRAD),[2] using an agile approach. The core functionality of SIRAD is the ability to hash sensitive identification numbers and separate PII at load time, and we designed it around this from the start, while incorporating additional features through agile sprints. Additional features, such as the RIIPL ID, were identified, specified, and validated through a tight feedback loop between the data integration team building SIRAD and the researchers using successive versions of the RI 360 database. The rapid development of SIRAD and the RI 360 database was facilitated by housing both the data integration team and the research team in the same lab.

SIRAD uses a simple layout file for each incoming table to describe the metadata for each of its columns. The layout file describes the original column name, the type (e.g. date, string, or numeric), the date format (if applicable), a flag for whether a column is a sensitive numeric

[2] https://github.com/riipl-org/sirad-example

identifier that needs to be automatically hashed at load time, and another flag indicating whether the column contains personally identifiable information (PII) and the standardized name of the PII (such as first name, last name, or DOB). These files are version controlled in git to retain the full history of loads and transformations.

Extract Load Transform (ELT) approach

SIRAD uses an Extract Load Transform (ELT) rather than a more typical Extract Transform Load (ETL) approach [5]. In practice, this means that the de-identified data are loaded into the RI 360 database in as close to their original format as possible. The majority of transformations are added later after researchers have a chance to perform preliminary analyses to assess data quality and understand the data-generating processes underlying the administrative systems. Often, this is an iterative process that involves our government data partners, and can both confirm data quality issues that the partners are already aware of, as well as identify newly discovered issues that may require further investigation or improvements to the administrative system.

Loading the original data without applying extensive transformations has several benefits when using administrative data for research. First, it retains the provenance of data and the values in the RI 360 database can be assumed to be the original information from the administrative system, not a derivative value created by a transformation process that may not be readily available to the researcher viewing the data. Second, transformations do not need to be defined upfront, which can be both time consuming and rigid, especially as research needs can evolve and change rapidly. The data are minimally transformed and made available to researchers within a short period of time. Finally, this approach is flexible for researchers; transformations can be created, changed or dropped within the RI 360 database without requiring interventions by the data integration team, or time-consuming reprocessing steps.

Optimizing the data for research

After the data are loaded, we employ several “data enabling” [3] processes to optimize them for use by researchers. The first of these processes is creating a set of “derived tables” which perform analysis on the administrative data in its original form, and transform it into derivative data and statistics often needed in research and analysis.

As an example, researchers at RIIPL are currently studying the factors that predict a child’s performance on standardized tests in reading and math in third grade. These factors can be used to point policy makers towards new policy levers that could help raise proficiency on the test, supporting the government’s goal of helping all children meet their potential[3]. The data needed for this project are spread across dozens of tables in inconsistent formats that cannot easily be joined together. One hypothesis in the study is that the timing or duration of events in a parent’s life, such as unemployment, disability, or incarceration could have a predictable effect on a child’s test scores which could be abated through additional outreach or support program

[3] http://www.kids.ri.gov/documents/Third%20Grade%20Reading%20Goal.pdf

targeting. Testing this hypothesis requires first determining when a parent is unemployed or incarcerated. Although this sounds like a simple query, administrative data do not always readily provide the required program participation indicator. For example, the incarceration system and its administrative backend have complex relationships between charges, sentences, incarceration, and parole. Determining from these databases, which are designed for prison operation, when an anonymized individual is incarcerated or released requires a complex transformation.

To perform this transformation and simplify the Department of Corrections data for use by an approved researcher, we create a *derived table* for incarceration spells, which lists all year-months in which an anonymized individual is incarcerated with a summary of the charges and sentences related to that incarceration. A derived table is a materialized view that aggregates, normalizes, and/or combines data from multiple original tables in the RI 360 database into a single table that facilitates a specific analysis need – in this case, determining if a parent is incarcerated or released at one, six, or 12 months prior to a child's taking a standardize test assessment in third grade.

