RECOMMANDATIONS ET MÉTHODOLOGIES POUR LA CONSTRUCTION D'UN CORPUS TEXTUEL ANNOTÉ

Bahdja Boudoua, Nadia Guiffant, Mathieu Roche, Maguelonne Teisseire, Annelise Tran

RECOMMANDATIONS ET MÉTHODOLOGIES POUR LA CONSTRUCTION D'UN CORPUS TEXTUEL ANNOTÉ

Bahdja Boudoua[1,2], Nadia Guiffant[1,2], Mathieu Roche[1,3], Maguelonne Teisseire[1,2], Annelise Tran[1,2,4]

[1] TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
[2] INRAE, UMR TETIS, Montpellier, France
[3] CIRAD, UMR TETIS, F-34398 Montpellier, France
[4] UMR ASTRE, Univ. Montpellier, CIRAD, INRAE, Montpellier, France

Mot-clés: Guide d'annotation, Données textuelles, Corpus

Résumé

Ce document, fondé sur les retours d'expérience de l'UMR TETIS et la littérature scientifique, propose une méthodologie générique pour la création d'un guide d'annotation et la production de jeux de données textuelles (corpus) annotés. Il couvre les aspects méthodologiques, de stockage, de partage et de valorisation des données, en intégrant des définitions et des exemples pour guider chaque étape, offrant ainsi un cadre complet pour accompagner la constitution et l'exploitation de corpus dans divers contextes de recherche.

Introduction

Les bonnes pratiques pour constituer un corpus annoté couvrent l'ensemble du cycle de vie de la donnée : collecte, traitement, analyse, stockage, partage et réutilisation. Ce guide inclut des recommandations fondées sur les retours d'expérience issus de l'UMR TETIS et sur la bibliographie. Son objectif est de synthétiser le processus de création d'un guide d'annotation, de production de jeux de données textuelles (corpus) et de gestion de leur stockage dans les grandes étapes.

1.1 Définition du corpus annoté

D'après la définition du Centre National de Ressources Textuelles et Lexicales[1] :

- En linguistique, un corpus est en ensemble de textes rassemblés selon un principe de documentation exhaustive, un critère thématique ou exemplaire en vue de leur étude linguistique.

[1] https://www.cnrtl.fr/

- En informatique, un corpus désigne un ensemble de données exploitables dans une expérience d'analyse ou de recherche automatique d'informations.

Annoter un corpus consiste à ajouter une ou plusieurs couches d'interprétation linguistique aux données du corpus.
Les annotations sont réalisées lors de campagnes d'annotation réunissant plusieurs annotateurs humains, plus ou moins experts, à l'aide d'un guide d'annotation.

La constitution et l'évaluation de corpus annotés ainsi que des méthodes automatiques d'annotation suscitent un intérêt grandissant en linguistique et en Traitement Automatique du Langage (TAL ou Natural Language Processing (NLP) en anglais), au vu notamment du développement croissant de l'usage du Machine Learning.

Dans le cadre de ce guide et des exemples fournis, nous adoptons une approche d'annotation à l'échelle du document ou de la phrase. D'autres approches, telles que l'annotation à l'échelle de la phrase ou du mot, sont envisageables. Un exemple de ces différents types d'annotation dans le domaine de l'épidémiologie est illustré dans la figure 1 ci-dessous : l'article de presse choisi comme exemple est successivement annoté à l'échelle du document, de la phrase, puis du mot.


Reuters

By Reuters

August 25, 2018 6:13 PM GMT+2 · Updated 6 years ago

African Swine Fever hits Romania's biggest pigs farm

"We've been focusing on mainland and the virus might have emerged from the waters," he said. Romania has reported hundreds of outbreaks of the disease among **pigs** kept in backyards and smallholdings as well as several large private farms located especially in the south of the country.
About 100,000 pigs have been culled so far.
African swine fever affects **pigs** and **wild boar** and has spread in Eastern Europe in recent years. It does not affect humans.

Figure 1: Extrait d'un article de Reuters. L'article complet est disponible à l'adresse suivante : https://www.reuters.com/article/us-romania-swineflu-pigs-idUSKCN1LA0LR/.

Type d'annotation	**Label / Exemple**
À l'échelle du document	Pertinent (Description de foyers épidémiologiques)
À l'échelle de la phrase (phrase surlignée)	Bilan / conséquence
À l'échelle du mot (Pigs, Boars)	hôtes affectés

Tableau 1: Types d'annotations possibles appliquées à l'exemple illustré en Figure 1.

