La protection des données est la protection des personnes

PROTECTION DES DONNÉES ET GUIDES DE CONFIDENTIALITÉ

MERCY CORPS

Désidentification des données

Ce guide fournit un exemple de suppression d'informations personnelles identifiables (PII) d'un ensemble de données. Il existe plusieurs façons de « désidentifier » les données, c'est-à-dire les activités ou les méthodes de traitement visant à empêcher la révélation de l'identité d'une personne concernée. Deux types courants de désidentification sont l'« anonymisation » et la « pseudonymisation ».

L'**anonymisation** est le processus par lequel les données personnelles sont rendues anonymes de sorte qu'un individu (ou « personne concernée ») n'est plus identifiable : il s'agit d'une modification permanente des données. Les méthodes courantes consistent à supprimer les informations d'identification personnelle ou à brouiller les valeurs de certains ensembles d'IPI.

> Exemple : imaginez qu'une organisation dispose de données d'enquête contenant des champs pour le nom, le numéro d'identification national, le nom du village, l'affiliation ethnique, l'âge, le niveau d'éducation et les indicateurs de santé. Dans ce cas, la suppression du nom et du numéro d'identification national serait la première étape pour rendre les données anonymes, car ces « attributs directs » sont des données personnelles qui identifient directement une personne. Les « attributs indirects » que sont le nom du village, l'appartenance ethnique, l'âge, le niveau d'éducation et les indicateurs de santé seraient conservés.

Toutefois, même si certains attributs semblent « anonymes », ils peuvent ne pas l'être Si l'enquête a été collectée dans un très petit village où seuls deux résidents s'identifient comme une affiliation ethnique particulière, et qu'ils sont tous d'âges différents, alors l'utilisation de ces deux attributs indirects pourrait permettre d'identifier ces individus ! Le processus par lequel tous les attributs sont examinés pour réduire le risque de réidentification d'une personne concernée est appelé contrôle statistique de la divulgation. La première étape de ce processus est l'évaluation du risque de divulgation et le Humanitarian Data Centre dispose d'un tutoriel en ligne pour réaliser une évaluation du risque de divulgation.

La **pseudonymisation**, quant à elle, décrit le traitement des données personnelles de telle sorte que celles-ci ne puissent plus être attribuées à une personne concernée spécifique sans l'utilisation d'informations supplémentaires, telles qu'un code clé.

> Exemple : imaginez qu'une enquête contienne votre nom, votre adresse électronique, votre âge, votre nationalité et votre lieu de travail. La pseudonymisation prend les données qui vous identifient spécifiquement (votre nom, votre adresse électronique, votre âge) et les rend inaccessibles et distinctes des données non identifiantes, comme votre nationalité. Les données pseudonymes peuvent être reconstituées à un moment donné afin que toutes les informations puissent être reliées à une source ou à une personne spécifique. C'est pourquoi la pseudonymisation exige que les informations supplémentaires soient conservées séparément et fassent l'objet de mesures techniques et organisationnelles visant à garantir que les données personnelles ne sont pas attribuées à la personne concernée.

Faut-il choisir l'anonymisation ou la pseudonymisation ?

L'anonymisation sera généralement plus sûre et réduira le risque d'exposer les informations personnelles identifiables (PII). Cependant, cela peut parfois rendre les données trop générales, ce qui peut ne pas les rendre utiles pour des programmes tels que l'aide sous forme de bons en espèces. Dans le cas des programmes de santé qui impliquent des vaccinations ou d'autres traitements, il peut être important de contacter les personnes pour un suivi du traitement. Dans ces deux cas, la pseudonymisation serait le meilleur choix, car il est toujours possible de reconstituer les données pour identifier un individu en cas de besoin.

Il n'existe pas de réponse unique et correcte quant au choix d'une méthode plutôt qu'une autre et il est important de comprendre pourquoi les données ont été collectées, les risques potentiels associés à la détention de ces données et les besoins du programme, avant de choisir la manière de dépersonnaliser vos données.

Il est également important de comprendre que les techniques utilisées pour rendre les données anonymes et pour les pirater deviennent de plus en plus sophistiquées et que même les données dépersonnalisées ne sont pas toujours sûres à cent pour cent. En cas de doute, contactez votre équipe chargée des données ou de l'informatique pour obtenir de l'aide.

