OCHA centre for humdata

THE CENTRE FOR HUMANITARIAN DATA

SÉRIE DE NOTES D'ORIENTATION
LA RESPONSABILITÉ DES DONNÉES DANS L'ACTION HUMANITAIRE

CONTRÔLE DE LA DIVULGATION DE DONNÉES STATISTIQUES

POINTS CLÉS :

- Les données provenant d'enquêtes auprès des ménages ou d'évaluations des besoins, ou d'autres formes de microdonnées constituent un volume croissant de données dans le secteur humanitaire. Ce type de données est essentiel pour définir les besoins et les perspectives des populations affectées par les crises, mais présente également des risques particuliers.
- Même si une organisation supprime des microdonnées d'identifiants directs tels que le nom ou le numéro de téléphone d'une personne, la combinaison de variables clés telles que la localisation ou l'origine ethnique peut encore permettre la réidentification d'individus et de groupes vulnérables.
- Le contrôle de la divulgation de données statistiques (« CDS », de l'anglais « Statistical Disclosure Control - SDC ») est une technique pour évaluer et réduire le risque qu'une personne ou une organisation soit réidentifiée au cours de l'analyse des microdonnées. L'application du contrôle de la divulgation de données statistiques aux microdonnées permet aux organisations de partager les données plus largement sans mettre les personnes affectées en danger.
- Le CDS peut être utilisé pour réduire le risque de réidentification à un seuil convenu pouvant varier en fonction du contexte dans lequel se déroule la réponse humanitaire. La valeur informative globale ou l'utilité d'un jeu de données seront toujours affectées lors de l'application du contrôle de la divulgation statistique ; un juste équilibre entre le risque de réidentification et la perte d'informations est essentiel pour assurer une utilisation sûre, éthique et efficace des données.
- Pour commencer à utiliser le contrôle de la divulgation de données statistiques, les organisations doivent investir dans (1) la sélection d'un outil approprié, (2) l'intégration dudit contrôle dans les flux de gestion des données existants et (3) l'amélioration de la pratique à travers l'apprentissage continu.

QUELLE EST LA DÉFINITION DES MICRODONNÉES HUMANITAIRES ?

On entend par « microdonnées », les données relatives aux caractéristiques des unités statistiques d'une population - telles que les individus, les ménages ou les établissements - recueillies lors de recensements ou d'enquêtes.[1] Dans la réponse humanitaire, ce type de données est généralement recueilli dans le cadre d'exercices tels que les enquêtes sur les ménages et les évaluations des besoins. Les microdonnées constituent un volume de données toujours plus élevé dans le secteur humanitaire et sont essentielles pour définir les besoins et les perspectives des personnes affectées par les crises.[2] Les organisations humanitaires doivent comprendre comment évaluer et gérer la sensibilité de ce type de données afin d'en assurer l'utilisation responsable dans différents contextes d'intervention.

[1] Survey Design and Statistical Methodology Metadata, Software and Standards Management Branch, Systems Support Division, United States Bureau of the Census, Washington D.C., August 1998, Section 3.4.4, page 39.

[2] Au moment de la rédaction du présent document, une recherche du mot « enquête » sur le site Humanitarian Data Exchange a renvoyé 1 198 résultats sur les 9 805 jeux de données de la plateforme ; une recherche du mot « évaluation » a renvoyé 1 399 résultats.

Les microdonnées brutes peuvent contenir à la fois des données à caractère personnel et des données à caractère non personnel sur une variété de sujets, y compris des sujets sensibles tels que la violence basée sur le genre, les maladies infectieuses et d'autres problématiques qui pourront être enregistrés dans des champs de texte libre. La plupart des organisations humanitaires reconnaissent la sensibilité des données personnelles tel qu'un nom, une donnée biométrique ou un numéro d'identification, et de manière standard, anonymisent les jeux de données en conséquence. Il est toutefois possible de réidentifier une personne ou de divulguer des informations confidentielles en combinant différentes observations dans un jeu de données, même après l'application d'une telle anonymisation.

RISQUE DE RÉIDENTIFICATION ET DE DIVULGATION

Une combinaison d'observations dans un jeu de données peut permettre une réidentification, soit de façon isolée, soit combinée avec une compréhension contextuelle de base. Les techniques avancées d'analyse de données peuvent également permettre d'extraire des informations plus sensibles que celles ressortant d'une analyse de base, ce qui augmente la sensibilité potentielle des microdonnées dans le secteur humanitaire.

