Data Protection is *People* Protection

DATA PROTECTION & PRIVACY GUIDE:

Deidentifying Data

This guide provides an example of removing personally identifiable information (PII) from a dataset. There are several ways to "de-identify" data, which refers to the processing activities or methods that work to prevent a Data Subject's identity from being revealed. Two common types of de-identification are "Anonymization" and "Pseudonymization."

Anonymization is the process by which Personal Data is rendered anonymous so that an individual (or "data subject") is no longer identifiable: it is a permanent change to the data. Common methods include removing personally identifiable information or scrambling values across certain sets of PII.

> Example: imagine an organization has survey data that contains fields for name, national ID number, village name, ethnic affiliation, age, education level, and health indicators. In this case, removing name and national ID number would be the first step in making the data anonymous since these "direct attributes" are personal data that directly identify an individual. The "indirect attributes" of village name, ethnic affiliation, age, education level, and health indicators would remain.

However, even though some attributes seem "anonymous" they may not be. If the survey was collected in a very small village where only two residents identify as a particular ethnic affiliation, and they are each of different ages, then using those two indirect attributes could allow for those individuals to be identified! The process by which all attributes are examined to reduce the risk of re-identifying a data subject is called Statistical disclosure control. The first step in this process is a disclosure risk assessment and the Humanitarian Data Centre has an online tutorial for conducting a disclosure risk assessment.

Pseudonymization, on the other hand, describes the processing of personal data in a way that personal data can no longer be attributed to a specific data subject without the use of additional information, such as a key code.

> Example: imagine a survey contains your name, email address, age, nationality, and workplace. Pseudonymization takes the data that's identifiable about you specifically (your name, email address, age) and makes it inaccessible and separate from non-identifying data, like your nationality. Pseudonymous data can be put back together at some point so that all information can be linked back to a specific source or person. This is why pseudonymization requires that the additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to the data subject.

Should you choose Anonymization or Pseudonymization?

Anonymizing will generally be safer and reduce the risk of exposing PII. However, this can sometimes make the data too general, which may not make it useful for programs such as cash voucher assistance. In the case of health programs that involve vaccinations or other treatments, it may be important to contact individuals for follow-up treatment. In both of these cases pseudonymization would be the best choice since you can always put the data back together to identify an individual when needed.

There is no single right answer about when to choose one method over another and it is important to understand why the data were collected, the potential risks associated with holding that data, and the needs of the program, before choosing how to deidentify your data.

It's also important to understand that the techniques used to both anonymize data and to hack data are becoming ever more sophisticated and that even de-identified data isn't always one hundred percent secure. When in doubt, contact your data or IT team for assistance.

☆ Importance

Recent data breaches at the international Committee of the Red Cross, email hacks at the U.S. Agency for International Development, and improper data sharing by the U.N. High Commissioner for Refugees all show several ways in which humanitarian data are at risk. Data from household surveys, needs assessments and other forms of microdata make up an increasingly significant volume of data in the humanitarian sector. These types of data are critical to determining the needs and perspectives of program participants and the communities we work in, but these data also present risks. Understanding how to assess and manage the sensitivity of these data is essential to ensuring that they are used in a safe, ethical and effective manner in different response contexts.

Some advantages of using anonymized data over personal data include:

- protecting against inappropriate disclosure of personal data;
- fewer legal restrictions apply to anonymized data; and
- allowing organizations to create open or publicly accessible data while still complying with their data protection obligations.

Principles

De-identifying data is part of data processing, and personal data processing undertaken by humanitarian organizations should comply with the following principles.

- **Fairness and lawfulness of processing:** methods must comply with regional, national, or local legislation or policies that may limit what data can be de-identified and how certain technologies are used. Any processing of personal data should be transparent for the data subjects involved.
- **Purpose limitation:** humanitarian organizations should determine and set out the specific purposes for which data are processed. These purposes should be explicit and legitimate.
- **Proportionality:** ensure each particular activity related to the processing of personal data is appropriate for the stated goal. For example: is only the minimum required amount of data being collected? Are appropriate technical and organizational measures in place to reduce the risks associated with data processing?
- **Technology changes:** new datasets and new tools for analyzing them change and advance rapidly, and so do the means by which data are hacked or stolen. It is important to understand new and emerging risks to your data and continue to adjust your methods and practices accordingly.

Pseudonymization

This is an example of one way to de-identify data in a spreadsheet. There are a wide variety of ways to perform de-identification and this example uses a "key code" to remove personally identifiable information found in direct identifiers and keep it in a separate file. Personally Identifiable Information (PII) is information that can be used to identify an individual. Common examples are name, address, phone number, date of birth, and social security or national ID number.

