



DATA  
INFRASTRUCTURE

*Développer  
la culture des données*



Sébastien Oliveau, directeur de l'IR\* Progedo  
Maître de conférences HDR à l'université d'Aix-Marseille



13 - 16 mars 2023

**Love Data Week 2023**

Université de Lorraine



# Qu'est-ce qu'une donnée ?

# Qu'est-ce qu'une donnée?



Matériaux: ce que l'on récolte sur le terrain



Donnée : matériau mis en forme (et forme spécifique)

Aujourd'hui, la donnée est implicitement numérique



Donnée secondaire ? Donnée issue du traitement d'une donnée existante (données agrégées par exemple)

# Qu'est-ce qu'une donnée pour PROGEDO?

## Qualifier la donnée

- Quantitative : mesure possible
- Qualitative : décrit les caractéristiques
- Textuelle : un champ à part ?

## Les données dans le cadre de PROGEDO:

- issues d'enquêtes (ou de registres administratifs)
- quantitatives
- représentatives
- (il peut y avoir des exceptions, c'est le principe des règles)

« statistique »

## Enjeux particuliers autour de la microdonnée



DATA  
INFRASTRUCTURE

*Développer  
la culture des données*



**Produire, gérer, informer, accompagner**



DATA  
INFRASTRUCTURE



European  
*Values*  
Study



# Enquêtes internationales



DATA  
INFRASTRUCTURE



POPP

Océrisation des  
recensements



Des dispositifs nationaux





DATA  
INFRASTRUCTURE

*Développer  
la culture des données*



Produire, **gérer**, informer, accompagner



# Gérer les accès



L'ICPSR conserve et diffuse les données de plus de **9.500 enquêtes** en sciences sociales et comportementales. Il héberge 21 collections spécialisées de données dans les domaines de l'éducation, du vieillissement, de la justice, etc.



Le LIS acquiert des ensembles de données sur le revenu, la richesse, l'emploi et les données démographiques de nombreux pays à revenu élevé et intermédiaire, les harmonise pour permettre des comparaisons transnationales et les rend accessibles au public dans deux bases de données : la **Luxembourg Income Study Database (LIS)** et la **Luxembourg Wealth Study Database (LWS)**.

Diffuser mais aussi...  
protéger les données

# Les données en SHS

## Cadre légal français

Loi pour une République Numérique

Règlement Général pour la Protection des Données

## Implications

Diffusion obligatoire des données publiques

Protection des données individuelles

## Application

Développement de l'Open Data

Développement de procédures d'accès restreints

## Accès aux données

nakala; recherche.data.gouv; opendata.gouv.fr;  
zenodo; etc.

Quetelet-Progedo-Diffusion; CASD; Health Data Hub;  
etc.

