Data Protection Toolkit (DPT) Resource Collection

Attachment 2
Data Collection Instrument

OMB #1850-0803 v.261

Submitted by
National Center for Education Statistics (NCES)
U.S. Department of Education

February 2020

Contents

INSTRUCTIONS

In the Data Protection Toolkit Request Form for Existing Resources below, provide links (URLs) or examples (documents) of tools, manuals, checklists, templates, etc. that you or your organization use or recommend for assessing, managing, and mitigating the risk that individuals or enterprises are re-identified from the release of confidential data.

The URLs or documents you provide will be used in a website, so the material must be able to be shared with the public; the material is what you use and can be provided in the public realm.

A Reference Guide is included below to help remind you of the types of information that are being requested. Please briefly scan this guide to ensure that your responses are as informative as possible and that you don't miss reporting something on the Request Form.

Please refer to the Glossary of Terms document if any concept or term is unclear.

Within the form, please provide a link to the existing resource along with a brief one-sentence description. As an alternative, enter the file name of any attachments that you may provide in response. Lastly, indicate for each resource if it is currently in active use. If you have more than three URLs/files for any resource type, just add additional lines as necessary.

If any materials apply to more than one pre-defined category, please list the material in as many categories as applicable. Additionally, if there are any other materials or thoughts you want to share related to data protection that might not fit into one of the pre-defined categories, feel free to include it in 'other'.

Please complete and return the information form to Michael Hawes (michael.b.hawes@census.gov), Peter Meyer (meyer-peter@norc.org), and Rickita Walley (rwalley@sanametrix.com) by **April 3, 2020**.

GLOSSARY OF TERMS

Automated tools. Primarily meant to be data-computation programs, inventory systems, etc., that assist with the statistical confidentiality process.

Differential privacy. See http://privacytools.seas.harvard.edu/files/privacytools/files/pedagogical-document-dp_0.pdf.

Exhaustive tabulations. An example is to conduct all 4-way tabulations among 20 variables. Violations of k-anonymity could be tallied, for example, to identify high risk records, or categories.

Governance. The documents used to guide the functioning of the Confidentiality Officer or Disclosure Review Boards.

Log-linear approach. Skinner, C.J. and Shlomo, N. (2008). Assessing Identification Risk in Survey Microdata Using Log-linear Models. Journal of American Statistical Association, 103, 989–1001.

Manual/Rules. A document that contains your standards, rules, instructions, or computations.

Mu-Argus. See https://joinup.ec.europa.eu/solution/sdctools-tools-statistical-disclosure-control/news/mu-argus-version-513

OAS. Online analysis system, sometimes referred to as flexible table generators or table builders. Aggregated results are computed in real-time from underlying microdata.

PRAM. Post-RAndoMization is a statistical disclosure control treatment approach to add noise to categorical variables for microdata, or to tabular estimates.

Random perturbation. Perturbation approaches involve applying a controlled random treatment procedure to replace a subset of the original data values by other values, with the aim of introducing just enough noise or uncertainty into the microdata to reduce the disclosure risk to an acceptable level, while attempting to maintain multivariate associations.

Remote analysis system. A file is provided to the public with the same structure as the RUF. The file may or may not have analytical use. The user creates the program code and sends it to the data controller. The data controller runs the code off the RUF, reviews the output, and then sends the safe output to the user.

Static Tables. These are tables that are generated in-house, and then provided to the public.

SUDA. Special Unique Detection Algorithm. Based on Elliot, M. J., Manning, A. M., and Ford, R. W. (2002). A computational algorithm for handling the special unique problem. International Journal of Uncertainty, Fuzziness and Knowledge Based System, Vol 10, No. 5, pp 493–509.

Synthetic data approaches. Synthetic data approaches involve producing fully synthetic datasets or partially synthetic datasets that are mixtures of actual and multiply imputed values. Synthetic approaches typically replace original values with draws from appropriate probability distributions in a way that aims to retain the essential statistical features of the original data, including multivariate associations.

Template/checklist. A form that the customer would need to complete as a way of providing information to consider for statistical confidentiality.

U statistic. Woo, M., Reiter, J., Oganian, A., and Karr, A. (2009). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality* 1:111-124.

Verification servers. Similar to remote analysis system, but different in that the file is comprised of synthetic data of potential analytical use. The user feeds the code into the server, the server runs the code off the RUF and synthetic data and returns a message to the user relating to the usability of the output generated from the synthetic data.

