Data Protection Toolkit's Playbook as Demonstrated through the PIAAC Prison Study

Tom Krenzke[1], Jianzhu Li[1]
[1]Westat, 1600 Research Boulevard, Rockville, MD 20850

Abstract

This article demonstrates the Data Protection Toolkit's Playbook, which is a general operational framework for protecting confidential data in Federal agencies. It includes a step-by-step "how to" guide for data disclosure analysis from data protection planning to data dissemination. The basic steps of the process are to 1) write a disclosure analysis plan (DAP), 2) implement the DAP, 3) conduct an impact assessment, 4) write a post-mitigation report, and 5) disseminate data. The Playbook does not contain standard operating procedures for agencies to follow, but rather, it is a guideline for agencies to consider for their own best practice. The data used for the demonstration comes from the Prison Study that was conducted in 2014 for the National Center for Education Statistics (NCES) as part of the Programme for the international Assessment of Adult Competencies (PIAAC). The PIAAC Prison Study included an in-person assessment of literacy, numeracy and problem solving skills for adult inmates 16 to 74 years old from eligible federal and state prisons in the United States. The main challenge is to balance the needs for data users with the need to comply with laws and standards related to confidentiality. To address the challenges and following NCES standards, a risk assessment was conducted, and two risk mitigation approaches were implemented: data coarsening, and random perturbation through controlled data swapping. The impact of the perturbation was evaluated

Key Words: Statistical disclosure control, establishment, risk assessment

1. Introduction

The Data Protection Toolkit (DPT) is Action #15 of the Federal Data Strategy. The toolkit will address the need to maintain confidentiality and data privacy when providing access to federal data assets. Once available, it can be used by agencies to develop and implement cost effective data protection programs. The web-based toolkit will provide a repository for recommended confidentiality and data privacy practices. It is intended to be a central resource for guidance tools and templates. There are over 220 resources from over 25 contributors which are mainly federal agencies with some international organizations and contributors. Also the text in the web based toolkit will be a refresh of the FCSM Working Paper 22.

The Playbook was developed for the DPT and it provides a general operational framework for protecting confidentiality in federal agencies. It's a step-by-step how-to guide for data disclosure analysis, from data protection planning to data dissemination. It's patterned closely to the standards and protocols followed by the National Center for Education Statistics. This article is partly a preview of the DPT's Playbook, which will come out when the DPT comes out, which is TBD. Then the Playbook is applied to the Program for the International Assessment of Adult Competencies Prison Study public use file creation.

Figure 1 provides a flow of the Playbook's operational step-by-step guide. It begins with the disclosure analysis plan template. The template is a refresh of a widely used disclosure avoidance checklist that has been used throughout the Federal government for communicating the risks in data

products. After the template has been completed and discussed, it can inform the writing of the disclosure analysis plan called the DAP. The DAP gets submitted for approval by the government agency and once approved the DAP is implemented. That means all the coarsening or variable suppression or perturbation that is outlined in the DAP gets implemented. Next the impact assessment is conducted, which performs some quality checks by comparing results before and after the confidentiality treatments were employed. Then after the impact assessment has been completed, the risk mitigation report is written. Before the data can be disseminated the risk mitigation report needs to be approved.