☆ Importance

Les récentes violations de données au Comité international de la Croix-Rouge, les piratages de courriels s à l'Agence américaine pour le développement international et le partage inapproprié de données par l'ONU. Haut-Commissariat aux Réfugiés montrent tous plusieurs façons dont les données humanitaires sont en danger. Les données issues d'enquêtes auprès des ménages, d'évaluations des besoins et d'autres formes de micro données constituent un volume de données de plus en plus important dans le secteur humanitaire. Ces types de données sont essentiels pour déterminer les besoins et les perspectives des participants aux programmes et des communautés dans lesquelles nous travaillons, mais ces données présentent également des risques. Il est essentiel de comprendre comment évaluer et gérer la sensibilité de ces données pour garantir qu'elles sont utilisées de manière sûre, éthique et efficace dans différents contextes d'intervention.

Les avantages de l'utilisation de données anonymes par rapport aux données personnelles sont notamment les suivants :

- la protection contre la divulgation inappropriée de données à caractère personnel,
- moins de restrictions légales s'appliquent aux données anonymes et
- permettant aux organisations de créer des données ouvertes ou accessibles au public tout en se conformant à leurs obligations en matière de protection des données.

Principes

La dépersonnalisation des données fait partie du traitement des données, et le traitement des données personnelles entrepris par les organisations humanitaires doit respecter les principes suivants.

- **L'équité et la licéité du traitement :** les méthodes doivent être conformes à la législation ou aux politiques régionales, nationales ou locales qui peuvent limiter les données pouvant être dépersonnalisées et l'utilisation de certaines technologies. Tout traitement de données à caractère personnel doit être transparent pour les personnes concernées.

- **Limitation de la finalité :** les organisations humanitaires doivent déterminer et énoncer les finalités spécifiques pour lesquelles les données sont traitées. Ces objectifs doivent être explicites et légitimes.
- **Proportionnalité :** s'assurer que chaque activité particulière liée au traitement des données personnelles est appropriée à l'objectif déclaré. Par exemple : la collecte de données se limite-t-elle au minimum requis ? Des mesures techniques et organisationnelles appropriées sont-elles en place pour réduire les risques liés au traitement des données ?
- **Changements technologiques :** les nouveaux ensembles de données et les nouveaux outils pour les analyser changent et progressent rapidement, tout comme les moyens par lesquels les données sont piratées ou volées. Il est important de comprendre les risques nouveaux et émergents pour vos données et de continuer à adapter vos méthodes et pratiques en conséquence.

Pseudonymisation

Voici un exemple d'une façon de dépersonnaliser des données dans une feuille de calcul. Il existe de nombreuses façons de procéder à la dépersonnalisation. Cet exemple utilise un "code clé" pour supprimer les informations personnelles identifiables contenues dans les identifiants directs et les conserver dans un fichier séparé. Les informations personnellement identifiables (IPI) sont des informations qui peuvent être utilisées pour identifier un individu. Les exemples courants sont le nom, l'adresse, le numéro de téléphone, la date de naissance et le numéro de sécurité sociale ou d'identification nationale.

Instructions

Vous pouvez suivre ces instructions de pseudonymisation pour parcourir un exemple de base de pseudonymisation d'un ensemble de données. L'exercice utilise un ensemble de données type qui se trouve dans le dossier de données du guide en ligne.