Instructions

You can follow these Pseudonymization-instructions to walk through a basic example of pseudonymizing a data set. The exercise uses a sample data set found in the data folder of the online guide.

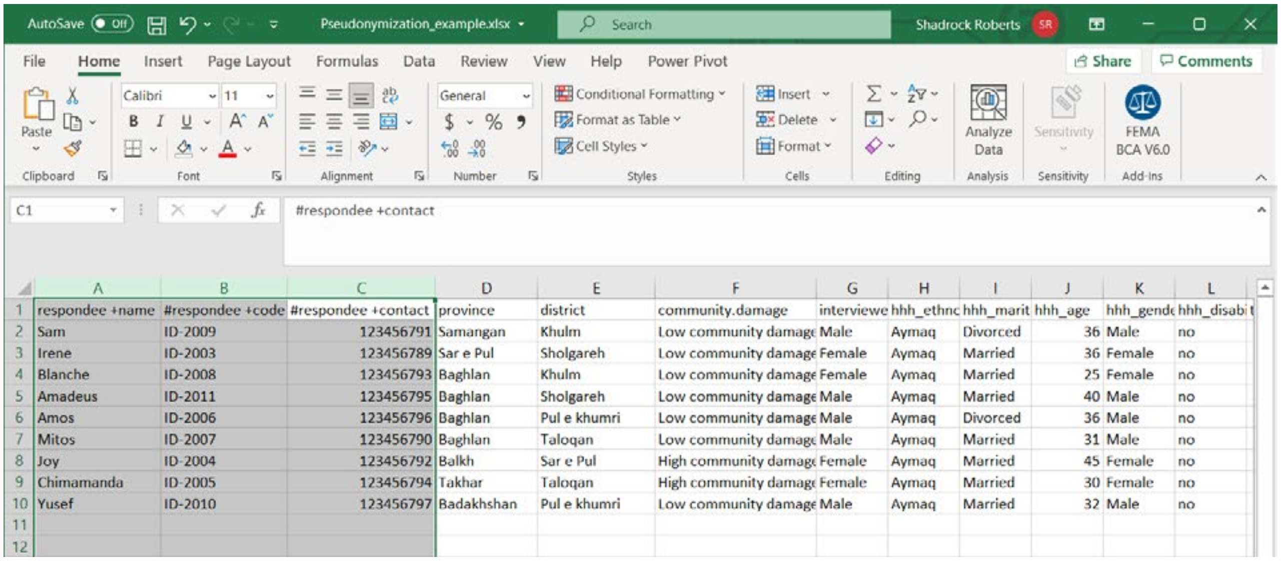
Once you have pseudonymized the sample data, you can continue with the Humanitarian Data Centre's tutorial for conducting a disclosure risk assessment.

Step 1 - Identify PII

Start by identifying PII in the data. Ideally, you will have metadata—data or a document that defines your data—to help you understand which fields contain PII. In the sample data, there are three columns that contain potential PII:

- **#respondee +name** appears to contain a name.
- **#respondee +code** likely contains an identification number of some kind.
- **#respondee +contact** possibly contains a mobile phone number

Each of these direct identifiers uses the Humanitarian Exchange Language for tagging data.