Matériel et méthode

1.2 Collecte des données :

Pour la constitution d'un corpus, des données existantes sont collectées (textes, portions de textes), très souvent sur Internet. Il est important de pouvoir justifier le choix des données (ici des documents textuels) collectées par rapport aux objectifs de recherche.

1.3 Création du guide d'annotation :

L'élaboration du guide d'annotation suit généralement un processus itératif (Figure 2) (Sabou et al. 2014). Ce processus implique plusieurs phases d'annotation réalisées sur un échantillon par des experts, avec pour objectif de produire un guide précis permettant à des annotateurs, même non-experts, d'annoter les articles sans rencontrer de problèmes d'ambiguïté.

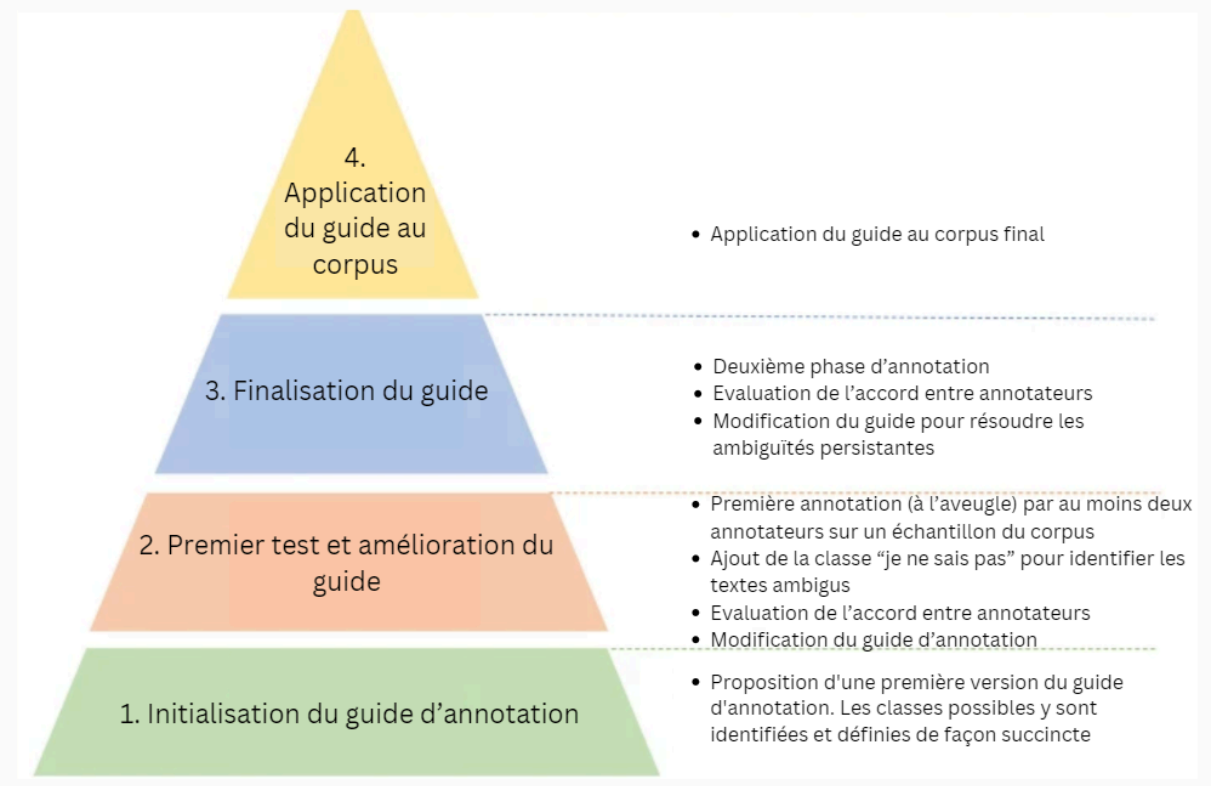


Figure 2 : Processus de création du guide d'annotation (adapté de Valentin et al. 2022)

1.3.1 Élaboration d'une version préliminaire du guide

La première étape consiste à établir une version initiale du guide d'annotation, contenant des définitions succinctes et claires des catégories d'annotation.

Remarque : Il est également conseillé d'inclure une catégorie "Je ne sais pas" afin de mettre en évidence les documents ou phrases ambigus présentant des difficultés d'interprétation.

Cette option permet d'éviter que les annotateurs se sentent contraints de classer des articles (ou phrases/segments) dans des catégories inappropriées.

