Projet Open OPSE

Quelles **précautions** avant de **publier** des données ?

Rappels & compléments suite à l'atelier juridique, en lien avec les entrepôts de données

Luc DECKER
IRD (service IST – MCST)
data@ird.fr

Quelles précautions avant de publier des données ?

Du point de vue juridique,

l'action de publier est une étape « critique »

quelque soit la plateforme : entrepôt de données, site web, réseau social, forum de discussion, livre, journal

Pourquoi ?

Lorsque des informations ont été rendues publiques (divulguées), il est en général difficile de les rendre à nouveau privées : les personnes qui y ont accédé peuvent les avoir mémorisées, copiées, rediffusées.

Or certaines informations ne peuvent pas être publiées pour diverses raisons, en particulier le risque du causer du tort à une organisation, à des personnes, ou encore des espèces protégées.

Approche « FAIR », générale et pragmatique

- Certains types de données peuvent (tout de même) être déposées dans un entrepôt de données, mais leur accès sera restreint, soumis à diverses conditions. Exemple : remplir un dossier de demande d'accès, qui est approuvé (ou non) par un comité, puis signer un engagement de confidentialité.
- Les métadonnées et documentations associées peuvent en général rester accessibles au public
- Avant tout : respecter les réglementations applicables. En cas de doute, s'abstenir et se renseigner.

« Aussi ouvert que possible, aussi fermé que nécessaire »

- Ne pas céder aux pressions éventuelles des éditeurs et des financeurs des projets : ils délivrent leurs consignes sans connaitre vos données, ni même parfois la déontologie de votre domaine de recherche.

Exemples de données interdites à la publication

à titre indicatif,
se référer à la législation locale applicable

- Données relatives à la sécurité publique, sureté d'un Etat, à sécurité d'un établissement (biens, personnes, informatique…)
- Secret défense
- Secret professionnel, secret des procédés, secret médical, secret de l'instruction, secret bancaire, …



Certains type de données peuvent être publiées sous conditions +/- strictes

se référer à la législation locale applicable

- Données personnelles
  - Données sensibles : origines ethniques; opinions politiques, philosophiques ou religieuses; appartenance syndicale, sexualité…. ➔ *anonymisation, comités d'éthique …*
  - Données de Santé ; Données de la Recherche médicale
- Données protégées par le droit d'auteur (œuvres originales) : textes, figures, photographies … ➔ *autorisation*
- Données qui impliquent un partenaire privé ou étranger
  - provenant de tiers ➔ *autorisation ou licence*
  - générées en collaboration avec des tiers ➔ *convention*
  - rassemblées pour le compte d'un tiers ➔ *contrat*



Certains type de données peuvent être publiées sous conditions +/- strictes

se référer à la législation locale applicable

- Informations pouvant avoir un impact sur la conservation de la biodiversité ➔ *éthique, déontologie*
- Données produites par des laboratoires travaillant sur des thématiques sensibles ➔ *autorisations*
- Données concernant les ressources génétiques et les « connaissances traditionnelles » associées ➔ *Protocole de Nagoya*
- Autres cas (liste non exhaustive)

Anonymisation des données personnelles
Principes généraux

- **Doit être irréversible et ne permettre l'identification d'une personne d'aucune façon, y compris par des méthodes statistiques complexes** : individualisation, corrélation, inférence, croisements avec des sources de données disponibles ailleurs
- Permet de s'affranchir ultérieurement de certaines règles (par exemple, le RGPD), mais pas avant et durant le traitement d'anonymisation
- Utilisation de pseudonymes/codes et de systèmes de cryptage : protège les personnes (déontologie) mais pas accepté en tant que méthode d'anonymisation

Anonymisation des données personnelles : en général, un travail d'expert

- Obligation de résultat : procéder avec rigueur et précaution, ne pas improviser, suivre des directives, contrôler
- Quand bien même on parviendrait à re-identifier seulement 1 personne sur 1000 (par exemple), la méthode est inadéquate

Dé-identification de données = pseudonymisation « avancée », plus simple à mettre en œuvre

Exemple : méthode *HIPAA - Safe Harbor* (USA) ≠ anonymisation

- ✓ Retirer tout types de dates, conserver seulement les années (et combiner tous les âges ≥ 90 ans en une seule catégorie)
- ✓ Pas d'indication géographique précise (noms de lieu de tout types) couvrant moins de 20.000 personnes
- ✓ Aucun élément d'identification : noms; tout types de numéros uniques : téléphone, sécurité sociale, immatriculations, certificats…; photos de visages; emails, coordonnées sur réseaux sociaux **…**
- ✓ Attention au contenu des champs de type « texte », « commentaires »

Pseudonymisation, dé-identification ou anonymisation de données personnelles ? Processus

Données brutes identifiantes (nominatives)
Déontologie : conservées si nécessaire, parfois uniquement sur les sites d'étude, non centralisées

Données pseudonymisées : accessibles aux chercheurs membres du projet
Supprimer (ou encrypter) tous les **identifiants directs**. Déontologie : protection de la confidentialité
~~Nom~~ → *code unique interne à l'étude* (correspondance avec le nom conservée sur site)
~~Adresse~~ ~~Téléphone~~ ~~Email~~ ~~Numéros identifiants uniques~~ ~~(Date de naissance ?)~~ .....

Données dé-identifiées : supprimer ou catégoriser **les identifiants indirects**
Diffusion restreinte possible (si prévue par le consentement des participants), par exemple après étude par un comité de demandes d'accès provenant d'autres chercheurs
~~*code unique interne à l'étude*~~ ~~Toute date précise~~ → année ou temps écoulé ~~Age~~ → classe d'âge
~~Textes libres~~ ~~Village~~ → région ~~Coordonnées GPS précises~~ .....

Données anonymisées « non-personnelles »
Impossibilité totale de réidentifier toute personne, y compris par des méthodes plus avancées : croisement avec d'autres données, déductions logiques, isolation de cas peu fréquents etc….
Vérifier *k-anonymat, l-diversité* … → modifier davantage les données (création de catégories)
~~données encryptées~~ ~~table(s) de correspondance~~

Difficulté technique et expertise croissante

ⓘ Un entrepôt de données peut (éventuellement) accueillir des données dé-identifiées si elles sont configurées «en accès restreint»