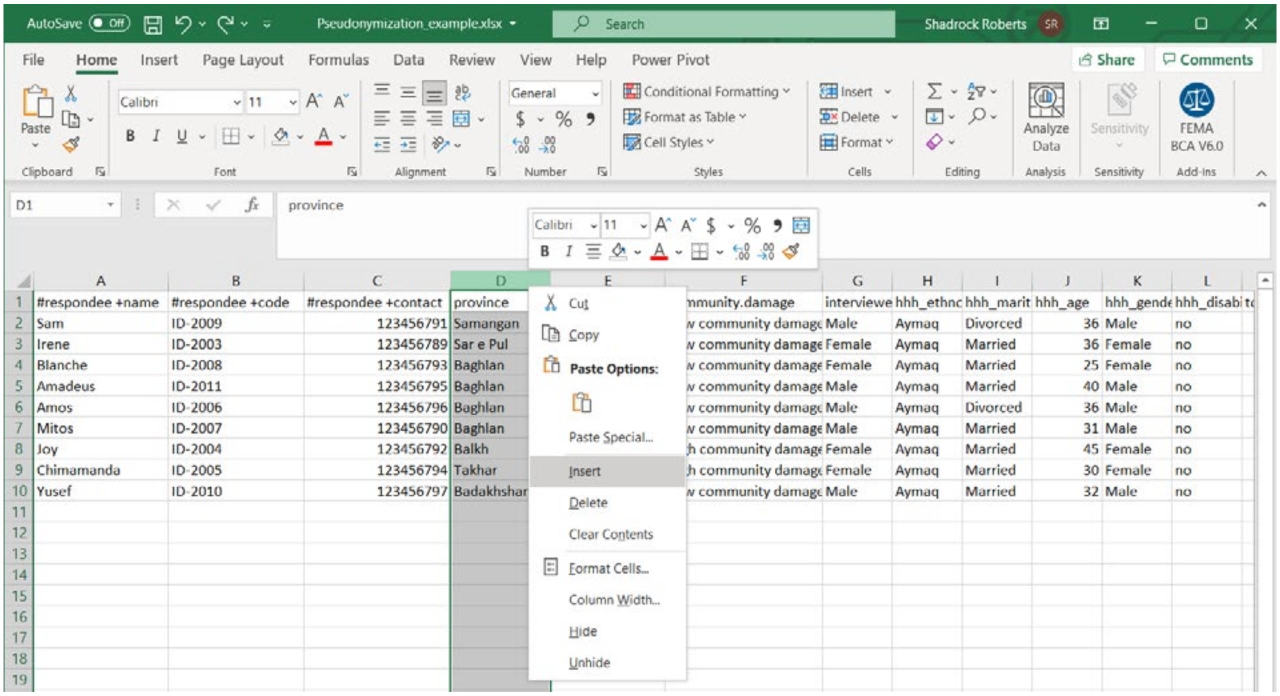


C1 #respondee +contact

	A	B	C	D	E	F	G	H	I	J	K	L
1	respondee +name	#respondee +code	#respondee +contact	province	district	community.damage	interviewe	hhh_ethnc	hhh_marit	hhh_age	hhh_gendc	hhh_disabi
2	Sam	ID-2009	123456791	Samangan	Khulm	Low community damage	Male	Aymaq	Divorced	36	Male	no
3	Irene	ID-2003	123456789	Sar e Pul	Sholgareh	Low community damage	Female	Aymaq	Married	36	Female	no
4	Blanche	ID-2008	123456793	Baghlan	Khulm	Low community damage	Female	Aymaq	Married	25	Female	no
5	Amadeus	ID-2011	123456795	Baghlan	Sholgareh	Low community damage	Male	Aymaq	Married	40	Male	no
6	Amos	ID-2006	123456796	Baghlan	Pul e khumri	Low community damage	Male	Aymaq	Divorced	36	Male	no
7	Mitos	ID-2007	123456790	Baghlan	Taloqan	Low community damage	Male	Aymaq	Married	31	Male	no
8	Joy	ID-2004	123456792	Balkh	Sar e Pul	High community damage	Female	Aymaq	Married	45	Female	no
9	Chimamanda	ID-2005	123456794	Takhar	Taloqan	High community damage	Female	Aymaq	Married	30	Female	no
10	Yusef	ID-2010	123456797	Badakhshan	Pul e khumri	Low community damage	Male	Aymaq	Married	32	Male	no
11												
12												

Step 2 - Create New Columns for the Key Code

We will use a key code, a value that we generate, to break out the PII. Since the direct identifiers are all grouped together, we'll create two new columns between columns C, **#respondee +contact** and column D, **province**. In Excel, we do this by highlighting a column to the right of where we want to insert new columns, right-click on the column and select **Insert**. Repeat this process again to create another empty column.