1.3.2 Phase d'annotation à l'aveugle

Une fois cette version préliminaire en place, un échantillon représentatif du jeu de données à annoter (par exemple de 40 à 50 articles) est sélectionné. Cet échantillon est ensuite annoté de manière indépendante et à l'aveugle par au moins deux experts. Cette étape permet de tester la cohérence des définitions.

Après cette étape, un premier calcul de l'accord inter-annotateur est effectué. Plusieurs méthodes peuvent être utilisées, comme le coefficient Kappa de Cohen (McHugh 2012), qui est approprié lorsque deux annotateurs sont impliqués, et le Kappa de Fleiss (Zheng et al. 2017), qui est utilisé lorsque plus de deux annotateurs participent.

1.3.3 Analyse et discussion des désaccords

Les désaccords observés dans les annotations sont ensuite discutés entre les experts. Cette discussion permet d'identifier les ambiguïtés ou imprécisions dans les définitions des catégories d'annotation. Il n'existe pas de règles strictes ; les ajustements dépendent du jeu de données. Par exemple, il peut être nécessaire d'ajouter de nouvelles catégories, ou de fusionner des classes existantes.

1.3.4 Révision et second tour d'annotation

Après la révision du guide, une nouvelle phase d'annotation peut être réalisée par les mêmes experts. Ce deuxième tour permet de vérifier que les ajustements apportés au guide d'annotation résolvent les problèmes. D'expérience, deux tours d'annotation sont généralement suffisants, mais il est essentiel de se fier à l'accord entre annotateurs pour évaluer l'efficacité de cette démarche et éventuellement organiser d'autres tours.

1.3.5 Finalisation du guide d'annotation

Lorsque l'accord entre annotateurs atteint un niveau jugé satisfaisant, la version finale du guide est établie. Pour le Kappa de Cohen, une valeur de 0,6 est généralement considérée comme un accord modéré, tandis qu'une valeur de 0,9 correspond à un accord parfait. Un seuil minimal de 0,8 est souvent recommandé pour garantir un accord inter-annotateurs acceptable (McHugh, 2012).

Ce guide final sera ensuite utilisé pour annoter l'ensemble du jeu de données.

En somme, ce travail aboutit à la production d'un guide d'annotation qui, une fois appliqué au corpus, permet d'annoter un jeu de données. Ces données doivent ensuite être stockées et organisées afin de faciliter leur exploitation et réutilisation. La section suivante aborde le stockage et le partage des données.

2. Stockage et partage des données :

2.1 Entrepôt des données

Une fois le corpus annoté produit, il est essentiel de le rendre accessible et réutilisable. Pour cela, l'attribution d'un identifiant persistant (PID) ou d'un identifiant unique (DOI) est recommandée, afin de permettre une localisation et une citation fiables des données. Ces identifiants facilitent également le suivi des citations et des réutilisations.

Un entrepôt pérenne et digne de confiance attribue automatiquement un identifiant à chaque jeu de données déposé. Le dépôt de données dans un entrepôt est gratuit. A minima, un entrepôt doit contenir les données annotées ainsi que le guide d'annotation associé. Il est également possible de stocker différentes versions des données, telles que les données brutes et les données annotées, pour assurer une meilleure traçabilité de leur évolution.

Depuis 2021, l'entrepôt Recherche Data Gouv est recommandé par les tutelles ministérielles. Il existe également des entrepôts thématiques dont certains dédiés aux corpus.
Pour l'unité TETIS, il est particulièrement recommandé d'utiliser les plateformes Dataverse (CIRAD)[2] et RechercheDataGouv (INRAE)[3].

Les agents peuvent y déposer les données et renseigner leurs métadonnées sans forcément les publier immédiatement, par exemple dans le cas d'un embargo si les données sont en attente de publication chez un éditeur ou si les producteurs de données estiment que les données doivent encore subir des modifications avant d'être partagées.

Lors de la mise en ligne des données, le choix d'une licence Creative Commons permet de préciser les conditions d'utilisation et vous donne le contrôle sur la manière dont vos données pourront être réutilisées par d'autres. Il existe plusieurs types de licences, allant de la plus permissive à la plus restrictive. Plus d'informations sont disponibles à l'adresse suivante : https://creativecommons.org/share-your-work/cclicenses/

Note : Les données collectées sont protégées par leurs producteurs par des licences ou Conditions Générales d'Utilisation (CGU). Il en existe différentes, qui précisent les droits accordés pour leur réutilisation (accès gratuit ou payant, droit de reproduire, copier, modifier, extraire, communiquer, re-distribuer, publier, etc.). Afin d'exploiter les données en toute légalité, il convient d'être vigilant et de s'assurer que leurs licences ou CGUs permettent bien d'effectuer les traitements souhaités.