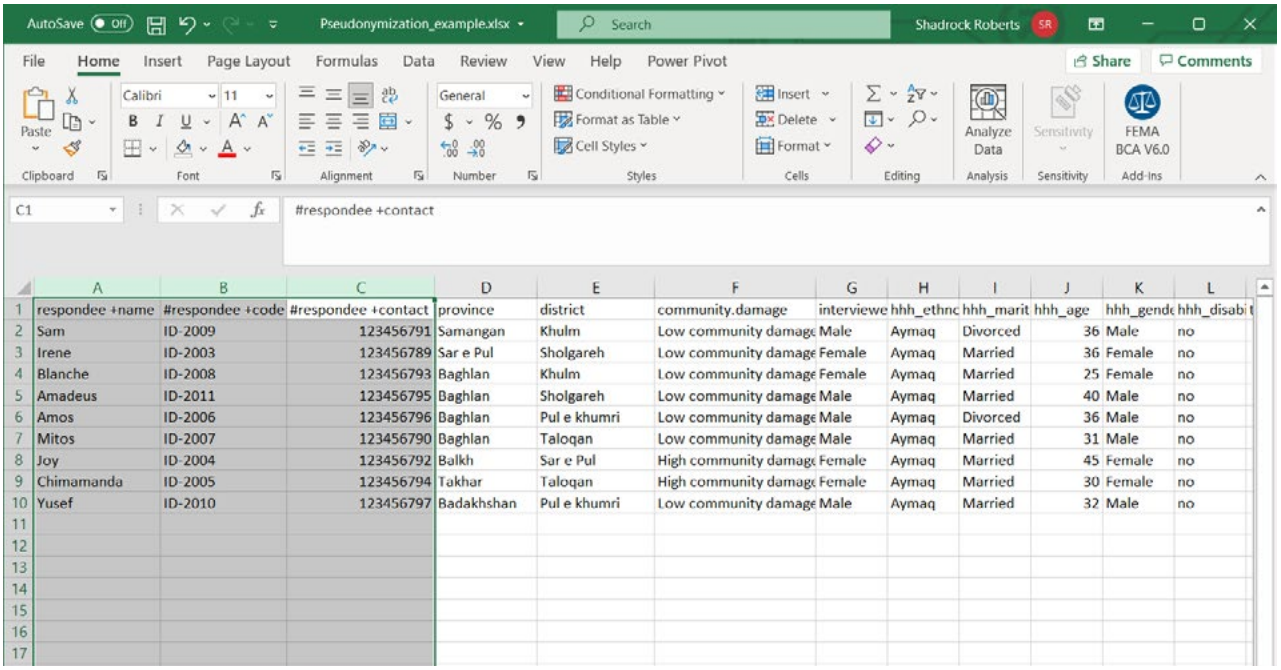
Une fois que vous avez pseudonymisé les données de l'échantillon, vous pouvez continuer avec le tutoriel pour effectuer une évaluation du risque de divulgation.

Étape 1 - Identifier les informations personnelles identifiables (PII)

Commencez par identifier les informations personnelles identifiables (PII) dans les données. Idéalement, vous disposerez de métadonnées - données ou document définissant vos données - pour vous aider à comprendre quels champs contiennent des informations personnelles identifiables (PII). Dans les données de l'échantillon, il y a trois colonnes qui contiennent des informations personnelles identifiables (PII) potentielles :

- **#respondee +name** semble contenir un nom.
- **#respondee +code** contient probablement un numéro d'identification quelconque.
- **#respondee +contact** contient éventuellement un numéro de téléphone mobile

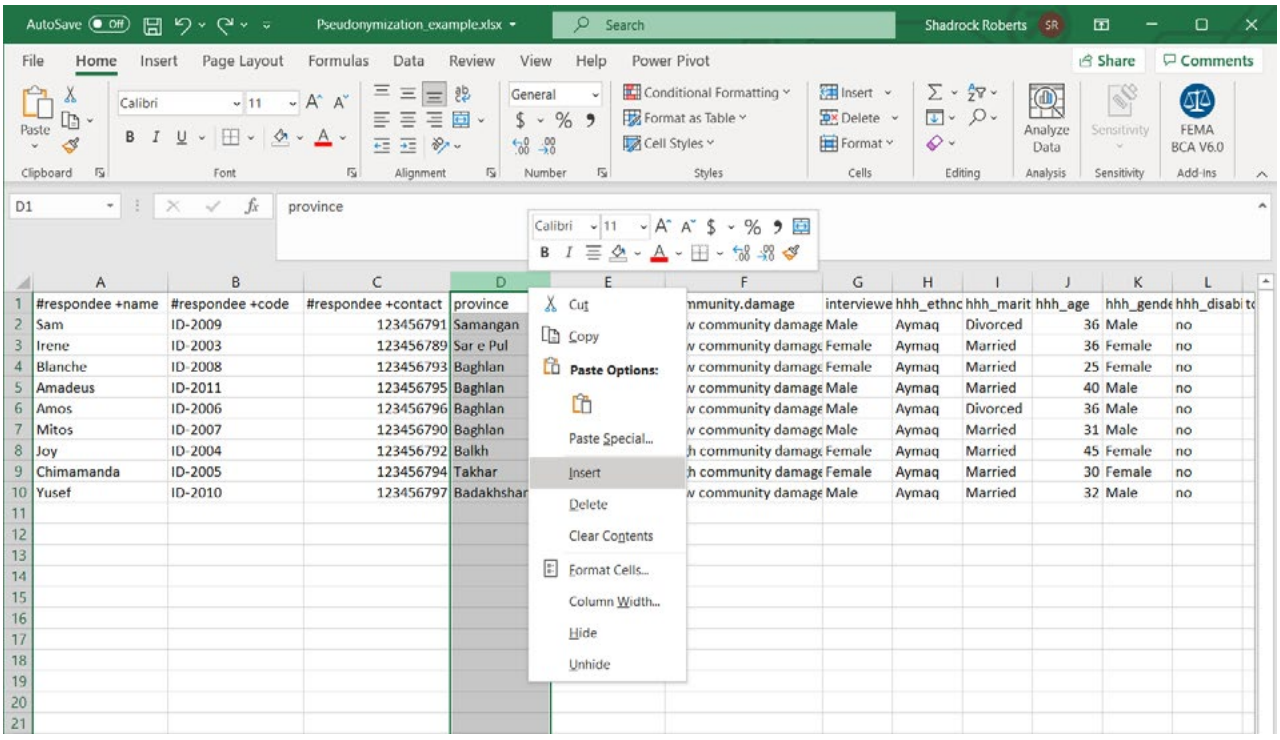
Chacun de ces identificateurs directs utilise le Humanitarian Exchange Language for tagging data.