Derived tables are typically indexed by individual (e.g. RIIPL ID) and by time (e.g. at a monthly granularity). We prioritize the development and validation of derived tables according to the priority of research projects. When starting a research project with a new data source, researchers perform exploratory data analysis of the original data to identify data quality issues. This process frequently involves government partners with deep knowledge of the operational uses of the data. Once a draft of the derived table has been constructed following the exploratory data analysis, we perform a peer code review to further validate the logic before officially adding the derived table to the RI 360 database. Peer review is requested and recorded using software development tools hosted within the data enclave.

Cross agency joins and the RI 360 Summary Table

To test how family incarceration events impact test score outcomes, we need to join Department of Corrections data on parents' incarceration spells and Department of Education data on children's test scores with Department of Human Services data on family structure. Like with incarcerations spells, the test scores and family structure have to be derived and transformed from multiple underlying tables in raw administrative program data. Using the RIIPL ID, we can join records anonymously across these derived tables. Ultimately, we are able to construct a single view with an observation for each child providing a primary male and female adult household member (which may be a parent, grandparent, or other relative) along with their incarceration history prior to the test while preserving the anonymity of all records.

To aid researchers with these kind of cross-agency joins, we roll up all of the derived tables into a single RI 360 summary table, which spans 20 years of history for the state's most important programs and outcomes, as well as demographic information about anonymized individuals (e.g. age, race, ethnicity, and gender). This table is a single source of information that on its own can provide the majority of variables needed for a new research project, exploratory analysis, or visualization. Like the derived tables, it is indexed on RIIPL ID and year-month.

Creating this derived table ensures that all research in the lab draws from common variable construction and definition, which is robust and reproducible, and that any changes based in new information are automatically pulled into all analyses in the lab.

Geocoding and anonymized neighborhood blocks

Another geospatial hypothesis in our study of third grade test scores is that characteristics of the neighborhood of the school or of the child's home may predict test scores. Because precise geographic coordinates of an individual's home could be used to identify an individual, we incorporate geographic information into the RI 360 database at an aggregated level, using the US Census Bureau's definitions of tracts, block groups, and blocks. A block can be geographically small enough to include a small number of individuals, so we salt and hash the block identifiers and provide a lookup table for joining Census demographics to the anonymized block identifier (Figure 1e). Therefore, a researcher can lookup block-level demographics for an individual's neighborhood without knowing the specific location of the block that the individual lives in.

Supporting research integrity

A fundamental requirement of scientific findings is that they can be independently replicated by other investigators [12]. Similarly, fact-based policy should be based on robust findings that are peer-reviewed and replicable. To facilitate future replication, we update and snapshot the RI 360 database approximately three times a year, creating what we call a *research version* (Figure 1f). The research versions are de-identified data and become the permanent archive of the RI 360 database. We have currently generated 11 such versions. Once a research version is validated by the lab, the encrypted original data used to create that version are destroyed, by shredding and wiping them from the encrypted tank. Every analysis is tied to a fixed research version of the database, and can be rerun against the research version at a later time to replicate the results. Additionally, to encourage reproducibility, analysis projects use a common project template to organize code and research results in a standardized way.[4] This template utilizes a software construction tool called SCons[5] to execute the full analysis pipeline for a project.

Since the documentation we receive from agencies varies considerably, we have built an automated system to generate codebooks in a standardized format for every table in a RI 360 research version. A codebook is a reference document that describes each column in a table along with summary information about the type and range of values and the proportion of missing values. Codebooks help researchers identify data quality issues, and enable faster onboarding of new team members and better knowledge exchange when a researcher joins a project.

[4] https://github.com/riipl-org/predictive-template

[5] http://www.scons.org

Controls

Even through the RI 360 database has been de-identified, our data sharing agreements restrict all research with anonymized individual-level records to the data enclave. Only aggregated or statistical results such as summary tables, plots, and regression coefficients can be exported from the enclave. All statistics must be aggregated such that they represent 11 or more distinct individuals. To ensure compliance with usage agreements and security, RIIPL has implemented two primary controls.