Rôle des DPD (DPO) et des administrateurs de données; exception de recherche

# Protéger les données personnelles

## Méthodes

protéger l'accès

pseudonymiser

Rendre anonyme

## Implications

Environnements  
adaptés

Potentiellement  
réidentifiant

Perte d'information

# Protéger l'accès

**Ordinateurs non connectés**  
dans les locaux du  
producteur

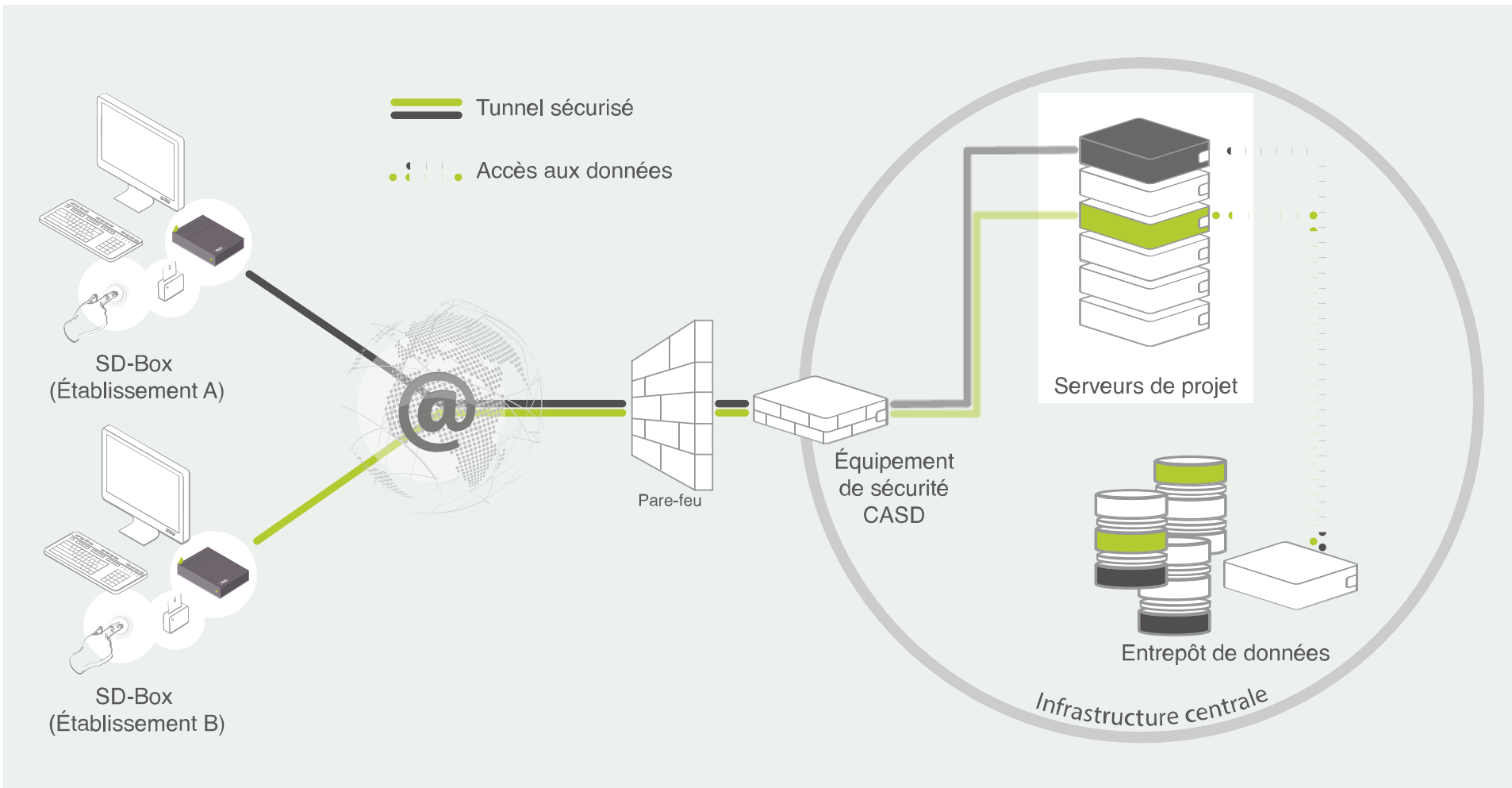
Le chercheur se déplace  
physiquement dans les  
locaux du producteur et  
accède à une machine où il  
peut travailler et  
imprimer/exporter ses  
résultats

**Exécution à distance**  
(*remote execution*)

Le chercheur envoie une  
ligne de commande sur un  
serveur distant qui opère  
sans donner accès aux  
données et renvoie les  
résultats

**Accès sécurisé à distance**  
(« bulles » informatiques)

Le chercheur se connecte  
sur un terminal qui lui  
donne accès aux données  
et à un environnement  
logiciel. Il peut exporter ses  
résultats.