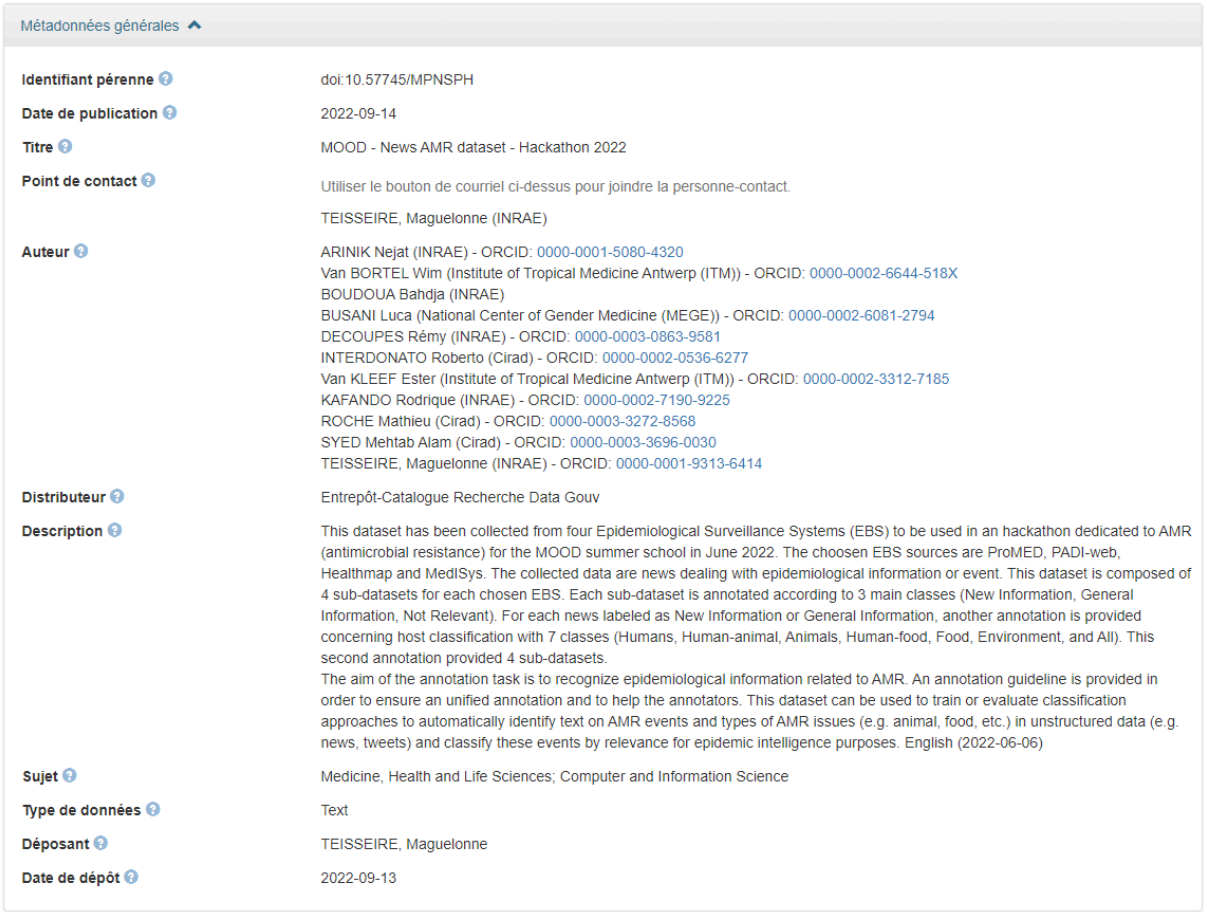
Astuce : Il existe des approches à adopter pour contourner ces restrictions tout en respectant les licences et CGU :

[2] https://dataverse.cirad.fr/
[3] https://entrepot.recherche.data.gouv.fr/dataverse/inrae

à faire	à proscrire
Partager des tweets sans mentionner les identifiants des utilisateurs	Diffuser des tweets dans leur intégralité
Diffuser des extraits de texte	Diffuser des textes complets sans autorisation
Diffuser des URL vers des articles et/ou les fragments des articles (“snippets”)	Diffuser des articles de média en ligne sans autorisation