First, no individual researcher has the ability to export files from the enclave. Copy and paste functionality has also been disabled within the enclave's user interface. Second, exports are subject to review and documentation to ensure that exported results conform to usage agreements (Figure 1g). Researchers create export requests in the lab's task management system, generating a task ID for the export. The request explains the files to be exported, linking them to a chartered project and deliverable. A senior lab member reviews the files listed in the request to ensure all requirements for aggregation are met. The approved files are tagged with the task ID and moved to an export server within the data enclave, which generates a real-time alert to senior leadership and allows the files to be downloaded from the enclave. A read-only snapshot of the exported files is archived in the enclave to facilitate future audits.

Conclusion

The insights gained from research with administrative data have the potential to transform the way that policymakers approach some of society's most important policy decisions. Robust evidence on previous policy outcomes and predictive modeling of future outcomes can guide policymakers to smarter policies with greater benefits at lower cost. We have described a comprehensive approach to overcoming the many challenges faced when integrating siloed state-wide databases into a research database, which is the first system of its kinds in the U.S. In the future, we hope that more systems of this kind will provide policymakers at all levels of government with a rich ecosystem and evidence base for the important decisions they make on behalf of the American people.

References

1. Raj Chetty, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler. 2016. The Association Between Income and Life Expectancy in the United States, 2001-2014. *JAMA* 315, 16 (April 2016), 1750–1766. DOI:https://doi.org/10.1001/jama.2016.4226
2. Commission on Evidence-Based Policymaking. 2017. The Promise of Evidence-Based Policymaking. Retrieved from https://www.cep.gov/cep-final-report.html.
3. Roxanne Connelly, Christopher J. Playford, Vernon Gayle, and Chris Dibben. 2016. The role of administrative data in the big data revolution in social science research. *Social Science Research* 59, (September 2016), 1–12. DOI:https://doi.org/10.1016/j.ssresearch.2016.04.015

4. Jonathan M.V. Davis and Sara Heller. 2017. *Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs*. National Bureau of Economic Research, Cambridge, MA. DOI:https://doi.org/10.3386/w23443

5. Umeshwar Dayal, Malu Castellanos, Alkis Simitsis, and Kevin Wilkinson. 2009. Data integration flows for business intelligence. 1. DOI:https://doi.org/10.1145/1516360.1516362

6. Liran Einav and Jonathan Levin. 2014. Economics in the age of big data. *Science* 346, 6210 (2014), 1243089.

7. P. Gauravaram. 2012. Security Analysis of salt||password Hashes. In *2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*, 25–30. DOI:https://doi.org/10.1109/ACSAT.2012.49

8. Katie Harron, Chris Dibben, James Boyd, Anders Hjern, Mahmoud Azimaee, Mauricio L Barreto, and Harvey Goldstein. 2017. Challenges in administrative data linkage for research. *Big Data & Society* 4, 2 (December 2017), 2053951717745678. DOI:https://doi.org/10.1177/2053951717745678

9. Justine S. Hastings and Jesse M Shapiro. 2017. How Are SNAP Benefits Spent? Working Paper No. 23112. National Bureau of Economic Research, Cambridge, MA. Retrieved from http://www.nber.org/papers/w23112.