Il existe trois formes communément reconnues[3] de risque de divulgation, chacune pouvant se manifester dans les microdonnées humanitaires:

- **Divulgation d'identité:** lorsqu'il est possible d'associer un individu connu à un enregistrement de données diffusées
- **Divulgation d'attribut:** lorsqu'il est possible de déterminer de nouvelles caractéristiques d'un individu sur la base des informations disponibles dans les données diffusées
- **Divulgation inductive:** lorsqu'il est possible de déterminer la valeur d'une caractéristique d'un individu plus précisément avec les données diffusées que cela aurait été possible autrement

CONTRÔLE DE LA DIVULGATION DE DONNÉES STATISTIQUES

Le contrôle de la divulgation de données statistiques (« CDS ») est une technique utilisée en statistique pour évaluer et réduire le risque[4] qu'une personne ou un groupe de personnes soit réidentifié à partir des résultats d'une analyse d'enquête ou de données administratives, ou lors de la diffusion de microdonnées. Cette technique a été utilisée principalement par les bureaux nationaux de statistique et d'autres services statistiques à l'égard des données de recensement.

L'application du CDS a une incidence sur la valeur informative globale ou l'utilité d'un jeu de données, et il est essentiel de trouver un équilibre approprié entre le risque de réidentification et l'utilité de l'information. Pour décider d'un juste équilibre entre les risques et l'utilité, les organisations doivent tenir compte des diverses utilisations possibles d'un jeu de données et du contexte dans lequel les données ont été recueillies.

L'application du CDS pour limiter le risque de divulgation dans les microdonnées comporte trois étapes distinctes:

1. **Évaluation du risque de réidentification**
   La première étape consiste à entreprendre une évaluation du risque de divulgation afin de déterminer la probabilité que la divulgation se produise pour les répondants dans un jeu de données particulier.[5] Savoir si cette probabilité de divulgation (on parle aussi de « pourcentage de risque ») est acceptable pour un jeu de données dépendra du contexte. Par exemple, dans un environnement de conflit, le pourcentage de risque autorisé sera généralement inférieur à celui d'une intervention dans un contexte de catastrophe naturelle.

[3] Pour plus de renseignements sur les risques liés à la divulgation et autres considérations techniques associées à l'évaluation et la gestion de ces risques par le contrôle de la divulgation statistique, consultez le document intitulé **Statistical Disclosure Control for Microdata: A Practice Guide**.

[4] Remarque : Le CDS a pour objet d'empêcher la divulgation d'identité et d'attribut, mais n'est pas spécifiquement conçu pour empêcher la divulgation inductive.

[5] Découvrez comment effectuer une évaluation des risques liés à la divulgation par le biais du parcours pédagogique du Centre sur le sujet : **https://centre.humdata.org/learning-path/disclosure-risk-assessment-overview/**.

2. **Réduction du risque de réidentification**
   L'étape suivante consiste à soumettre le jeu de données au processus de CDS, ce qui, grâce à différentes méthodes d'anonymisation, permettra de réduire le risque de réidentification. Ces méthodes relèvent de l'une de ces deux catégories : les méthodes perturbatives, qui ne suppriment pas les valeurs dans le jeu de données mais qui perturbent (c.-à-d., modifient) les valeurs pour limiter le risque de divulgation en créant une incertitude autour des vraies valeurs ; ou les méthodes non-perturbatives, qui réduisent le détail dans les données par généralisation ou suppression de certaines valeurs (c.-à-d., masquage) sans déformer la structure des données.

3. **Quantification de la perte d'informations**
   La dernière étape consiste à mesurer la perte d'informations résultant de l'application du CDS au jeu de données. L'objectif ici est de comparer la valeur informative du jeu de données d'origine avec la valeur informative finale.

En utilisant le CDS pour évaluer et réduire le risque de divulgation dans les microdonnées, les organisations humanitaires peuvent partager de manière plus responsable les données des enquêtes et des évaluations des besoins afin d'éclairer l'effort global de réponse.