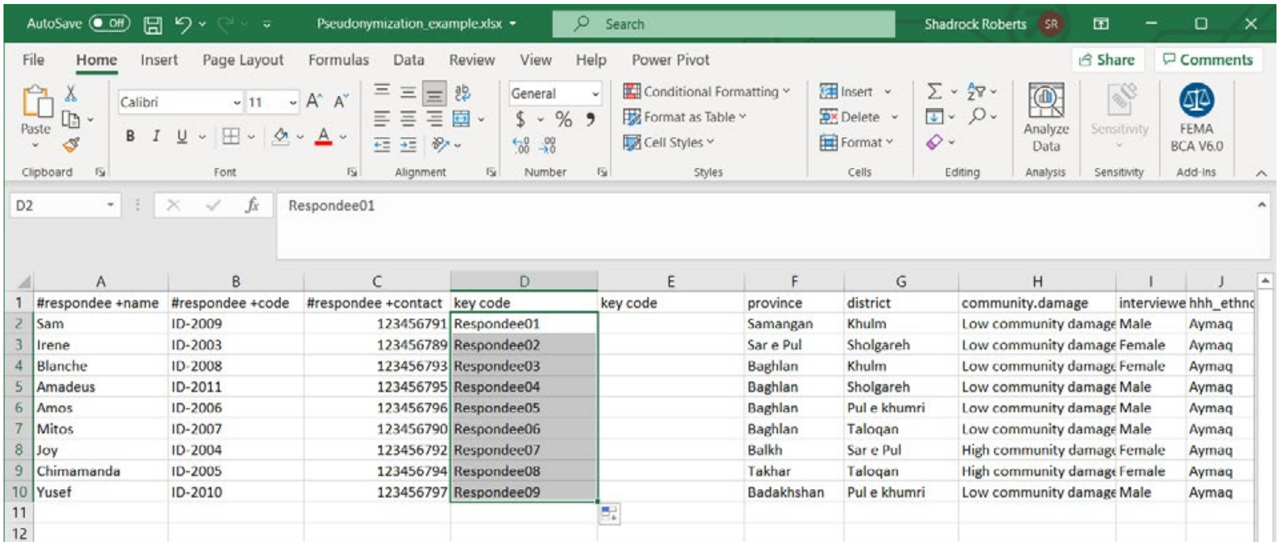


D1 province

	A	B	C	D	F	G	H	I	J	K	L
1	#respondee +name	#respondee +code	#respondee +contact	province	mmunity.damage	interviewe	hhh_ethnc	hhh_marit	hhh_age	hhh_gende	hhh_disabi
2	Sam	ID-2009	123456791	Samangan	w community damage	Male	Aymaq	Divorced	36	Male	no
3	Irene	ID-2003	123456789	Sar e Pul	w community damage	Female	Aymaq	Married	36	Female	no
4	Blanche	ID-2008	123456793	Baghlan	w community damage	Female	Aymaq	Married	25	Female	no
5	Amadeus	ID-2011	123456795	Baghlan	w community damage	Male	Aymaq	Married	40	Male	no
6	Amos	ID-2006	123456796	Baghlan	w community damage	Male	Aymaq	Divorced	36	Male	no
7	Mitos	ID-2007	123456790	Baghlan	w community damage	Male	Aymaq	Married	31	Male	no
8	Joy	ID-2004	123456792	Balkh	h community damage	Female	Aymaq	Married	45	Female	no
9	Chimamanda	ID-2005	123456794	Takhar	h community damage	Female	Aymaq	Married	30	Female	no
10	Yusef	ID-2010	123456797	Badakhshar	w community damage	Male	Aymaq	Married	32	Male	no

Step 3 - Create the Key Code

Start by naming your new columns. We'll use "key code" in each of them: each column will hold the same values. This would be a good time to update any metadata about this dataset to explain what **key code** means! Next, we'll use Excel's Auto Fill feature to create a simple code. Type **Respondee01** in the first cell. Next, highlight that cell, click on the drag handle in the lower right corner of the cell, and drag down to the end of the data set. This will automatically fill in the final number of each record so that each respondee now has a new code.