	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_gende	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 damag	Female	Aymaq	Married	45	Female	no
9	Chimamanda	ID-2005	123456794	Takhar	Taloqan	High community damag	Female	Aymaq	Married	30	Female	no
10	Yusef	ID-2010	123456797	Badakhshan	Pul e khumri	Low community damage	Male	Aymaq	Married	32	Male	no

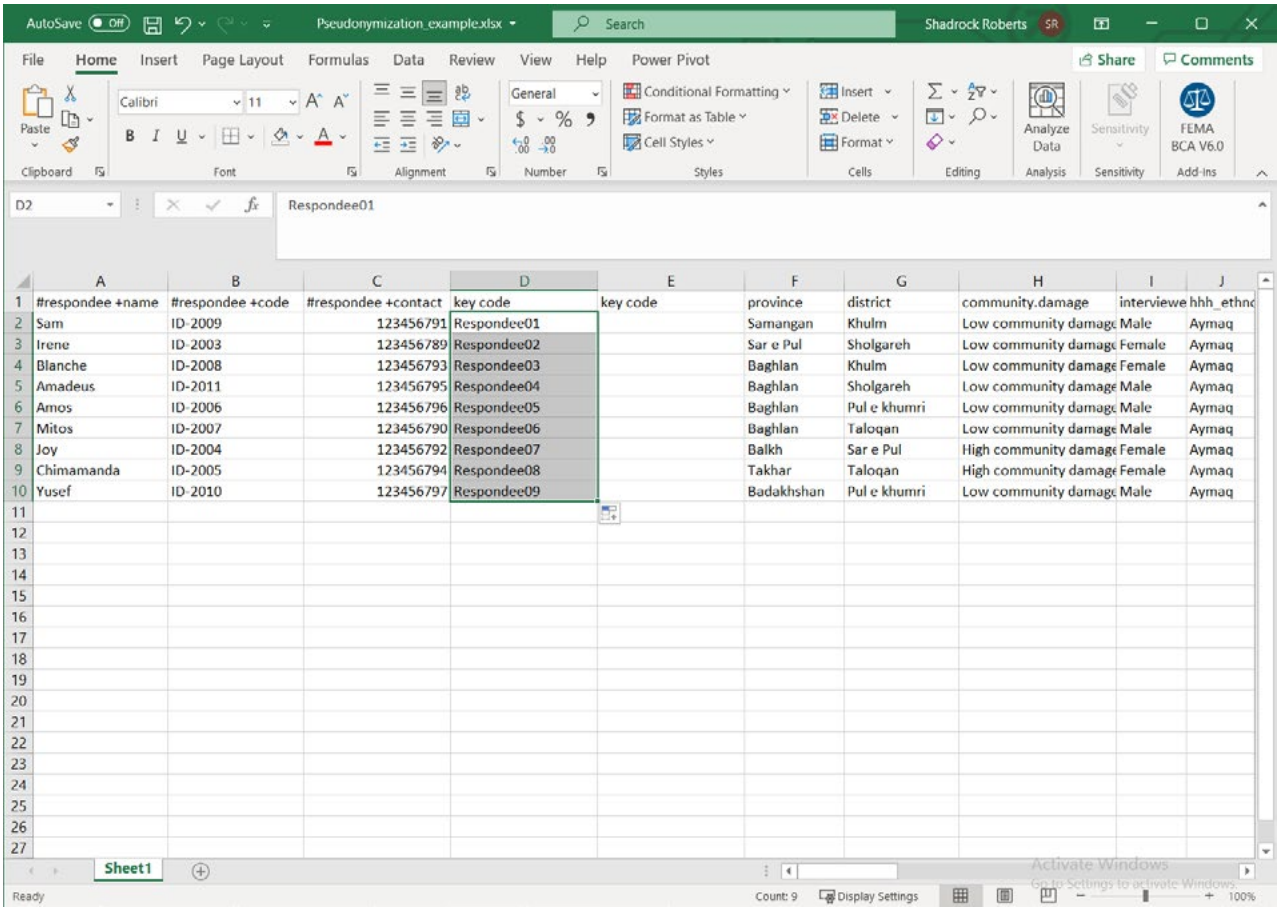
Étape 2 - Créer de nouvelles colonnes pour le code clé