Application du CDS aux données partagées sur HDX[6]

Depuis le début 2018, le Centre for Humanitarian Data (ci-après dénommé « le Centre ») a procédé à une évaluation des risques de 59[7] jeux de données publiés sur la plateforme Humanitarian Data Exchange (HDX). Le risque de divulgation de l'identité des répondants dans 38 de ces jeux de données a été jugé trop élevé pour une diffusion sur HDX. Les contributeurs de 14 de ces jeux de données ont accepté l'application du CDS à leurs données afin de réduire le niveau de risque. L'équipe HDX a appliqué le CDS à ces 14 jeux de données, pour lesquels le niveau de risque a été abaissé à un degré acceptable (c.-à-d. 5 % ou moins[8]). Pour 5 de ces 14 jeux de données, cela signifiait qu'ils pouvaient être rendus publics après anonymisation. Les 9 jeux de données restants ont été soit supprimés, soit partagés en privé sur HDX, tout comme les 24 jeux de données à haut risque pour lesquels le contributeur n'avait pas accepté de mener un CDS. Pour ces 24 jeux de données, de nombreuses organisations ont pris leurs propres mesures pour réduire le risque de réidentification, y compris parfois via la suppression de variables sensibles non essentielles.

APPLICATIONS DU CONTRÔLE DE DIVULGATION DE DONNÉES STATISTIQUES DANS LA GESTION DES DONNÉES HUMANITAIRES

En début d'année 2019, le Centre a interrogé des collaborateurs de sept organisations humanitaires qui effectuent des enquêtes et des évaluations des besoins afin de comprendre les pratiques existantes en matière de gestion des microdonnées. Certaines organisations, telles que le HCR (voir l'étude de cas ci-dessous) adoptent des approches relativement avancées et une expertise interne considérable pour la conduite du CDS sur différentes formes de microdonnées. Toutefois, la plupart des organisations interrogées ont besoin d'un soutien supplémentaire pour mener à bien ce travail.

[6] LPour en savoir plus sur l'utilisation par le Centre du CDS et du processus d'assurance qualité globale pour les données partagées sur HDX, cliquez ici : **https://data.humdata.org/about/hdx-qa-process**.

[7] Au moment de la rédaction du présent document.

[8] Le Centre a récemment ajusté son seuil par défaut de risque acceptable de réidentification de 5 % à 3 %. Le seuil exact pour un jeu de données en particulier est toujours contextuel et défini avec l'organisation qui contribue au jeu de données.

Gestion responsable de la conservation et de la gestion des microdonnées sur les réfugiés
Expérience du HCR

Le HCR recueille régulièrement des données sur les réfugiés et autres populations dans le cadre de son mandat. Ces données servent à évaluer les besoins et les vulnérabilités, à informer les programmes et à mieux cibler l'aide apportée. Bien que ces données n'aient traditionnellement pas été conservées dans des formats ou des emplacements qui pourraient les rendre facilement accessibles pour une utilisation future, le HCR est actuellement en train de créer une bibliothèque de microdonnées interne et une bibliothèque de microdonnées externe. En créant ces référentiels en ligne destinés à permettre à des utilisateurs internes et externes d'accéder aux microdonnées, le HCR vise à permettre une utilisation plus large des données par diverses parties prenantes et à, à partir de maintenant, empêcher le dédoublement dans les activités de collecte de données.

La diffusion publique de microdonnées présente de nombreux avantages potentiels, mais elle comporte également des risques. La diffusion sans mesures appropriées de contrôle de la divulgation peut permettre à des intrus d'identifier les entités (individus ou ménages) dont les données sont partagées, même si des identifiants directs tels que le nom ou l'adresse ont été supprimés. Conformément **à la politique de protection des données personnelles du HCR**, l'identité des personnes concernées doit être protégée et, par conséquent, les jeux de données doivent être dûment rendus anonymes avant de pouvoir être partagés. Les données du HCR sont particulièrement sensibles, car elles concernent des groupes de personnes particulièrement vulnérables, dont la protection est de la plus haute importance.