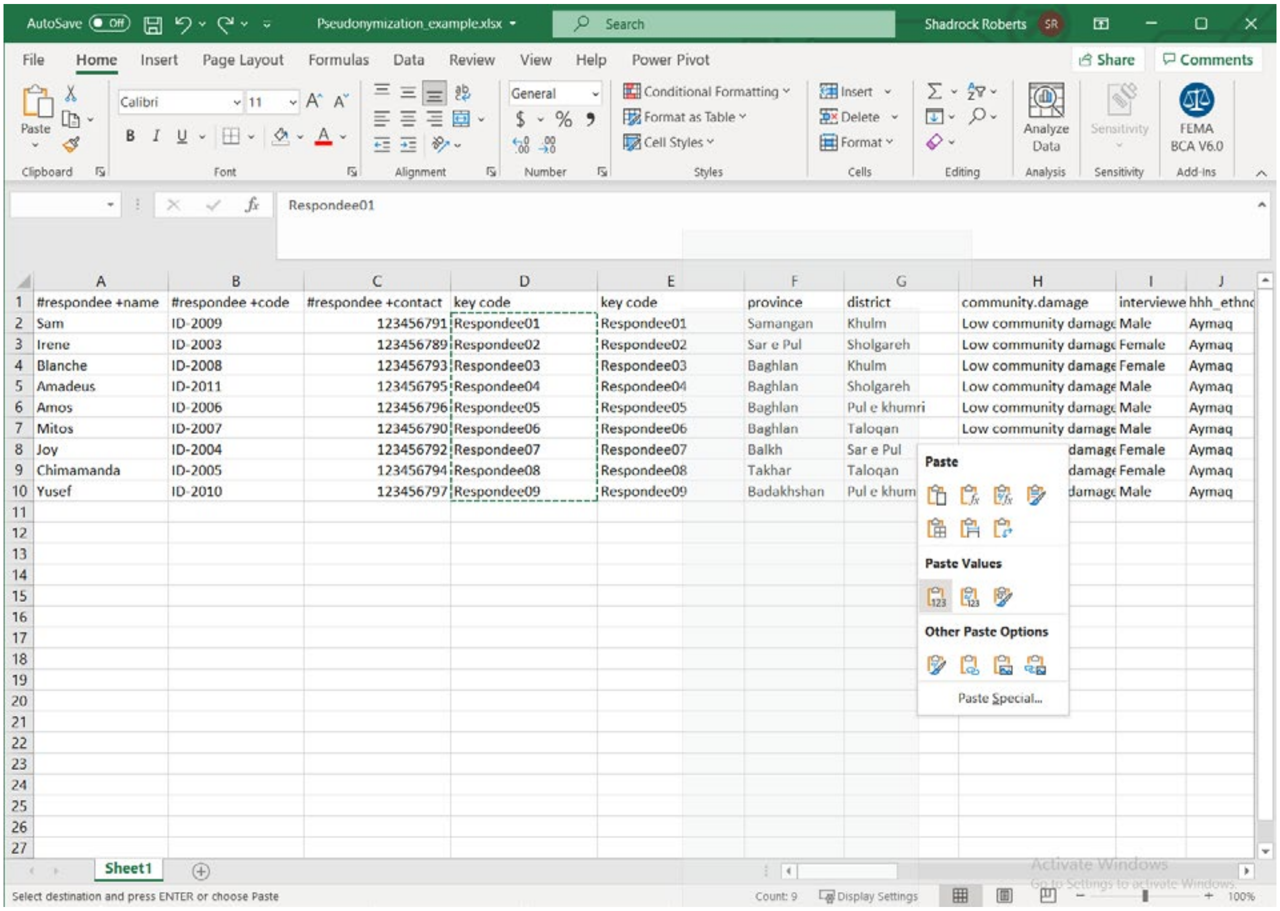


D2 Respondee01

	A	B	C	D	E	F	G	H	I	J
1	#respondee +name	#respondee +code	#respondee +contact	key code	key code	province	district	community.damage	interviewe	hhh_ethnc
2	Sam	ID-2009	123456791	Respondee01		Samangan	Khulm	Low community damage	Male	Aymaq
3	Irene	ID-2003	123456789	Respondee02		Sar e Pul	Sholgareh	Low community damage	Female	Aymaq
4	Blanche	ID-2008	123456793	Respondee03		Baghlan	Khulm	Low community damage	Female	Aymaq
5	Amadeus	ID-2011	123456795	Respondee04		Baghlan	Sholgareh	Low community damage	Male	Aymaq
6	Amos	ID-2006	123456796	Respondee05		Baghlan	Pul e khumri	Low community damage	Male	Aymaq
7	Mitos	ID-2007	123456790	Respondee06		Baghlan	Taloqan	Low community damage	Male	Aymaq
8	Joy	ID-2004	123456792	Respondee07		Balkh	Sar e Pul	High community damage	Female	Aymaq
9	Chimamanda	ID-2005	123456794	Respondee08		Takhar	Taloqan	High community damage	Female	Aymaq
10	Yusef	ID-2010	123456797	Respondee09		Badakhshan	Pul e khumri	Low community damage	Male	Aymaq

Step 4 - Duplicate the Key Code and Remove Formulas

Now we will copy the key code and paste it into the adjacent column. You can do this using basic keyboard commands such as **ctrl + C** or highlight the cells you want to copy, right click on them, and select **Copy**. In the adjacent column highlight the cells you want to paste the new key code into, right-click, and choose **Paste**. I've chosen to specifically paste only values. If you have used a formula to create a new code, then it will be important to retain *only the values* for use as a key code!