Nous utiliserons un code clé, une valeur que nous générons, pour extraire les informations personnelles identifiables (PII). Puisque les identifiants directs sont tous regroupés, nous allons créer deux nouvelles colonnes entre les colonnes C, **#respondee +contact** et la colonne D **province**. Dans Excel, nous faisons cela en mettant en surbrillance une colonne à droite de l'endroit où nous voulons insérer de nouvelles colonnes, en faisant un clic droit sur la colonne et en sélectionnant **Insérer**. Répétez ce processus pour créer une autre colonne vide.



Étape 3 - Créer le code clé

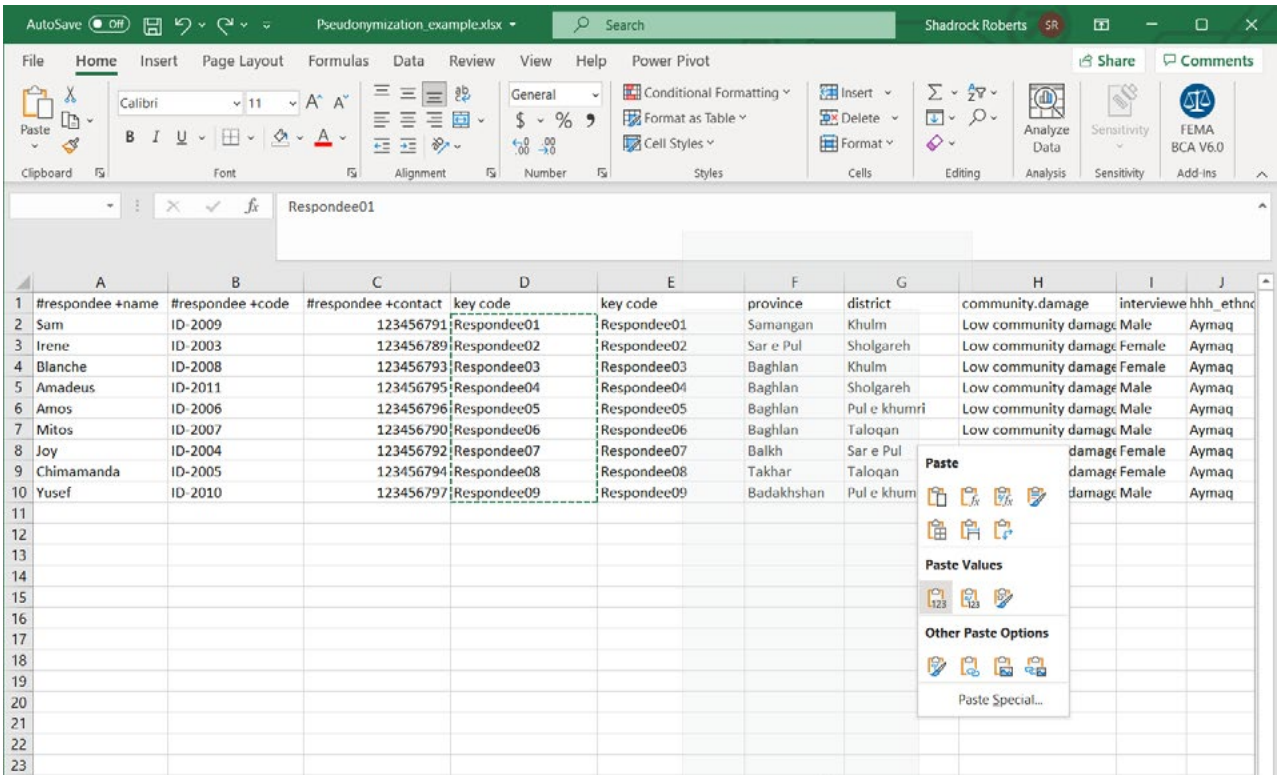
Commencez par nommer vos nouvelles colonnes. Nous utiliserons un « code clé » dans chacune d'elles : chaque colonne contiendra les mêmes valeurs. Ce serait le bon moment pour mettre à jour les métadonnées de cet ensemble de données afin d'expliquer la signification du **code clé** ! Ensuite, nous allons utiliser la fonction de remplissage automatique d'Excel pour créer un code simple. Tapez **Respondee01** dans la première cellule. Ensuite, mettez cette cellule en surbrillance, cliquez sur la poignée de déplacement dans le coin inférieur droit de la cellule et faites-la glisser jusqu'à la fin de l'ensemble de données. Le numéro final de chaque enregistrement sera automatiquement complété, de sorte que chaque personne interrogée aura désormais un nouveau code.