Pour assurer la protection et la diffusion responsable des microdonnées, le HCR utilise l'application sdcMicro en R pour calculer les risques de réidentification de ces données avant leur publication. Le processus est géré par l'équipe de curation des données du HCR, qui travaille avec les propriétaires des données pour identifier les variables clés, évaluer la sensibilité des données et établir un niveau de risque acceptable pour un ensemble de données particulier. Après l'anonymisation, les données modifiées sont comparées à l'original afin d'évaluer la perte d'information. Si le propriétaire des données juge que certaines variables modifiées sont essentielles pour les utilisateurs des données, les méthodes de contrôle de la divulgation peuvent être ajustées en conséquence. Par exemple, dans le cas de l'Enquête nutritionnelle standardisée élargie (SENS), l'équipe de curation a décidé de ne pas appliquer l'agrégation entre tranches d'âge qui le seraient habituellement, ces tranches étant essentielles pour caractériser la malnutrition selon l'âge en années et en mois pour les enfants. L'équipe a conservé la variable de l'âge mais a exclu la date de naissance et la date de l'enquête. Cela a mené à un scénario de risque acceptable tout en conservant l'aspect utile des données pour les nutritionnistes.

Le HCR continue d'investir dans ce processus en renforçant son équipe de curation et en renforçant l'expertise dans les techniques d'anonymisation au sein de l'organisation. Dans le cadre du plan actuel, la **bibliothèque de microdonnées du HCR** sera entièrement opérationnelle et sera remplie de microdonnées sur les déplacements forcés à la fin de 2019.

En travaillant avec des fournisseurs de données comme REACH (voir l'étude de cas ci-dessous) pour développer un modèle de service de CDS fiable et sécurisé, le Centre vise à soutenir le partage responsable de ces données et à démontrer la valeur de techniques plus robustes en matière d'évaluation des risques de divulgation et d'anonymisation des données. L'exposition à ces techniques aide également les organisations humanitaires à identifier les outils et les méthodes qu'elles peuvent intégrer à leurs propres processus de gestion des données, tout en contribuant à l'ensemble des connaissances au sein du secteur quant à la façon de gérer et de partager de manière plus responsable les microdonnées dans les contextes humanitaires.

Opportunities and challenges to incorporating SDC into an organization's workflow
Experience from REACH

REACH a commencé à explorer le potentiel du CDS en juin 2018, lorsque l'équipe HDX a appliqué pour la première fois le package sdcMicro R à un jeu de données publié sur la plateforme. Parmi les types de jeux de données pour lesquelles l'équipe HDX a appliqué le CDS pour REACH se trouvent les enquêtes auprès des ménages ainsi que les entretiens avec les informateurs clés (et les métadonnées associées). REACH n'a pas encore appliqué le CDS directement, mais est en train d'examiner comment le faireo.

Compte tenu de l'expérience acquise jusqu'à présent, REACH suggère que les organisations intéressées par l'intégration du CDS dans leur flux de travail examinent les questions suivantes :

- S'agit-il de la bonne méthodologie pour vos processus existants de gestion des microdonnées ?
- Dans quelle mesure l'application du CDS réduit-elle la validité et l'utilité des données ?
- Comment l'application du CDS affecte-t-elle la transparence ?
- Comment vous assurer que le personnel ne compte pas trop sur les résultats d'une évaluation des risques de divulgation d'un CDS et qu'il continue à faire preuve d'esprit critique face aux risques potentiels des différents types de données ?

REACH a établi qu'il serait réalisable sur le plan opérationnel de déployer relativement facilement les aspects techniques d'un CDS tant au niveau du siège que sur le terrain. Au siège, cela impliquerait d'exécuter un script sur tous les jeux de données produits ou publiés par REACH. Sur le terrain, cela exigerait des équipes pays qu'elles utilisent sdcMicro ou un outil similaire sur tous les jeux de données produits dans le pays.

Au-delà des aspects techniques du CDS, REACH voit un éventuel défi ou risque d'engorgement dans la composante manuelle du processus, selon laquelle l'équipe doit décider si une technique de contrôle de la divulgation particulière est adaptée, quelles variables supprimer ou sinon brouiller et comment interpréter et communiquer les résultats du processus. Ces décisions prennent du temps et nécessitent une compréhension du contexte auquel les données se rapportent.

À court terme, REACH continuera de collaborer avec l'équipe HDX pour mener des contrôles de la divulgation des données statistiques sur des données d'enquête et d'évaluation avant leur publication sur HDX. Cette expérience permettra à REACH de déterminer la meilleure façon d'intégrer à l'avenir le CDS dans ses propres flux de travail au niveau des pays comme au niveau global.