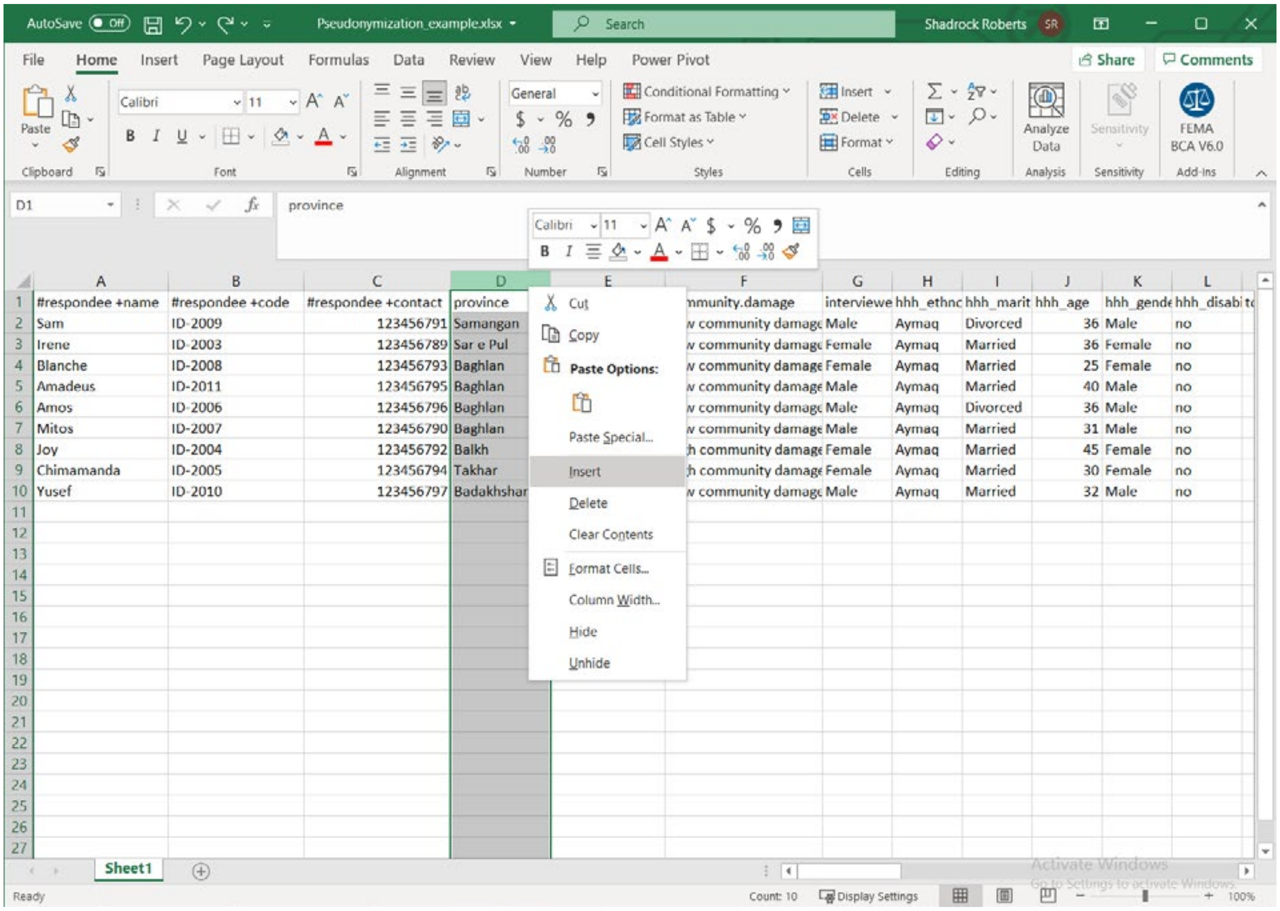
Respondee01

	A	B	C	D	E	F	G	H	I	J
1	#respondee +name	#respondee +code	#respondee +contact	key code	key code	province	district	community.damage	interviewe	hhh_ethnc
2	Sam	ID-2009	123456791	Respondee01	Respondee01	Samangan	Khulm	Low community damag	Male	Aymaq
3	Irene	ID-2003	123456789	Respondee02	Respondee02	Sar e Pul	Sholgareh	Low community damag	Female	Aymaq
4	Blanche	ID-2008	123456793	Respondee03	Respondee03	Baghlan	Khulm	Low community damage	Female	Aymaq
5	Amadeus	ID-2011	123456795	Respondee04	Respondee04	Baghlan	Sholgareh	Low community damage	Male	Aymaq
6	Amos	ID-2006	123456796	Respondee05	Respondee05	Baghlan	Pul e khumri	Low community damag	Male	Aymaq
7	Mitos	ID-2007	123456790	Respondee06	Respondee06	Baghlan	Taloqan	Low community damage	Male	Aymaq
8	Joy	ID-2004	123456792	Respondee07	Respondee07	Balkh	Sar e Pul	damage	Female	Aymaq
9	Chimamanda	ID-2005	123456794	Respondee08	Respondee08	Takhar	Taloqan	damage	Female	Aymaq
10	Yusef	ID-2010	123456797	Respondee09	Respondee09	Badakhshan	Pul e khum	damage	Male	Aymaq

