Title	Disclosure Risk Worksheet
Author	ICPSR
Date	August, 2020
Notice	***IMPORTANT*** These are meant as guidelines and not steadfast rules applied equally in all cases. Final decisions are often determined by many intersecting considerations taken together. ICPSR has removed macros from the published version of this worksheet. The macros are used internally to locate where a dataset fits in the Harm by Re-identification matrix
About ICPSR:	Established in 1962, the Inter-university Consortium for Political and Social Research (ICPSR) provides leadership and training in data access, curation, and methods of analysis for a diverse and expanding social science research community. The ICPSR data archive is unparalleled in its depth and breadth; its data holdings encompass a range of disciplines, including political science, sociology, demography, economics, history, education, gerontology, criminal justice, public health, foreign policy, health and medical care, education, child care research, law, and substance abuse. ICPSR also hosts several sponsored projects focusing on specific disciplines or topics. Social scientists in all fields are encouraged to archive their data at ICPSR. (https:\\www.icpsr.umich.edu)

Documenting the Decisions on Disclosure Risk Remediation

Purpose	The purpose of this worksheet is to work through and document the determination of what disclosure risk remediation steps may need to be taken and whether the planned release/access level* needs to be changed. The worksheet does not, and cannot, include all possible relevant scenarios. The worksheet does not tell you what to do. It helps you think through what you might need to do.

Steps	
1	Go to the Overview tab and fill in the study title, number, a brief abstract/description, any PI notes on confidentiality, initial release (access) level set in JIRA ticket, and JIRA ticket number.
2	Generate frequencies from the data and run spssvarlabs to fill in the Variables tab (you will need an SPSS version of your data). Include variables that pose possible disclosure risk issues. Frequencies are not required in that tab, but have them for reference.
3	Use the Detail tab to mark the characteristics that describe the dataset. Place an x in the Assessment column if the item applies to the dataset. You should make a selection in each section. Use the comments column as needed. Once you are done marking the attributes in the Assessment column, click the Run Summarize button at the bottom of the column. All of the attributes that you marked will appear in the Summary tab in a condensed version. Make sure to save your file.
4	Once you are done marking the attributes in the Assessment column, click the Run Summarize button at the bottom of the column. All of the attributes that you marked will appear in the Summary tab in a condensed version. Make sure to save your file.
5	Review the output in the Summary tab and think about what it means in terms of potential harm to and potential reidentification of respondents.
6	Review the Scales tab and determine where your dataset fits in Harm and Re-identification scales. Place one X in each Assessment column and click on the Fill In Matrix button.
7	In the Matrix tab locate where your dataset fits in the Harm by Re-identification matrix.

8	Lastly, in the Recommendations tab write a few notes that describe why you are making the recommendations that you are. Typically two, three, or maybe four attributes of the dataset will dominate your decision process. Be sure to indicate those items. Also, if any attributes of the dataset are in conflict (e.g., vulnerable populations are present; no geography below U.S. level is present), indicate why one was more important than the other.

Go to the Overview tab and fill in the study title, number, a brief abstract/description, any PI notes on confidentiality, initial release (access) level set in JIRA ticket, and JIRA ticket number.

Title
Study number
Abstract /Description
PI notes on confidentiality (eg from Deposit Viewer)
JIRA ticket #

Initial release (access) level

Generate frequencies from the data and run spssvarlabs to fill in the Variables tab (you will need an SPSS version of your data). Include variables that pose possible disclosure risk issues. Frequencies are not required in this tab, but have them for reference.

Were frequencies reviewed? ☐ YES ☐ NO

Frequency file location:

Copy/paste relevant variables from spssvarlabs (mind the spacing/formatting) and fill in whether Type of Variable is: ID*, geographic (geog), demographic (demo), or date. Sort results by type. If there are no variables with disclosure risk concerns, enter "None" (in cell B8).

*ID: direct identifier

DS# (if nec)	Variable Name + Label	Type	Notes on Specificity of Variables (e.g., DOB includes mm-DD-yyyy or just mm-yyyy)

Use the Detail tab to mark the characteristics that describe the dataset. Place an x in the Assessment column if the item applies to the dataset. You should make a selection in each section. Use the comments column as needed.
Once you are done marking the attributes in the Assessment column, click the Run Summarize button at the bottom of the column. All of the attributes that you marked will appear in the Summary tab in a condensed version. Make sure to save your file.