2.2 Documenter les données

Les métadonnées sont les données ou informations servant à décrire les données (ex : titre, date, auteur, contexte du projet, mots-clés, …).
Elles permettent de comprendre les données, d’en connaître l’origine, elles sont essentielles pour faciliter l’utilisation des données car elles sont souvent la seule forme de communication entre les étapes de production des données et d’analyse secondaire. Elles doivent donc être compréhensibles et fournir toutes les informations utiles à l’analyse et à la réutilisation des données.
La figure ci-dessous présente les métadonnées générales associées à un jeu de données produit dans le cadre du MOOD - News AMR dataset - Hackathon 2022 (Arınık et al. 2022).



Métadonnées générales

Identifiant pérenne	doi:10.57745/MPNSPH
Date de publication	2022-09-14
Titre	MOOD - News AMR dataset - Hackathon 2022
Point de contact	Utiliser le bouton de courriel ci-dessus pour joindre la personne-contact. TEISSEIRE, Maguelonne (INRAE)
Auteur	ARINIK Nejat (INRAE) - ORCID: 0000-0001-5080-4320 Van BORTEL Wim (Institute of Tropical Medicine Antwerp (ITM)) - ORCID: 0000-0002-6644-518X BOUDOUA Bahdja (INRAE) BUSANI Luca (National Center of Gender Medicine (MEGE)) - ORCID: 0000-0002-6081-2794 DECOUPES Rémy (INRAE) - ORCID: 0000-0003-0863-9581 INTERDONATO Roberto (Cirad) - ORCID: 0000-0002-0536-6277 Van KLEEF Ester (Institute of Tropical Medicine Antwerp (ITM)) - ORCID: 0000-0002-3312-7185 KAFANDO Rodrique (INRAE) - ORCID: 0000-0002-7190-9225 ROCHE Mathieu (Cirad) - ORCID: 0000-0003-3272-8568 SYED Mehtab Alam (Cirad) - ORCID: 0000-0003-3696-0030 TEISSEIRE, Maguelonne (INRAE) - ORCID: 0000-0001-9313-6414
Distributeur	Entrepôt-Catalogue Recherche Data Gouv
Description	This dataset has been collected from four Epidemiological Surveillance Systems (EBS) to be used in an hackathon dedicated to AMR (antimicrobial resistance) for the MOOD summer school in June 2022. The choosen EBS sources are ProMED, PADI-web, Healthmap and MedISys. The collected data are news dealing with epidemiological information or event. This dataset is composed of 4 sub-datasets for each chosen EBS. Each sub-dataset is annotated according to 3 main classes (New Information, General Information, Not Relevant). For each news labeled as New Information or General Information, another annotation is provided concerning host classification with 7 classes (Humans, Human-animal, Animals, Human-food, Food, Environment, and All). This second annotation provided 4 sub-datasets. The aim of the annotation task is to recognize epidemiological information related to AMR. An annotation guideline is provided in order to ensure an unified annotation and to help the annotators. This dataset can be used to train or evaluate classification approaches to automatically identify text on AMR events and types of AMR issues (e.g. animal, food, etc.) in unstructured data (e.g. news, tweets) and classify these events by relevance for epidemic intelligence purposes. English (2022-06-06)
Sujet	Medicine, Health and Life Sciences; Computer and Information Science
Type de données	Text
Déposant	TEISSEIRE, Maguelonne
Date de dépôt	2022-09-13

Figure 3 : Métadonnées générales associées au jeu de données produit dans le cadre du MOOD - News AMR dataset -Hackathon 2022.

⚠ Pensez à la documentation qui serait nécessaire pour permettre une réutilisation des vos données.
Il peut s'agir notamment de l'information sur la méthodologie utilisée pour constituer le corpus, sur les procédures et méthodes d'analyse utilisées, sur la définition des variables (dictionnaire des variables), des unités de mesure, etc.

Un stockage approprié garantit non seulement la sécurité et la pérennité des données, mais facilite également leur accessibilité et leur réutilisation future. Pour assurer une gestion optimale des données produites, il est recommandé de suivre les principes **FAIR** (**Findable, Accessible, Interoperable, Reusable**), qui définissent les bonnes pratiques de gestion des données.