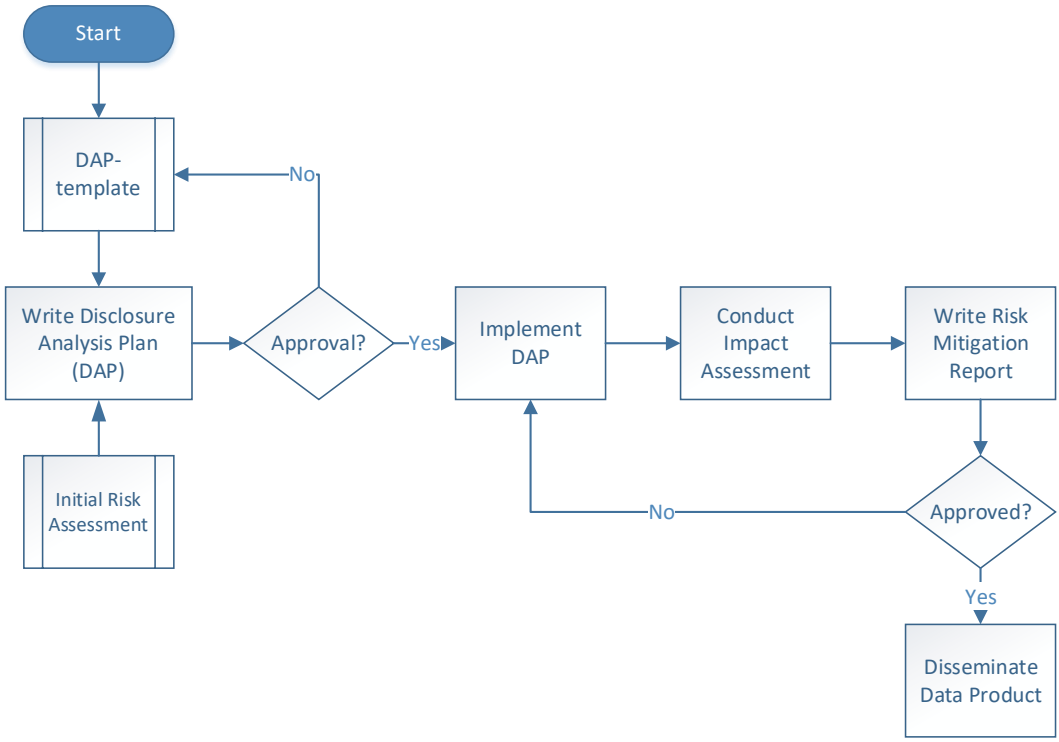

Figure 1. Flow of the Playbook's General Operational Framework for the Disclosure Avoidance Process

2. PIAAC Background

We retrospectively applied the Playbook as a step-by-step guide to the PIAAC Prison Study. PIAAC has two components: the household component and the prison component. The household component is part of an international assessment in literacy, numeracy and problem solving for adults 16 to 65 year old. It is sponsored by the Organization for Economic Co-operation Development. It includes a background questionnaire, and an assessment in adult skills. Cycle 1 included 38 countries that conducted data collection across three rounds conducted in 2012, 2014 and 2017. Cycle 2 is starting and currently includes 33 countries in the first round to conduct the data collection in 2022. The Cycle I Prison Study was conducted among 16 to 74 year olds in 2014 in the U.S. only, and was sponsored by NCES.

The Prison Study was comprised of a two stage stratified probability proportionate to size design. In the first stage, prisons were selected with an oversample of female-only prisons. The sample frame resulted from an update to the 2005 Prison Census using the 2012 American Correctional Association (ACA) directory. Stratification occurred by gender of the prison, Census Region, facility type, security level, and size of the prison. At the end of the data collection efforts, there

were 18 female-only prisons and 80 male or coed prisons that cooperated. In the next stage, inmates were selected from a list of all inmates occupying a bed the night before inmate sampling was conducted, except for prisons from the Federal Bureau of Prisons, which were based on rosters of inmates a week before the visit. Inmates were selected using a systematic random sample. At the end of the data collection period, there were 1,319 completed cases and an 82% overall response rate was achieved.

3. Disclosure Analysis Plan

As outlined in the Playbook, the disclosure avoidance process starts with completing the disclosure analysis plan-template (checklist), which informs the writing of the DAP. For the PIAAC Prison Study, the DAP includes discussions about the sample design and source data. Then it outlines the disclosure analysis that is needed to produce the public use file. This includes an initial risk analysis, and plans for data coarsening and data swapping, which is a standard at NCES. The DAP also mentions any other data products such as the International Data Explorer and the restricted use file for PIAAC. Next, the DAP is submitted to the NCES Disclosure Review Board for approval. For NCES, all data products must produce the same results. So the core file, which is the restricted use file, is perturbed, and all data products stem from that core file. The DAP also includes appendices. The appendices are comprised of a list of all variables and their treatments, a summary of the external source investigation (discussed below), and recodes that are proposed.