Select destination and press ENTER or choose Paste

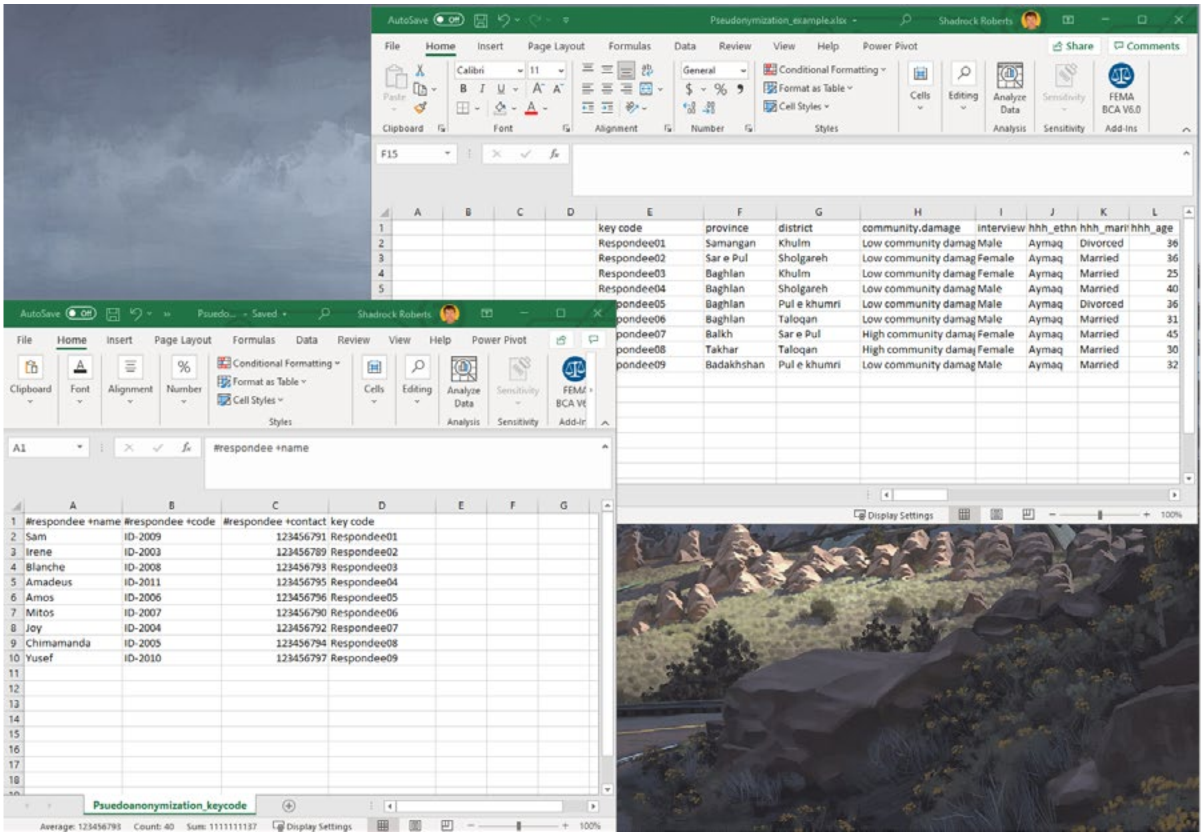
Step 5 - Separate Direct and Indirect Identifiers

Highlight the columns that contain the direct identifiers with PII along with one of the key code columns. In this example, we are highlighting columns A-D. Right-click on them and select **Cut**.



	A	B	C	D	E	F	G	H	I	J	K	L
1	#respondee +name	#respondee +code	#respondee +contact	province		mmunity.damage	intervieweе	hhh_ethnc	hhh_marit	hhh_age	hhh_gende	hhh_disabi
2	Sam	ID-2009	123456791	Samangan		w community damage	Male	Aymaq	Divorced	36	Male	no
3	Irene	ID-2003	123456789	Sar e Pul		w community damage	Female	Aymaq	Married	36	Female	no
4	Blanche	ID-2008	123456793	Baghlan		w community damage	Female	Aymaq	Married	25	Female	no
5	Amadeus	ID-2011	123456795	Baghlan		w community damage	Male	Aymaq	Married	40	Male	no
6	Amos	ID-2006	123456796	Baghlan		w community damage	Male	Aymaq	Divorced	36	Male	no
7	Mitos	ID-2007	123456790	Baghlan		w community damage	Male	Aymaq	Married	31	Male	no
8	Joy	ID-2004	123456792	Balkh		h community damage	Female	Aymaq	Married	45	Female	no
9	Chimamanda	ID-2005	123456794	Takhar		h community damage	Female	Aymaq	Married	30	Female	no
10	Yusef	ID-2010	123456797	Badakhshar		w community damage	Male	Aymaq	Married	32	Male	no

Next, open a new spreadsheet and paste these values using the keyboard shortcut **ctrl + V**, or other method. Save the new spreadsheet. You now have two spreadsheets: one of them contains indirect identifiers while the new sheet contains the direct identifiers with PII. Both datasets contain a key code for each record in the data so that all data can be recombined when necessary.