RECOMMANDATIONS POUR ACCROÎTRE L'UTILISATION DU CDS DANS LES CONTEXTES HUMANITAIRES

Dans cette note d'orientation, le Centre et ses collaborateurs recommandent aux organisations d'investir dans les trois domaines suivants pour réussir l'adoption du CDS :

1. **Sélection de l'outil approprié**
   On trouve en ligne et gratuitement toute une gamme d'outils pour effectuer du CDS. D'après le Centre et les autres organisations humanitaires consultées au cours des recherches de ce dernier, le logiciel le plus utilisé actuellement est **sdcMicro**.

   Le Centre a choisi sdcMicro pour sa capacité d'évolution et de gestion d'importants volumes de données ; et parce qu'il est gratuit et open source. D'autres outils gratuits et open source incluent **µArgus** et **ARX**. Lors de la sélection de l'outil approprié pour effectuer un CDS, les organisations doivent tenir compte de la flexibilité de l'outil dans la sélection des variables clés, de la capacité de l'outil à gérer les grands ensembles de données, de la fonctionnalité intégrée de compromis risque-utilité et de la facilité avec laquelle le personnel serait en mesure de naviguer l'interface utilisateur de l'outil.

2. **Intégration du CDS dans les flux de travail de gestion des données existants**
   La mise en place d'un processus d'application de CDS dans les flux existants est essentielle à l'adoption durable de la méthode. Le CDS exige la participation d'un personnel possédant des connaissances et des compétences différentes, notamment un spécialiste technique pour appliquer les méthodes statistiques et un spécialiste programme ayant une compréhension du contexte des données afin de déterminer l'équilibre acceptable entre les risques et l'utilité. Un flux de travail bien organisé permettra d'améliorer l'efficacité du processus, et d'éviter une interprétation erronée ou une dépendance excessive aux résultats de CDS.

3. **Améliorer la pratique par l'apprentissage continu**
   À mesure que les organisations appliquent le CDS, elles se familiariseront avec la sensibilité des différentes variables clés, le niveau de risque approprié à viser, le niveau acceptable de perte d'information et d'autres considérations qui doivent être prises en compte dans le processus. La tenue d'un registre de chaque application de CDS et la documentation des leçons apprises aideront à affiner le processus au fil du temps. Le partage de ces connaissances en interne et, le cas échéant, avec la communauté humanitaire en général peut favoriser une gestion plus cohérente et responsable des microdonnées dans le secteur.

Dans le cadre de ses efforts visant à soutenir une gestion et un partage plus responsable des données humanitaires sensibles, le Centre améliore son modèle de service[9] pour la conduite de CDS. Ce travail inclut l'introduction d'un processus automatisé de détection des risques pour toutes les données partagées via HDX, qui, lorsqu'il est effectué manuellement, peut prendre plusieurs heures pour les feuilles de calcul volumineuses. Grâce à ce processus, un script s'exécutera sur toutes les données téléchargées sur la plateforme pour identifier les microdonnées et autres formes de données potentiellement sensibles. Les données à haut risque seront envoyées dans un flux de travail dédié, et ces données seront évaluées et, si nécessaire, modifiées par l'intermédiaire du CDS afin de réduire le risque de réidentification avant que les données ne soient partagées plus largement.

Pour en savoir plus sur le travail du Centre sur le CDS, contactez : **centrehumdata@un.org**.

CONTRIBUTEURS : **UNHCR** ET **REACH INITIATIVE**.

Le **Centre for Humanitarian Data** (ci-après dénommé le « Centre »), en collaboration avec des partenaires clés, publie une série de huit notes d'orientation sur la Responsabilité des données dans l'action humanitaire au cours de 2019 et 2020. La série de notes d'orientation fait suite à la publication du **projet de directives opérationnelles sur la Responsabilité des données du Bureau de la coordination des affaires humanitaires des Nations Unies (UNOCHA)** en mars 2019. Par le biais de cette série, le Centre vise à fournir des orientations supplémentaires sur des questions, des processus et des outils spécifiques pour la mise en pratique de la Responsabilité des données. Cette série est rendue possible grâce au généreux soutien de la Direction générale de protection civile et opérations d'aide humanitaire européennes (DG ECHO).

La traduction de ces notes a été facilitée par CartONG grâce au soutien du Ministère français de l'Europe et des Affaires Etrangètres.

[9] Pour en savoir plus sur l'approche du Centre à l'égard du CDS, cliquez ici : **https://humanitarian.atlassian.net/wiki/spaces/HDXKB/pages/1381498881/ Statistical+Disclosure+Control+on+HDX**.