Virtual analysis zones. Secure virtual areas with tools and data needed for users to conduct analyses under restricted use agreement.

REFERENCE GUIDE

(Marked entries have a corresponding Glossary entry)

Risk Assessment	**Data Access and Sharing**	**Risk Mitigation**	**Impact Assessment**	**Governance** ^{g}
Probabilistic matching	Restricted	Microdata	Risk – rerun risk assessment	Reviews (e.g., how to sign-off on product)
Re-identification studies	Restricted use files (RUFs)	Remove identifiers	Utility	Disclosure Review Board
Reconstruction studies	Remote analysis system ^{g}	Geographic population thresholds	Ad hoc before vs after	Agency Review
Microdata risk metrics	Research Data Center	Data rounding: random	Hellinger's Distance	Legislatively controlled review
Threshold rules (k-anonymity)	Data user agreements	Data rounding: controlled	Correlations	Confidentiality Officer review
Exhaustive tabulations ^{g}	Virtual analysis zones ^{g}	Sampling	Multivariate associations	Other
Special unique detection algorithm (SUDA)^{g}	Verification servers ^{g}	Variable suppression	Overlapping confidence intervals	
Mu-Argus ^{g}	Other	Value suppression	U statistic (Woo et al, 2009)^{g}	
Log-linear approach ^{g}	Public access	Data coarsening (recodes, top-codes, bottom-codes, etc)	Kolmogorov-Smirnov 2-sample test	
Other	Public use files (PUFs)	Data swapping	Fischer's z-transformation	
Static tables ^{g}	Running an OAS^{g} from RUF	Rank swapping	Chi-square approximation	
Households	Running an OAS from PUF	Data shuffling	Data utility assessment	
Establishments	Static tables ^{g} from RUF	Add random noise	Other	
p% rule	Other	Blank and impute		
pq rule		Blurring (e.g., microaggregation)		
(n,k) rule		Other random perturbation ^{g}		
(1,.6) rule		Fully synthetic data ^{g}		
Other		Partially synthetic data ^{g}		
Online analytic system (OAS)^{g}		Differential privacy^{g}		
Longitudinal data		Other		

Risk Assessment	Data Access and Sharing	Risk Mitigation	Impact Assessment	Governance [g]
Administrative data		Tables (frequency or magnitude)		
		Pre-tabular (see microdata)		
		Post-tabular		
		Roll-up (combine categories)		
		Cell suppression		
		Rounding		
		Random rounding		
		Controlled rounding		
		Controlled tabular adjustment		
		Add noise		
		Post-RAndoMization (PRAM)[g]		
		Differential privacy [g]		
		Other		

Data Protection Toolkit – Request Form for Existing Resources

Name and Organization__________________________________

Topic	**Resource Type**	**Link (URL) or file name (of file attached to email response) with a title and brief description.**	**Is the resource currently in active use?**	
			Yes	**No**
Risk Assessment	Automated tools		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]
	Manuals/ Rules		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]
	Templates/ Checklists		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]
	Governance		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]

Topic	Resource Type	Link (URL) or file name (of file attached to email response) with a title and brief description.	Is the resource currently in active use?	
			Yes	No
Data Access and Sharing	Automated tools		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]
	Manuals/ Rules		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]
	Templates/ Checklists		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]
	Governance		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]

			Is the resource currently in active use?	
Topic	**Resource Type**	**Link (URL) or file name (of file attached to email response) with a title and brief description.**	**Yes**	**No**
Risk Mitigation	Automated tools		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]
	Manuals/ Rules		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]
	Templates/ Checklists		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]
	Governance		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]

			Is the resource currently in active use?	
Topic	**Resource Type**	**Link (URL) or file name (of file attached to email response) with a title and brief description.**	**Yes**	**No**
Impact Assessment	Automated tools		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]
	Manuals/ Rules		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]
	Templates/ Checklists		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]
	Governance		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]

Topic	Resource Type	Link (URL) or file name (of file attached to email response) with a title and brief description.	Is the resource currently in active use?	
			Yes	No
Governance	Automated tools		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]
	Manuals/ Rules		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]
	Templates/ Checklists		[ ]	[ ]
			[ ]	[ ]
			[ ]	[ ]
Other				