	E	F	G	H	I	J	K	L
1	key code	province	district	community.damage	interview	hhh_ethn	hhh_mari	hhh_age
2	Respondee01	Samangan	Khulm	Low community damag	Male	Aymaq	Divorced	36
3	Respondee02	Sar e Pul	Sholgareh	Low community damag	Female	Aymaq	Married	36
4	Respondee03	Baghlan	Khulm	Low community damag	Female	Aymaq	Married	25
5	Respondee04	Baghlan	Sholgareh	Low community damag	Male	Aymaq	Married	40
	pondee05	Baghlan	Pul e khumri	Low community damag	Male	Aymaq	Divorced	36
	pondee06	Baghlan	Taloqan	Low community damag	Male	Aymaq	Married	31
	pondee07	Balkh	Sar e Pul	High community dama	Female	Aymaq	Married	45
	pondee08	Takhar	Taloqan	High community dama	Female	Aymaq	Married	30
	pondee09	Badakhshan	Pul e khumri	Low community damag	Male	Aymaq	Married	32

A1: #respondee +name

	A	B	C	D
1	#respondee +name	#respondee +code	#respondee +contact	key code
2	Sam	ID-2009	123456791	Respondee01
3	Irene	ID-2003	123456789	Respondee02
4	Blanche	ID-2008	123456793	Respondee03
5	Amadeus	ID-2011	123456795	Respondee04
6	Amos	ID-2006	123456796	Respondee05
7	Mitos	ID-2007	123456790	Respondee06
8	Joy	ID-2004	123456792	Respondee07
9	Chimamanda	ID-2005	123456794	Respondee08
10	Yusef	ID-2010	123456797	Respondee09

Average: 123456793 Count: 40 Sum: 1111111137

Next Steps

Both files contain a key code that will allow them to be put back together. One way to do this in Excel, is to use the VLOOKUP function to automatically populate cells based on the value of other cells. In this case, you could populate empty cells in the original file with the missing PII based on the **keycode** value.

Because the new file contains the direct identifiers containing PII, it must be stored securely. One excellent way to do this is to encrypt the file and to use cloud storage to limit who has access to the file (**see the Encryption and File Sharing Best Practices guides**).

Remember: while the original spreadsheet has been deidentified by removing the direct identifiers that contain obvious PII, the other indirect identifiers have the potential to be combined with other data or analyzed in such a way as to allow for an individual to be identified.

For this reason, both files should still be stored securely. If you wanted to share the original, non-PII, file more widely it would be critical to perform a *disclosure risk assessment* to ensure the minimum amount of risk that the data could be re-identified. The Humanitarian Data Centre has an online tutorial for conducting a disclosure risk assessment using the open source statistical software "R". Additionally, Poverty Action Lab's De-identification for data publication web page provides an excellent discussion of data de-identification and includes sample code for the statistical software Stata. For Mercy Corps staff, Draft Guidance from T4D is available internally and provides additional Excel formulas.

Finally, all of these steps together help mitigate risk or exposing PII, so they should be listed in the PIA (**see the Privacy Impact Assessment guide**) so that others understand how these data are being protected.

Further Assistance

Deidentifying data is part of good data management practices and the larger data life cycle, which is the overall activities for individual data collection as part of a program or response. The following resources are excellent places to start for a more complete understanding of managing your data responsibly.

- The Cash Learning Partnership's Data Responsibility Toolkit is designed for cash and voucher practitioners specifically, but is a gold standard in guidance for responsible data. The Toolkit is available in English, Arabic, French, and Spanish.
- The Electronic Cash Transfer Learning Action Network's *Data Starter Kit for Humanitarian Field Staff* provides a series of data tip sheets for understanding various aspects of good data management and protection practices.
- The International Committee of the Red Cross' *Handbook on Data Protection in Humanitarian Action* is a detailed guide to almost every aspect of humanitarian data. Chapter 2 specifically deals with deidentifying data.
- The Engine Room's *Handbook of the Modern Development Specialist* is a good overview of data in the context of international development activities. The section on *Sharing Data* specifically deals with de-identification.

CONTACT

HEATHER LOVE
Director, Global Data Protection and Privacy | IT
hlove@mercycorps.org

SHADROCK ROBERTS
Data Protection Specialist | IT
shroberts@mercycorps.org

About Mercy Corps
Mercy Corps is a leading global organization powered by the belief that a better world is possible. In disaster, in hardship, in more than 40 countries around the world, we partner to put bold solutions into action—helping people triumph over adversity and build stronger communities from within. Now, and for the future.

MERCY
CORPS

Global Headquarters
45 SW Ankeny Street
Portland, Oregon 97204
888.842.0842
mercycorps.org

European Headquarters
40 Sciences
Edinburgh EH9 1NJ
Scotland, UK
+44.131.662.5160
mercycorps.org.uk