Description	Assessment	Comments	Examples
Previously Reviewed for Disclosure Issues			
If the dataset was previously reviewed by a reputable disclosure review board (e.g., at the U.S. Census Bureau), then ICPSR is likely to give their assessment substantial consideration.			
Previously Reviewed - determined OK for public release		Specify by whom:	*Public Libraries in the United States Survey, 2015 (ICPSR 37138)*
Previously Reviewed - determined **not** OK for public release		Specify by whom:	
Not reviewed/unknown			
Living Persons / Active Organizations			
Do the data contain information on living persons?			
Do the data contain information on deceased persons with living relatives?			*Capital Punishment in the United States*
Do the data contain information on active organizations?			
Unknown			
Vulnerable Population			
Vulnerable populations are composed of individuals who may be more susceptible to coercion or undue influence, such as children, prisoners, pregnant women, mentally disabled persons, or economically or educationally disadvantaged persons. (See Common Rule definitions, http://research-compliance.umich.edu/glossary/common-rule)			
Children			
Prisoners			
Pregnant women			
Persons with diminished capacity (those who cannot understand things to the fullest extent, especially informed consent)			
Students			
Persons with disabilities			
Minorities			
AIDS/HIV+ Subjects			
Hospital patients			
Undocumented persons			
None			
Other: Mark this with an "x" and include Comments.			
Unknown			
Expectation of Privacy			
Much of the data that ICPSR archives was collected with an expectation that privacy would be maintained. Sometimes this expectation is unambiguous because data collectors have explicitly promised it to respondents. Other times there is a reasonable expectation of privacy.			
No expectation of privacy			
Expectation of privacy.			
Unknown			
Data Type/Level			
The type of data involved can greatly affect the decision of how to release the data collection.			
Aggregate data			*Uniform Crime Reporting Program Data Series*
Administrative records			
Survey data			*Annual Survey of Jails*
Census data			
Event/transaction data			
Sampling			
Geospatial			
Audiovisual			
Qualitative			
Biometric			
Administrative Census Sampling			
Transactional			Number of website hits; number of tweets
Observational			Cell phone tracking
None			
Other: Mark this with an "x" and include Comments.			

Unit of Analysis	
The unit of analysis describes the level at which information is disclosed, typically people, households, or transactions of some type.Transactions could be hospital admissions, cell phone conversations, arrests, medical treatments, etc. The more that specific people or organizations are identifiable in the data the greater the disclosure risk	
The unit of analysis refers to an individual.	
The unit of analysis refers to an organization.	
Unknown	

Sampling (Representation in Population/Universe)	
Many of the studies we archive are samples from a larger population/universe (e.g., News polls are samples taken from the U.S. general population). If every record in the dataset is represented in the population/universe by thousands of other persons/households/organizations, then the risk of reidentification is relatively low. The risk is low because no one can definitively say that a person in the sample is a specific person in the larger population/universe.	
Respondents in the data represent many in the population/universe.	e.g. National sample
Respondents represent few in the population/universe	e.g. Sub-sample in sub-national region
The dataset includes the entire population/universe; it's a census	*Capital Punishment in the United States*
Unknown	

Longitudinal Data	
Longitudinal data contains repeated measures taken over at least two periods of time. It also includes repeated samples over time. Longitudinal data poses additional disclosure risks because individuals become more unique with each additional measurement. Also, the number of potential external databases to link someone to increases.	
Data are primarily longitudinal	
Data are partially longitudinal	
Data are not longitudinal	
Unknown	

Data Available Elsewhere	
An important consideration affecting release method is the ability of the archived data to be matched to other data (at ICPSR or other sources) which make it more disclosive. Also keep in mind whether Googling pieces of information in the data will result in identifying the respondent.	
Other known data sources make these data more disclosive. If so, indicate them in the Comments.	*Substance Use Among Violently Injured Youth in an Urban Emergency Department: Services and Outcomes in Flint, Michigan, 2009-2013 (ICPSR 36769)*
There are currently no other known sources that these data match to.	
Unknown	

Social Relationships	
Social relationships are present in social science data whenever the dataset contains information about paired or multiple individuals with a common link within the dataset. Information about the individuals may be within one dataset or in multiple datasets that link by an identifier; however, both individuals need to be in the data for there to be a disclosure concern. Disclosure risk can be present due to two factors: 1) more information about the relationship through separate and overlapping characteristics reported by the respondents, and 2) ability to reidentify one respondent leads to greater ability to also reidentify others in their social relationship. A related issue is “third parties” in human subject research. A third-party is someone “about whom researchers obtain information from human subjects but who themselves have no interaction with investigators or their agents”.	
Entire household	
Mother and infant dyads	*Maternal Lifestyle Study (ICPSR 34312)*
Caregiver and youth respondents	*Project on Human Development in Chicago Neighborhoods (PHDCN Series)*
Spouse/partner dyads	*Victim Participation in Intimate Partner Violence Prosecution (ICPSR 30741)*
MSM/sexual partners	*Sexual Acquisition and Transmission of HIV Cooperative Agreement Program (SATHCAP) (ICPSR 29181) | Latino MSM Community Involvement: HIV Protective Effects (ICPSR 34385)*
Index child/mother/father/sibling respondents	*Center for Education and Drug Abuse Research (CEDAR) (ICPSR 33444)*
Teacher/principal and student respondents	
Victim and offender reports	
Offender/officer respondents	

Supervisor/officer respondents

Social network data (including respondent-driving sampling)

Other: Mark this with an "x" and include Comments.