3. Partage et Valorisation des Données de Recherche

La publication sous forme de **data paper** est une option privilégiée et recommandée pour partager les données de manière formelle et informer la communauté scientifique de leur disponibilité.
Un **data paper** est un article scientifique destiné à décrire et documenter un jeu de données. Publié en libre accès dans une revue scientifique classique ou un **data journal** (une revue dédiée exclusivement aux data papers), il se distingue par une structure spécifique, mettant l'accent sur **la méthodologie de collecte et les analyses techniques qui valident la qualité des données**.
Plusieurs ressources, comme la base « Où Publier »[4] du CIRAD, répertorient des revues spécialisées dans les data papers pour faciliter leur diffusion dans la communauté scientifique.

Exemples de datapapers produits dans l'UMR TETIS

1. Valentin, S., Arsevska, E., Vilain, A. *et al.* Elaboration of a new framework for fine-grained epidemiological annotation. *Sci Data* **9**, 655 (2022). https://doi.org/10.1038/s41597-022-01743-2

2. Arınık, N., Van Bortel, W., Boudoua, B., Busani, L., Decoupes, R., Interdonato, R. Teisseire, M. (2023). An annotated dataset for event-based surveillance of antimicrobial resistance. *Data in Brief*, *46*

3. Koptelov, M., Holveck, M., Cremilleux, B. *et al.* A manually annotated corpus in French for the study of urbanization and the natural risk prevention. *Sci Data* **10**, 818 (2023). https://doi.org/10.1038/s41597-023-02705-y

[4] https://coop-ist.cirad.fr/fr/gerer-des-donnees/publier-un-data-paper/4-choisir-la-revue

4. Holveck, Margaux; Koptelov, Maksim; Roche, Mathieu; Teisseire, Maguelonne, 2023, "Lisez_Moi.pdf", *Segments textuels - Textual Segments - Hérelles Project*, https://doi.org/10.57745/KT6JAB

5. Boudoua, El Bahdja; Richard, Manon; Roche, Mathieu; Teisseire, Maguelonne; Tran, Annelise, 2023, "AI_Annotation Guideline.pdf", *Annotated datasets from PADI-web for event-based surveillance of Avian Influenza, African Swine Fever, and West-Nile Virus Disease*, https://doi.org/10.57745/X3IZT3

Références:

1. Arınık, N., Van Bortel, W., Boudoua, B., Busani, L., Decoupes, R., Interdonato, R. Teisseire, M. (2023). An annotated dataset for event-based surveillance of antimicrobial resistance. *Data in Brief*, *46*
2. McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-282.
3. Sabou, M., Bontcheva, K., Derczynski, L., & Scharl, A. (2014, May). Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC* (pp. 859-866).
4. Valentin, S., Arsevska, E., Vilain, A. *et al.* Elaboration of a new framework for fine-grained epidemiological annotation. *Sci Data* **9**, 655 (2022).
5. Xie, Z., Gadepalli, C., & Cheetham, B. M. (2017). Reformulation and generalisation of the Cohen and Fleiss kappas. *LIFE: International Journal of Health and Life-Sciences*, *3*(3), 1-15.

Annexe : Points de vigilance juridique et éthique

Pour rappel, la Charte française de déontologie des métiers de la recherche[5] impose entre autres les principes suivants :

- liberté d'expression
- indépendance de la recherche
- reproductibilité des recherches (citation des références, accès aux données brutes ET travaillées)
- limitation du stockage à l'essentiel (minimisation de l'empreinte écologique)

Aussi, certaines données ne pourront être diffusées : données scientifiques protégées ou à risques (sécurité état, sécurité des populations, etc), données liées à l'intelligence économique (secret industriel et commercial), données soumises au secret statistique.

5 https://comite-ethique.cnrs.fr/wp-content/uploads/2020/01/2015_Charte_nationale_d%C3%A9ontologie_190613.pdf

L’accès à ces données peut néanmoins être accordé à des personnes autorisées, à condition de stocker ces données sur un espace sécurisé dont l’accès est contrôlé, et de spécifier clairement les restrictions d’usage.

Pour plus d'informations sur les aspects juridiques et éthiques de la gestion des données dans la recherche, consultez ce lien : https://doranum.fr/aspects-juridiques-ethiques/

Par ailleurs, pour les spécificités concernant les données de santé, vous pouvez consulter les recommandations de la CNIL à cette adresse :
https://www.cnil.fr/fr/quelles-formalites-pour-les-traitements-de-donnees-de-sante

Remerciements:

Ce travail a été réalisé dans le cadre du projet MOOD (MOnitoring Outbreak events for Disease surveillance in a data science context. https://mood-h2020.eu/), financé par le programme Horizon 2020 de l'Union européenne (Grant Agreement MOOD n° 874850).