In addition to completing the DAP-template, an initial risk analysis is conducted to also inform the writing of the DAP. In the Prison Study initial risk analysis, we identified the major disclosure risk factors. For instance, identification number of the prison or inmate, which is a direct identifier, and other direct identifiers were removed. Prison-level and geographical variables were listed, and all were suppressed except for Census Region. Then variables relating to the sample design, weighting and variance estimation were identified, and all were suppressed except final weights, replicate weights, and variance estimation codes. The risk assessment also included an evaluation of external data, and a review of combinations of factual variables.

4. DAP Implementation

As outlined in the Playbook, after the DAP is approved and all the data are ready, it is time to implement the DAP. Figure 2 presents a flow of the DAP implementation process. First, external data sources are evaluated with consideration for probabilistic record linkage or exact matching if the external data are seen as a threat. Then using the initial survey file, the re-identification risk in the population is measured and the individual risk is measured as well, because some records are more risky than others. These risk measurements are useful for creating recodes and variable suppressions that are applied and then the risk assessments are repeated. If more risk reduction is needed then we go back through and recode more variables or suppress more variables.

It is a standard at NCES to perturb the data through controlled random swapping. Then an impact assessment is conducted and reviewed prior to disseminating the data. The risk assessment contains two main phases. In the first phase, the external sources are evaluated, and in the second phase, combinations of identifying variables are evaluated.