Unknown

None

Health Consequences of Long-Term Injection Heroin Use Among Aging Mexican American Men in Houston (ICPSR 34896)

Institutional Environment

Institutional environment is a parallel measure to geospatial environment that describes the institution in which the person is housed. Characteristics are provided about the institution that could then re-identify the person. Group disclosure vs. individual disclosure: you can assume attributes of the individual by knowing attributes of the group.

For example, we may not initially know who a respondent to a survey is; however, if we know, or can find out, an institution that the respondent is in, then we are closer to identifying the respondent.

Prison

Hospital

School

Other: Mark this with an "x" and include Comments.

Unknown

None

Specificity of Geographic Environment

Knowing about the geographic space helps you identify an individual. Also a matter of distance and density measurements, which can help identify an individual (e.g., how many blocks do you live from school A). Important to conceptualize in terms of scope of study (national, regional), scale of study (block-group attributes vs. county-level attributes), Metropolitan Statistical Area status (urban vs. rural), distance vs. contextual vs. number.

Geographic detail

Country level only

State level

County

City

Other: Mark this with an "x" and include Comments.

Not Applicable

Unknown

Date Specificity

Having specific dates or many dates in the data make them more of a disclosure risk because it becomes easier to tie events in time to a person.

Dates include month, day and year.

Dates include month and year.

Dates include year.

Not Applicable

Unknown

Harm/Information Sensitivity

The terms confidential information and sensitive information are often confused. Sensitive information can cause harm or legal jeopardy; cause financial loss or damage reputation. Confidential data are information that has been promised to be kept secret. Direct Identifiers may or may not be sensitive but are confidential. Names are usually not sensitive; however, credit card numbers are. A dataset could contain sensitive information, but that doesn't mean those data could not be made public.

Health history (including sexual history) — *National Survey on Drug Use and Health*

Drug use — *Monitoring the Future*

Criminal record

Criminal victimization

School record — *National Longitudinal Study of Adolescent to Adult Health (Add Health)*

Other: Mark this with an "x" and include Comments.

Not Applicable

Unknown

Small or Distinct Populations

The concern with small or distinct populations is that a user may be able to reconstruct the sampling frame, thus increasing the likelihood of reidentification. There are three dimensions to consider: (1) universe size, (2) geographic/institutional scope, and (3) sampling RATE (overall vs. subsampling vs. census of subgroups)

The Vermont Study on Physician Aid-in-Dying, 2016-2018 (ICPSR 37209)

- Doctors
- University presidents
- Prosecutors
- Sheriffs
- Religion
- Minorities (ethnic, sexual orientation)
- Professional registries
- Other: Mark this with an "x" and include Comments.
- Not Applicable
- Unknown

Add any new items ABOVE this line.

Run Summarize

Review the output in the Summary tab and think about what it means in terms of potential harm to and potential reidentification of respondents.

Description	Assessment	Comments

Review the Scales tab and determine where your dataset fits in Harm and Re-identification scales. Place one X in each Assessment column and click on the Fill In Matrix button.

Place one X in the Assessment column for each scale based on your review of the data. Then click on the "Fill In Matrix" button.

Harm scale

				Assessment
Low	Very Low	0	No Harm	
		1	Little Harm	
	Low	2	Humiliation	
		3	Reputation Damage	
Moderate	Moderate	4	Financial Loss	
		5	Health Threat	
		6	Legal Jeopardy	
High	High	7	Prison	
		8	Physical Injury/Impairment	
	Very High	9	Disfigurement	
		10	Death	

Re-identification scale

				Assessment
Low	Very Low	0	Negligible risk	
		1	Unique profiles* without links*	
	Low	2	Unique profiles with slim chance of links	
		3	Unique profiles with small chance of links	
Moderate	Moderate	4	Unique profiles with possible links	
		5	Unique profiles with probable links	
		6	Unique profiles with known links	
High	High	7	Unique profiles with possible lookups*	
		8	Unique profiles with probable lookups	
	Very High	9	Unique profiles with known lookups	
		10	Personally Identifiable Information (PII)	

***Unique profile:** Set of variables when combined together form a profile which can be used to link data from different sources.
***Links:** Other sources of information that can be linked to data. Links increase the chances of re-identification and may enable the formation of a profile for lookup.
***Lookups:** Information that translates profiles into identities.

In the Matrix tab locate where your dataset fits in the Harm by Re-identification matrix.

Expectation of Privacy: YES

Harm x Re-identification

	V Low Re-ID	Low Re-ID	Mod Re-ID	Hi Re-ID	V Hi Re-ID
V Low Harm					
Low Harm					
Mod Harm					
Hi Harm					
V Hi Harm					

Lastly, in the Recommendation tab write a few notes that describe why you are making the recommendations that you are. Typically two, three, or maybe four attributes of the dataset will dominate your decision process. Be sure to indicate those items. Also, if any attributes of the dataset are in conflict (e.g., vulnerable populations are present; no geography below U.S. level is present), indicate why one was more important than the other.

Why - Describe, briefly, your disclosure risk decisions and if you are recommending a different release/access level.

Description: