Data Quality and Standards

ADRF Network Working Group Participants

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

Recommended Citation
ADRF Network Working Group Participants, "Data Quality and Standards" (2018). ADRF Network Working Group Reports. 3. https://repository.upenn.edu/admindata_reports/3

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/admindata_reports/3
For more information, please contact repository@pobox.upenn.edu.
Data Quality and Standards

Description
The Data Quality and Standards Working Group determined where current administrative data quality standards exist and where additional guidance are needed. The group used a hypothetical example to illustrate how improved data quality can make administrative data research better.

Chair and Lead Author: Amy O’Hara (Stanford University)

Creative Commons License
This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License.
Data Quality and Standards

By Working Group Participants
Acknowledgments

The following people served as Working Group participants. They contributed to group discussions, shared important resources, generated valuable insights, and helped with writing and editing this report.

- Amy O’Hara, Stanford University (chair and lead author)
- Katie Barghaus, Penn Child Research Center
- Steve Cook, University of Wisconsin
- Rashida Dorsey, U.S. Health and Human Services
- Dylan Ellis, Camden School District
- Bhargavi Ganesh, Urban Institute
- Amy Hawn Nelson, Actionable Intelligence for Social Policy
- Monica King, ADRF Network
- Kimberly Korejko, Delaware Valley Regional Planning Commission
- Maggie Levenstein, ICPSR
- Kathryn McNamara, U.S. Census Bureau
- Rohit Naimpally, J-PAL North America
- Shawna Waugh, U.S. Energy Information Administration
- Jun Zhu, Urban Institute

We also thank the Alfred P. Sloan Foundation for making this work possible.
## Contents

Acknowledgments ............................................................................................................. 1  
General Findings ............................................................................................................... 3  
Motivation: A Hypothetical Example .................................................................................. 3  
How could this be different? ............................................................................................ 5  
Recommendations and Conclusion ................................................................................... 9  
References .......................................................................................................................... 12  
Appendix 1: Viewing the six issues through private sector data ..................................... 13
General Findings

The Working Group determined where data quality standards exist and where additional guidance can enhance operational and statistical uses of administrative data. We considered incentives that are needed to encourage data producers and analysts to adopt standards and follow guidelines. We considered how standards may differ across data from local, state, and federal government agencies and from the private sector.¹

We observed that critical information is missing at data origin. There are seldom standards or documentation where data are generated. There is little consistency in catalogs or inventories of administrative data. There are no metadata standards uniformly applied, nor information on data transformations or provenance measures. The lack of information about individual datasets hampers linkages across datasets.

We noted difficulties faced by different groups, analysts versus data generating agencies versus policy makers. Analysts face access problems. Once data is accessed, researchers lack repositories to find information on data standards, documentation, and code. Some repositories exist but are not known across user groups and lack common tags to hasten discovery. Generating agencies have concerns about security, whether they can trust analysts and their security measures, and whether data use will spur public or political backlash. Generating agencies face resource constraints. Policymakers and, in many cases, researchers lack understanding of which questions can be answered with available data based on data quality and availability.

Motivation: A Hypothetical Example

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. Basic data quality and documentation issues could pose obstacles to understanding the drivers of low matriculation rates. Administrative data on test scores and advanced placement courses may have gaps or outright errors, while the absence of demographic data would preclude an understanding of how

¹ We considered open data and federal enforcement and surveillance uses out of scope.
the issue breaks down for certain populations. Basic data infrastructure would need to be in place for the district to understand why students enroll in college at lower rates, and what programs might effectively combat that.

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. Data for each of these areas would be needed to assess whether the program achieved its intended goals; however these data would likely be siloed.

Of course, monitoring or evaluating student outcomes will be easier if all of the necessary data is gathered in an IDS (Fantuzzo et al., 2017). But regardless of whether data is in an IDS or is gathered from multiple data providers and then matched for this specific project, the usefulness of this data will depend on the quality of the data and the documentation available.

Data that have been entered by caseworkers or school employees may have gaps for certain populations, or outright errors. Such gaps and errors are more common in fields that are considered inconsequential for program administration. In our example, high school workers may collect family income data for promise grant applicants, but not bother to record this information for students who do not qualify due to academic shortcomings. Even for those students whose family income data is collected, there may be significant reporting error in these numbers since the data collectors have no way to verify, and income does not determine grant eligibility. On the other hand, those students may have very high-quality income data recorded when they apply for additional financial aid at their chosen colleges, since this data is verified by tax records, and is an important determinant of eligibility for other assistance. These differences in quality for different data sources may be well known to program staff, but unless this
information is documented the gaps and errors would not be obvious to the evaluation team.

Good documentation should indicate not only how data fields are coded and what they mean, but also what is known about the data collection process and how that may affect data reliability. This will allow program monitors and evaluators to prioritize the use of data fields from higher quality sources (here, one would prefer income data from financial aid applications over those collected by the high schools), in addition to appropriately using and interpreting that data.

If data is being put together on an ad hoc basis for this particular program then these data quality issues will need to be dealt with on an ad hoc basis as well: known data quality issues will need to be collected from data providers; documentation will need to be reviewed for accuracy and completeness, quality assessments may need to be conducted. The advantage of a pre-constructed integrated data system (IDS) is that these steps can be taken in advance, assuring that high-quality, documented data is available when needed, and the costs of this data quality and documentation work can be amortized across all the projects that access the IDS (in addition to the costs of establishing data use agreements, data collection, matching, data security, etc.).

Assume that the scholarships proceed, but access to data on the students’ earnings, program participation, and justice involvement falters. Perhaps the family income and program participation are known through the scholarship assignment process. Perhaps after a decade, the subjects’ state unemployment insurance wage data are evaluated to measure long-term impacts of the intervention, but that data source only includes subjects who remained in-state and in covered employment. The initial broad goal of observing current and future impacts of the scholarships remains unmet.

How could this be different?

How could better data quality, standards or guidelines make the project easier and the evaluation more effective in identifying good policy? Our working group identified six topics in need of standards or guidelines. Below, we describe each topic and indicate how addressing each topic would have helped our hypothetical scholarship pilot project.

1) **Metadata.** Metadata are the technical information that describes and explains a dataset
and makes it easier for users to access and use. Metadata standards already exist. There are multiple approaches that reflect best practices (Green & Humphrey 2013). For example, the Data Documentation Initiative (DDI) has been adopted by many data intermediaries. However, metadata standards have not been adopted by many data generators. Our example above would have benefitted if metadata were available for the data sources needed in the evaluation. Analysts would have known which data were documented, which were not (and would likely be difficult or time consuming to work with). This would have improved planning for the evaluation, including what outcomes could be measured and, consequently, what questions are answerable.

2) **Data quality dimensions.** Data quality challenges such as data entry error, missing data, and duplicate records are common. There are no recognized technical standards for administrative data, but guidelines do exist. Numerous frameworks have been published on determining data quality dimensions. Looking abroad, we find Eurostat identifying six quality indicators—relevance, accuracy, timeliness and punctuality, accessibility and clarity, comparability, and coherence—in assessing data quality (Bergdahl et al., 2007). In the United States, Iwig et al. (2013) developed a similar framework that include the following dimensions—relevance, accessibility, coherence, interpretability, accuracy, and institutional environment. The recent report from the National Academies of Sciences, Engineering, and Medicine Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods suggests that these dimensions, especially timeliness and granularity, should be part of a federal agencies’ broader framework for evaluation administrative data quality. These dimensions will help capture the benefits of administrative data often reflected in their timeliness and granularity, but also its challenges with linkage errors and less structured data types (National Academies of Sciences, Engineering, and Medicine, 2017).

Another framework for thinking about issues of data quality is defined by the concepts of internal and external validity. We can think of internal validity as capturing whether your data makes sense and can be used to generate the measures you need (including the necessary metadata and documentation needed to use the data consistently), while external validity reflects how accurate your data is to individual’s
and program’s real situation.

In our scholarship pilot example, data quality information would have revealed which sources were fit for the stated use. For example, when the pilot evaluates long term employment and earnings outcomes, a study of data quality dimensions would reveal coverage issues such as state Unemployment Insurance (UI) data missing study participants who move out of state and missing information on contractors and the self-employed.

3) **Data Wrangling.** By data wrangling, we refer to data transfer, cleaning, standardization, and linkage activities. There are no standards, but guidelines do exist. Data cleaning may include imputing missing data and harmonizing sources. We recommend a process standard, informing analysts of issues to watch out for but not specifying solutions to those issues. For data standardization, there are best practices in certain domains. For example, the USPS has postal addressing standards, and there are multiple recognized approaches for parsing person name data. We recommend a standard process and compilation of resources to help users determine how to handle their data. Again, for data linkage, no standards exist. However, University College London releases the GUidence for Information on Linked Datasets (GUILD) to help researchers review and document issues surrounding data linkage (Gilbert et al., 2018).

These data processing standards would have benefitted our hypothetical scholarship example, identifying the issues that the researchers should consider, and referrals to methods and code to help with cleaning and standardization.

4) **Retaining Data.** Existing data repositories such as the Inter-University Consortium for Political and Social Research (ICPSR) and Dataverse have processes for archiving social science research and data. According to ICPSR, data archiving is a process that begins even before the project starts with important milestones to be met throughout the data lifecycle that ends with depositing the data (Inter-University Consortium for Political and Social Research, 2012). In the scholarship example where restricted data are linked to produce a research dataset, researchers should aim to archive the linked data with permission from the data producers. These data could then be accessed by other trusted
researchers in a data enclave or under other restricted-use requirements. Proper archiving of these data and project would support research replication, ensure longevity of the data, and promote their longer-term visibility and impact while ensuring protection for the research subjects.

5) **Retaining code and documentation.** Proper code documentation and retention are crucial for research transparency and replication. Best practices for code documentation have emerged in the data science world across a variety of uses and programming languages, but there are relatively few published guides for researchers. A recent guide from Innovations for Poverty Action (IPA) describes how to adopt best practices in coding and code documentation in the context of research projects (Chuang et al., 2015). The IPA guide first recommends creating a living ReadMe document that explains key code and data files of the project and variable interpretations. The guide also encourages research teams to utilize existing web platforms such as GitHub for collaborating on code writing/revision and storing data documentation. Adopting GitHub and its version control and file tagging features could go a long way in ensuring researchers are retaining code properly in the future.

Proper documentation, retention, and archiving is also essential to aid replication and allow similar projects to understand methods and assumptions in a project. Alongside preregistration plans, we recommend repositories or catalogs of codebooks, data use agreements, data quality assessments, and sensitivity analyses.

Proper code documentation would have benefitted the scholarship project, making joins across school files or human services datasets more efficient. Code retention at the time of a pilot would benefit the outcome study conducted a decade later, providing insights and methods to align or correct issues with the later research team. Proper documentation retention would also benefit the project, especially if data use agreements and quality assessments were known at the planning stage.

6) **Retaining Results.** We recommend that researchers seek information on projects that have used administrative data to identify sources that may be fit for their use. While researchers will encounter some of this information when conducting a literature review,
there are many analyses outside of peer review journals using administrative datasets, and there may be multiple data sources for a needed concept (e.g., earnings from self-report, IRS, or state UI wage files). It is useful to know where data access has been permitted, for which studies, and the results of such studies.

The American Evaluation Association (AEA) recommends that results of research and evaluation should be disseminated to both the public and the relevant government agencies in a timely fashion (American Evaluation Association, 2013). Following dissemination, the AEA also suggests tracking the uses of these findings to understand whether results have an impact on long term policymaking. Along a similar vein, researchers and data archivists can help make data discoverable (Vardigan, 2012). Researchers should engage in data citation practices, which are currently encouraged by a small number of peer-reviewed journals. Data repositories and intermediaries should also create databases that link data to publications that use the data.

Understanding which data sources had been used in research and evaluation projects would have benefitted our scholarship example by showing the research team which projects have harmonized human services and justice data or have used state longitudinal data system student data in previous studies. Seeing the results and limitations of the various sources would help the project team with the research plan.

**Recommendations and Conclusion**

The working group is not striving to make or impose “standards” for administrative data uses and users. We propose identifying existing resources and approaches, then sharing that information with multiple communities, including evaluators, researchers, and administrators. We eagerly anticipate the report on data quality standards from the Federal Committee on Statistical Methodology (FCSM). FCSM has been instrumental in setting standards and guidelines for surveys and statistical programs. The ADRF Network can support the testing and implementation of federal data quality standards as we explore best practices for other data sources.

We feel that the topics identified above should not be viewed as by-products of projects and operations. They are critical to the success and quality of analyses using administrative
data. Overall, we recommend that the ADRF Network advocate use of guidelines and process standards that improve administrative data quality. This should involve outreach to professional organizations (including organizations that support data generators such as county, state, and federal agencies), the statistical community, journal editors, and funders with the objective of establishing norms for administrative data researchers. When they ask how to meet that objective, we must be prepared with a long-run vision and specific short-run recommendations.