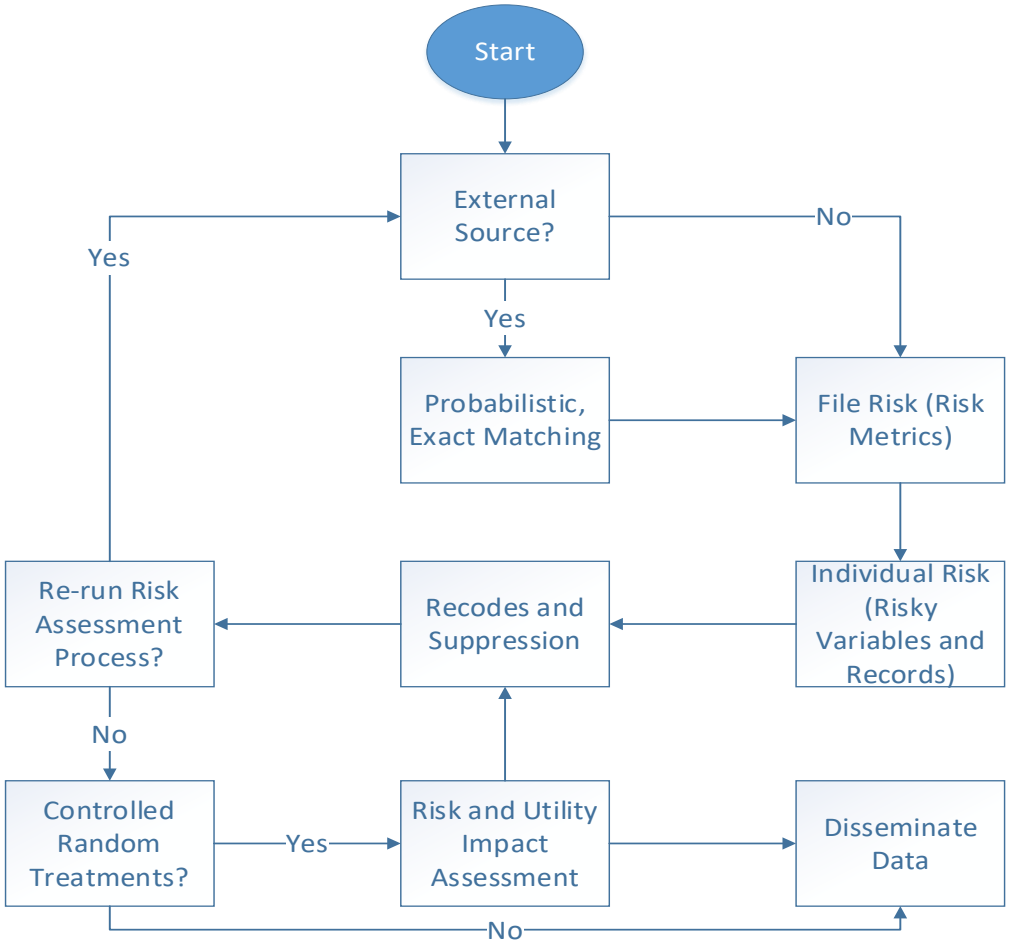

Figure 2. Flow of the DAP Implementation Process

To assess the risk of external sources, typically a match is conducted between the sample data and the population data. As shown in Figure 3, in the sample data, common variables between the sample and the population data are identified in green. The sample also contains responses to sensitive questions (yellow). In the publically available population data, the variables in common between the sample and the population data are identified and the population data contain PII (red). The sample and the population data are merged using the common variables as the matching key to bring the direct identifiers together with the responses to sensitive questions through the merging process. The goal is to assess the risk of an intruder successfully bringing the direct identifiers and sensitive data together.

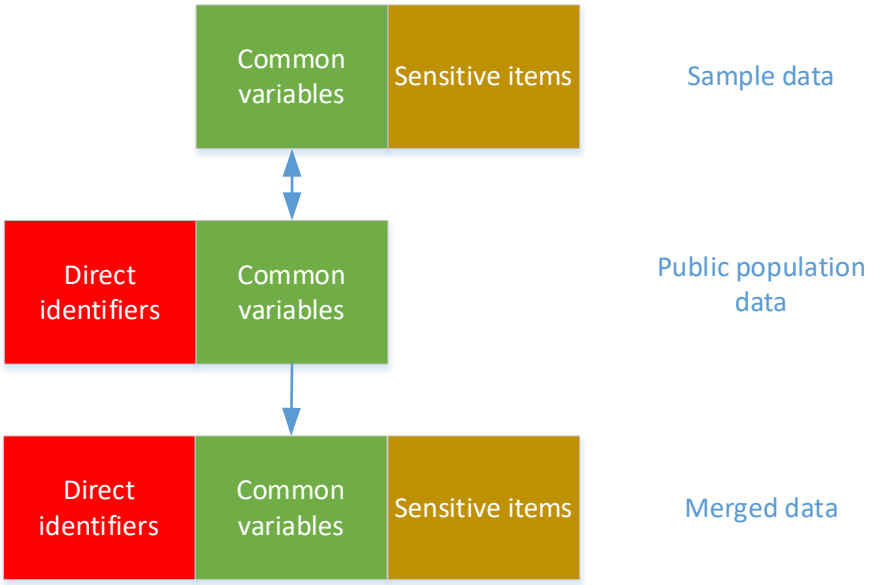

Figure 3. Illustration of Intruder Attempt of Matching To Population Data

For the PIAAC prison study, as mentioned above, the sample data will not have any prison-level data provided on the public use file except for Census Region, and the variance strata and variance units from the jackknife replication approach. In the jackknife approach, the first stage units are the prison, therefore the combination of the variance strata and variance unit provides a unique ID for clusters of inmates that mostly aligned with a prison.

For the Prison Study, the related population data is very limited and not intruder friendly. The 2012 American Correctional Association manual was used to update the 2005 prison census. The manual is very thick and is only available via hard copy. Furthermore, the definition of the sampling frame is hard to replicate. For example, the ACA includes all units within a prison as separate entities, such as annexes, work camps, medical wards. The most likely threat from external sources may be to match information for clusters to the population data toward the re-identification of prisons. To do this, we can roll-up the PIAAC inmate survey responses to the variance cluster level, which essentially would be the prison level. This process results in estimated percentages by race/ethnicity and gender, for example. And then probability-based record linkage could be used to identify the most likely prison to match those rolled up estimates and calculate the percentage of correct matches. One thing to note here is that those percentages would be unstable because of the low number of inmates (average of 13.5 completes per prison) contributing to those estimates. That all being said, it was agreed to not conduct this process due to the instability of the estimates, and reasons stated above about the population data's inaccessibility to an intruder.