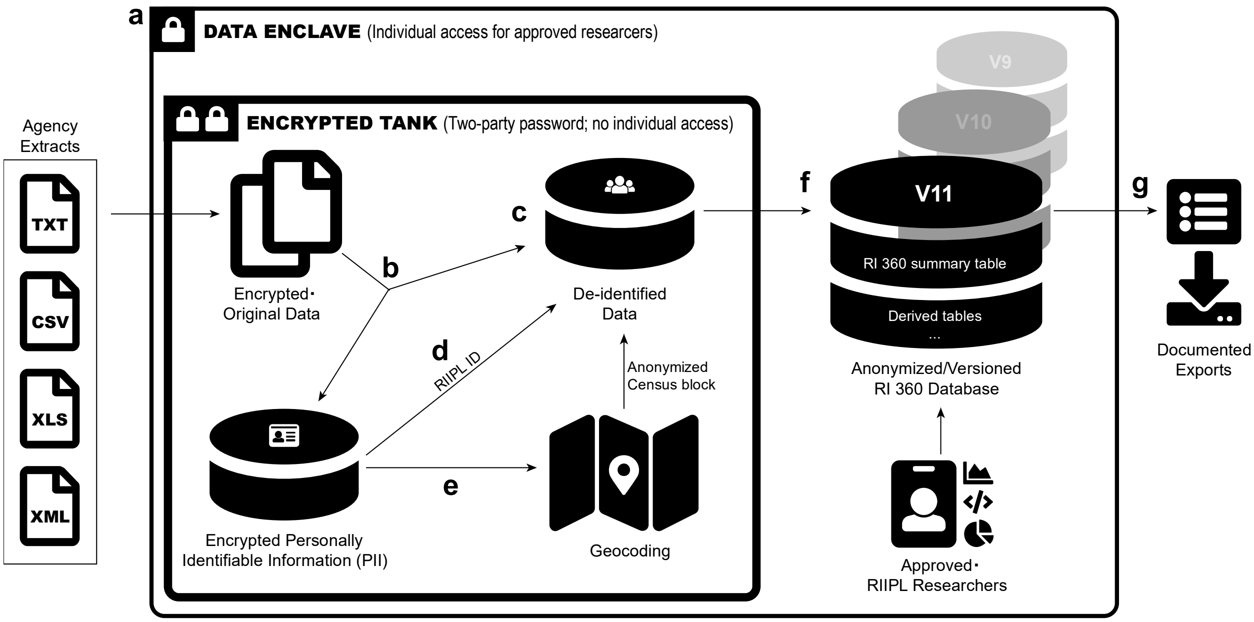
10. Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human Decisions and Machine Predictions. *Q J Econ* 133, 1 (February 2018), 237–293. DOI:https://doi.org/10.1093/qje/qjx032

11. Julia Lane and Stephanie Shipp. 2007. Using a Remote Access Data Enclave for Data Dissemination. *International Journal of Digital Curation* 2, 1 (2007), 128–134. DOI:https://doi.org/10.2218/ijdc.v2i1.20

12. Roger D. Peng. 2011. Reproducible Research in Computational Science. *Science* 334, 6060 (December 2011), 1226–1227. DOI:https://doi.org/10.1126/science.1213847

13. C. Russell Robert. 1918. The Soundex coding system. Patent No. US1261167.

Figure 1. Overview of the processing steps to secure, integrate, and conduct anonymized research with administrative data. Agencies securely transfer data extracts to an encrypted tank inside the data enclave **(a)**. These data are split **(b)** into personally identifiable information (PII) and de-identified data **(c)**. PII are used to construct an anonymized global identified called the RIIPL ID **(d)**. PII are also used to geocode home addresses to construct an anonymized neighborhood identifier **(e)**. De-identified data are used to construct research versions of the RI 360 database **(f)**, which can be accessed by approved RIIPL researchers from inside the data enclave. Research findings can be exported from the data enclave through a documented review process **(g)**.


a
DATA ENCLAVE (Individual access for approved researcers)
Agency Extracts
TXT
CSV
XLS
XML
ENCRYPTED TANK (Two-party password; no individual access)
Encrypted Original Data
b
c
De-identified Data
d
RIIPL ID
Anonymized Census block
Encrypted Personally Identifiable Information (PII)
e
Geocoding
f
V9
V10
V11
RI 360 summary table
Derived tables
...
Anonymized/Versioned RI 360 Database
Approved RIIPL Researchers
g
Documented Exports