The long-run vision: Data and metadata will be linked and available to researchers. Data inventories will exist with adequate metadata aiding data discovery. Datasets will have digital object identifiers and citations, which are linked to other repositories containing data use agreement templates/precedents and results of previous studies using that dataset. The code and documentation (including data quality measures and checklists) from previous studies will be connected to facilitate searching in code repositories. These linkages will expedite proposal development and review and data access, and replicability.

How do we get there?

1. We recommend that the ADRF Network build on existing IDS efforts to incorporate more data, code, and results retention by providing a roadmap and technical assistance.

2. We recommend that ADRF Network members with expertise in metadata and working with state and local agencies develop quick-start guides to capturing metadata for jurisdictions lacking any documentation.

3. We recommend having ADRF Network intermediaries interacting with private sector firms to discuss documentation development—whether metadata created through external research projects is useful to the firm, and whether it can be shared to motivate additional studies.

4. We recommend that a group of ADRF Network intermediaries (1) test existing data quality checklists over the next year and report on their experiences, and (2) test application of FCSM data quality standards across a variety of data types.
5. We recommend that the ADRF Network survey the community to determine which data wrangling tools they are using, the level of effort they apply (given their workload processed), and their degree of satisfaction with the results.

6. We recommend that the ADRF Network establish guidelines and process standards for data management under our six topic areas where none exist.

7. Finally, we recommend that the ADRF Network review our issues and recommendations with the other working groups and we recommend that the ADRF Network conduct an assessment of how known intermediaries are addressing the issues that we identified. There is a tremendous opportunity to make research using administrative data more efficient and robust.
References


Appendix 1: Viewing the six issues through private sector data

We anticipate differences between government and private sector data in data quality and access issues. We will need to consider whether there are incentives for firms to document or improve data quality for research and evaluation, or if they will permit external researchers to develop documentation and benchmarks.

Consider a large financial services company with 150 million transactions per day. They manage data for operations with a high degree of efficiency and security. Their internal research group addresses problems in their industry and develops new technical approaches to advance the firm’s global market share in existing and new products. This firm has cautiously shared through industrial contracts with university researchers. Approved research team members use company laptops with data extracts (no data may be added or removed from the devices).

How do the six topics in need of standards or guidelines identified by our working group fit proprietary data?

**Metadata:** The firm did not offer a codebook or data dictionary to the university researchers. Their staff answer questions but care is taken not to burden the company researchers. The research assistants on the project develop documentation, intending to share materials back to the company. This is metadata for external research, since the company didn’t need it and may not devote resources to review or verify its accuracy. There is little incentive for the company to have comparable metadata with other firms in their sector, but this information is needed to facilitate external research.

**Data quality dimensions:** The firm has a lot of data on global consumers and businesses. Processing 13,000 transactions per second, the firm has already optimized data quality – for their purposes. From a researcher’s perspective, many quality dimensions used for government administrative data will be difficult to evaluate. Data relevance and timeliness will probably be endogenous to the research question, the data are sought to answer a question and cover an appropriate timeframe. As described above, metadata may be missing, so data clarity is lacking. A researcher may assume that the transactions data are accurate, but they may need to assess representativeness (whether the company’s data reflect true population values). This also affects
measures of coherence: the firm’s data reflect expenditures, debt, and sales, but analysis is needed to determine how the data complement other sources. Comparability may be measurable by viewing data stability across geography and over time, and benchmarks to others in the industry exist (especially when demanded by investors). However, accessibility may challenge both coherence and comparability if only a narrow or time-specific data snapshot is available or data use is limited to one research team.

**Data wrangling:** The company will have cleaned and recoded data for their purposes, which may not align with external research needs. Efforts will be needed by the university researchers, who may struggle to interpret and join data appropriately. In other private sector data projects, files may already be joined and cleaned before sharing. Researchers must trust that firms are sharing the full datasets, are transparent about edited and modeled values, and that any processing or data capture changes over time are shared with the researchers.

**Retaining data:** The company wants the laptops back when the analysis is complete. The external research team lacks the authority to save the data to an archive (at this time).

**Retaining code and documentation:** The external project team may retain code and documentation for the project, provided that contract terms allow.

**Results of research and evaluation projects:** External research teams publish with the data, and their findings join internal company research findings that are published in journals or technical papers. Replication may be challenging if access was based on relationships. It will also rely on archiving of project data and version control on data extracts.

With proprietary data, we may need to survey research teams who have had data access to explore these issues. The ADRF network will need to continue conversations with the firms to establish a framework and greater trust to use their data for research and evaluation.