	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

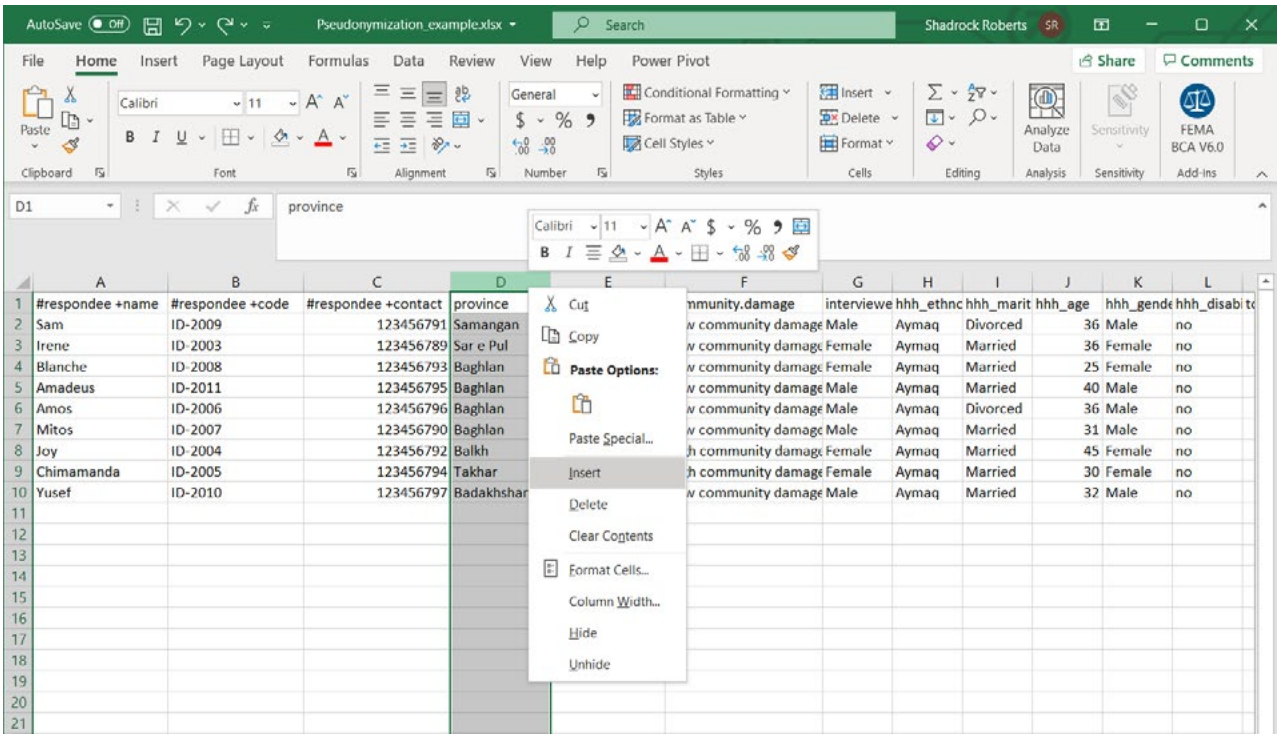
Étape 4 - Dupliquer le code clé et supprimer les formules

Nous allons maintenant copier le code clé et le coller dans la colonne adjacente. Pour ce faire, vous pouvez utiliser des commandes clavier de base telles que ctrl + C ou mettre en surbrillance les cellules que vous souhaitez copier, faire un clic droit dessus et sélectionner Copier. Dans la colonne adjacente, mettez en surbrillance les cellules dans lesquelles vous souhaitez coller le nouveau code clé, faites un clic droit et choisissez Coller. J'ai choisi de ne coller spécifiquement que les valeurs. Si vous avez utilisé une formule pour créer un nouveau code, il sera important de ne conserver *que les valeurs* pour les utiliser comme code clé !