After assessing the risk due to external data sources, the risk from combinations of identifying variables was evaluated. First, the global or file risk was quantified, and then the relative risk among the records was estimated. For the global risk, we chose indirect identifiers and used a log linear modeling approach to estimate the re-identification risk (Skinner and Shlomo, 2008). Then using the log linear modeling approach, we estimated the proportion of the sample that are population uniques. This can now be done using the R package SDCNway[1], which was developed as part of the DPT. The estimated proportion that are population uniques is a measure of the probability of re-identification. A determination is made from the estimate as to whether to do more variable suppression or data coarsening, or to proceed with the next step.

[1] Available at: https://CRAN.R-project.org/package=SDCNway

Next, we measured the relative risk among records and categories of variables. First, we conducted all one-way tabulations and identified categories with less than 25 observations, which were subject to recoding or variable suppression. Next, we conducted all two-way and three-way tabulations among the indirect identifiers. While doing this, violations of the "Rule of 3" were tallied for each record, and for each category of each variable. This approach was done using NCES software *InitialRisk*, for which the method can also be found in the aforementioned R package SDCNway.

5. Risk Mitigation

The risk assessments helped to inform the risk mitigation, that is, the confidentiality treatments. For example, categories causing the most violations were subject to recoding or variable suppression. Also following NCES standards, data swapping was done using the software *DataSwap* (Kaufman, Seastrom and Roey, 2005)[2]. Using the software, very high risk records were targeted for swapping, and controlled random swapping was conducted on the remaining records. Also data records with more violations in the exhaustive tabulations were given a higher chance of being swapped. The swapping partners were carefully chosen within classes formed by variables related to their proficiency levels.

After confidentiality treatments were completed, the impact assessment occurred. The software *DataSwap* contains the following impact assessment measures (Dohrmann et al, 2009), which compare the original data with the swapped data.

- Hellinger's distance, which is used to compare results from contingency tables,
- Weighted cell counts and weighted cell means
- Measures of associations, including Cramer's V, Pearson's product correlation, Pearson's contingency coefficient, and regression coefficients

Five separate runs were processed, and reviewers selected the best run based on the metrics above.

5. Summary

After the risk mitigation and impact assessment were completed, it was time to write the risk mitigation report. The report goes into details about the swapping parameters for the prison study, while emphasizing the changes that occurred from the DAP and the reasons why the changes happened. The report also provides the *DataSwap* output results. Justification is given for the selected run.

Once the risk mitigation report is approved, a safe-to-release memo is prepared and approved and the data product is disseminated. The PIAAC public use file can be accessed here https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2016337REV. As guided by NCES standards, it is important to note that the swapping was conducted on the restricted use file, then the public use file was derived by implementing the variable suppressions and the recoding on the perturbed restricted use file. All data products are ultimately derived from the restricted use file.

In summary, the DPT will contain the following tools: the DAP template, the Playbook, the SDCNway R package, *DataSwap*, and a safe-to-release memo template. This case study also demonstrated the NCES standards for maintaining confidentiality, and the DPT's Playbook.

[2] Available at: https://github.com/Westat-Stats/DataSwap.

References

Dohrmann, S., Krenzke, T. Roey, S., and Russell, N. (2009). Evaluating the impact of data swapping using global utility measures. Proceedings of the Federal Committee on Statistical Methodology Research Conference.

Kaufman, S., Seastrom, M., and Roey, S. (2005). Do disclosure controls to protect confidentiality degrade the quality of the data? Proceedings of the Joint Statistical Meetings, American Statistical Association Section on Government Statistics. Alexandria, VA: American Statistical Association.

Skinner, C. J. and N. Shlomo (2008). Assessing identification risk in survey microdata using log-linear models. Journal of American Statistical Association. 103, no. 483 (2008): 989–1001.