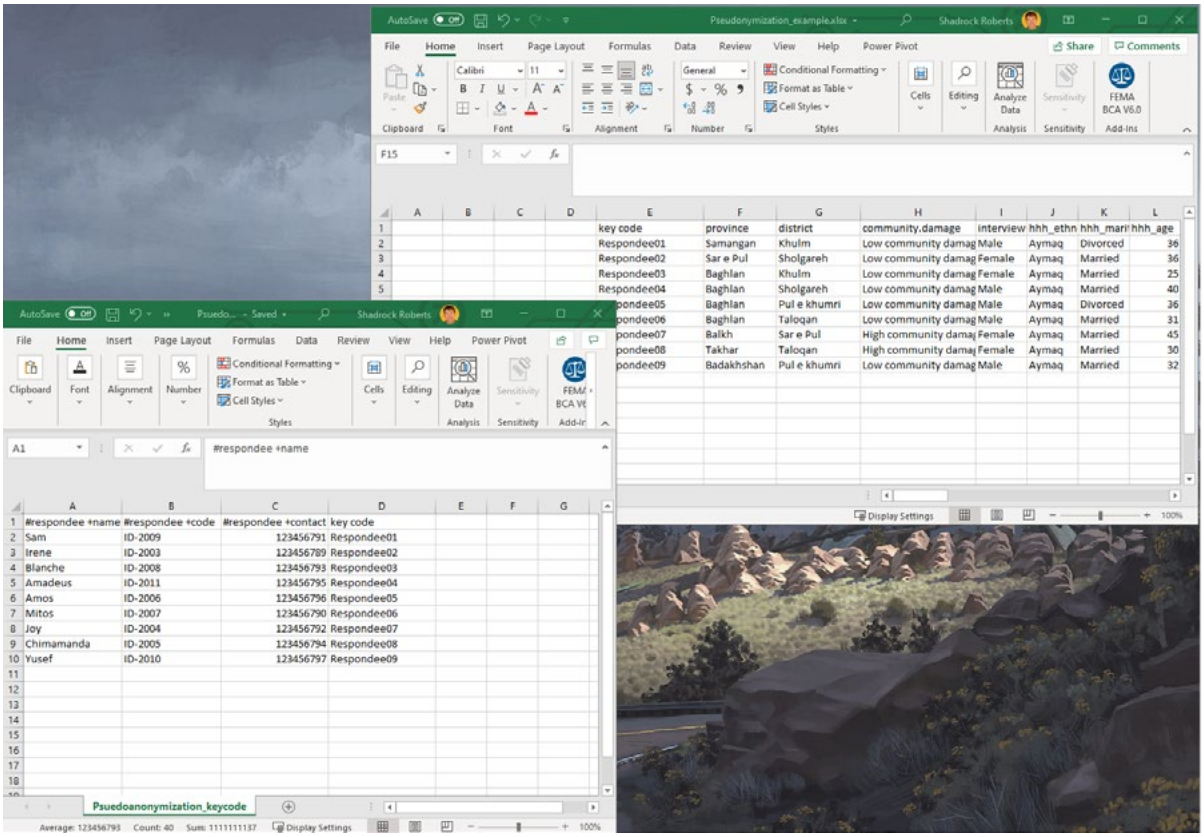


Étape 5 - Séparer les identifiants directs et indirects

Mettez en évidence les colonnes qui contiennent les identificateurs directs avec les informations personnelles identifiables (PII) ainsi qu'une des colonnes de code clé. Dans cet exemple, nous mettons en surbrillance les colonnes A-D. Cliquez dessus avec le bouton droit de la souris et sélectionnez **Couper**.



Ensuite, ouvrez une nouvelle feuille de calcul et collez ces valeurs en utilisant le raccourci clavier **ctrl + V**, ou une autre méthode. Enregistrez la nouvelle feuille de calcul. Vous avez maintenant deux feuilles de calcul : l'une d'elles contient les identifiants indirects tandis que la nouvelle feuille contient les identifiants directs avec les informations personnelles identifiables (PII). Les deux ensembles de données contiennent un code clé pour chaque enregistrement des données afin que toutes les données puissent être recombinées si nécessaire.



Next Steps

Les deux dossiers contiennent un code clé qui permettra de les reconstituer. Une façon d'y parvenir dans Excel est d'utiliser la fonction VLOOKUP pour remplir automatiquement les cellules en fonction de la valeur d'autres cellules. Dans ce cas, vous pouvez remplir les cellules vides du fichier d'origine avec les informations personnelles identifiables (PII) manquantes en vous basant sur la valeur du **code clé**.

Comme le nouveau fichier contient les identifiants directs contenant les informations personnelles identifiables (PII), il doit être stocké de manière sécurisée. Un excellent moyen d'y parvenir est de crypter le fichier et d'utiliser le stockage en nuage pour limiter les personnes ayant accès au fichier (**voir les guides des meilleures pratiques en matière de cryptage et de partage de fichiers**).

N'oubliez pas : si la feuille de calcul originale a été dépersonnalisée en supprimant les identifiants directs qui contiennent des informations personnelles identifiables (PII) évidentes, les autres identifiants indirects peuvent être combinés avec d'autres données ou analysés de manière à permettre l'identification d'une personne.

Pour cette raison, les deux fichiers doivent toujours être stockés de manière sécurisée. Si vous souhaitiez partager plus largement le fichier original, sans DPI, il serait essentiel de procéder à une évaluation du risque de divulgation afin de garantir le risque minimum que les données puissent être réidentifiées. Le Humanitarian Data Centre dispose d'un tutoriel en ligne permettant de réaliser une évaluation du risque de divulgation à l'aide d'un du logiciel statistique à code source ouvert « R ». De plus, la page web Poverty Action Lab's De-identification for data publication fournit une excellente discussion sur la dé-identification des données et inclut un exemple de code pour le logiciel statistique Sata. Pour le personnel de Mercy Corps, l'ébauche de directive de T4D est disponible en interne et fournit des formules Excel supplémentaires.

Enfin, toutes ces mesures contribuent à atténuer le risque d'exposition des DPI. Elles doivent donc être mentionnées dans l'évaluation des incidences sur la vie privée (**voir le guide de l'évaluation des incidences sur la vie privée**) afin que les autres comprennent comment ces données sont protégées.

Assistance supplémentaire

La désidentifcation des données fait partie des bonnes pratiques de gestion des données et du cycle de vie des données au sens large, c'est-à-dire l'ensemble des activités de collecte de données individuelles dans le cadre d'un programme ou d'une intervention. Les ressources suivantes sont d'excellents points de départ pour une compréhension plus complète de la gestion responsable de vos données.

- La trousse à outils sur la responsabilité des données du Cash Learning Partnership est conçue spécifiquement pour les praticiens du secteur des transferts monétaires, mais elle constitue une référence en matière de conseils sur les données responsables. La trousse à outils est disponible en anglais, arabe, Français et espagnol.
- La *trousse à outils de données pour le personnel de l'humanitaire de l'Electronic Cash Transfer Learning Action Network's* fournit une série de fiches-conseils sur les données permettant de comprendre les différents aspects des bonnes pratiques de gestion et de protection des données.
- Le Manuel sur la protection des données dans l'action humanitaire du *Comité international de la Croix-Rouge* est un guide détaillé sur presque tous les aspects des données humanitaires. Le chapitre 2 traite spécifiquement de la désidentification des données.
- Le manuel *du spécialiste en développement moderne de The Engine Room* est un bon aperçu des données dans le contexte des activités de développement international. La section sur le *partage des données* traite spécifiquement de la désidentification.

CONTACT

HEATHER LOVE
Directrice, Protection des données mondiales et confidentialité | IT
hlove@mercycorps.org

SHADROCK ROBERTS
Spécialiste de la protection des données | IT
shroberts@mercycorps.org

À propos de Mercy Corps

Mercy Corps est une organisation mondiale de premier plan animée par la conviction qu'un monde meilleur est possible. Nous nous associons dans plus de 40 pays du monde pour mettre en œuvre des solutions audacieuses, en aidant les gens à triompher de l'adversité lors de catastrophes et de situations difficiles, et à bâtir des communautés plus fortes de l'intérieur. Maintenant et pour l'avenir.

MERCY CORPS

Siège mondial
45 SW Ankeny Street
Portland, Oregon 97204
888.842.0842
mercycorps.org

Siège européen
96/3 Commercial Quay
Édimbourg,
EH6 6LX Écosse, RU
+44.131.662.5160
mercycorps.org.uk