# Protéger l'accès

**Ordinateurs non connectés**  
dans les locaux du  
producteur

**Exécution à distance**  
(*remote execution*)

**Accès sécurisé à distance**  
(« bulles » informatiques)

Nécessité de se déplacer  
Inégalités d'accès

Moindre convivialité  
Moindre souplesse  
(difficile de proposer des choix  
de logiciels en dehors de ce qui  
est proposé, idem pour les  
données)

Environnements adaptés  
Coûts



# Pseudonymiser les données

## Les fichiers « Production et recherche »

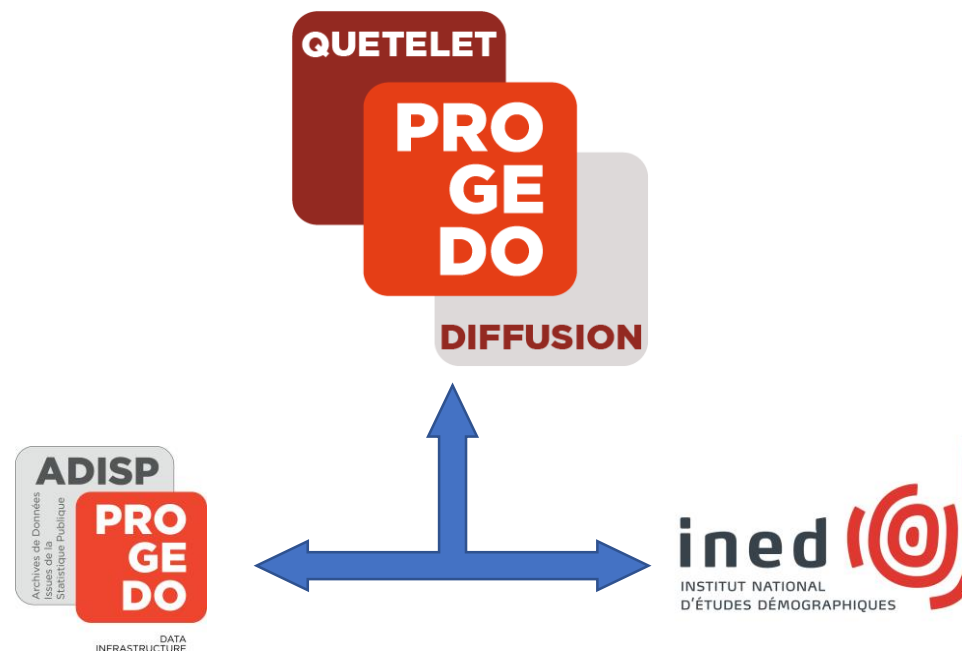
En termes de détails, les fichiers "FPR" offrent un niveau intermédiaire d'information entre les données standards anonymisées (plus agrégées) et les données confidentielles plus détaillées.

Les données pseudonymisées sont considérées comme potentiellement ré-identifiantes. A ce titre, elle nécessite une autorisation administrative auprès du **Comité du secret statistique**

La procédure d'habilitation auprès du Comité du Secret Statistique est gérée par l'intermédiaire des équipes de Quetelet-Progedo-Diffusion. L'accès aux données se fait ensuite depuis l'ordinateur du chercheur.

# Gérer et diffuser les données

<https://data.progedo.fr>



*PROGEDO assure le financement de Quetelet-Progedo-Diffusion et la coordination scientifique des équipes membres*

*Huma-Num assure l'accueil de l'infrastructure informatique de PROGEDO sur ses serveurs*



Institut national de la statistique  
et des études économiques

Mesurer pour comprendre

Coordonne

**16 services statistiques  
ministériels (SSM)**

**Service de la Statistique Publique**

Autorise l'accès  
aux données



Comité du Secret Statistique

concourt à l'établissement du programme de la  
statistique publique pour l'année suivante

organise la concertation entre les producteurs  
et les utilisateurs de la statistique publique

Examine les enquêtes



**Comité du label de la statistique publique**

Veille à l'indépendance professionnelle  
dans la conception, la production et la  
diffusion des statistiques publiques

**L'Autorité de la  
statistique publique**



**Conseil national  
de l'information statistique**

sollicite

sollicite



**DEPS**  
Ministère de la Culture

**DEPP**  
MINISTÈRE  
DE L'ÉDUCATION  
NATIONALE

**DSED**  
[interieur.gouv.fr](http://interieur.gouv.fr)  
MINISTÈRE DE L'INTÉRIEUR

**SIES**  
Ministère de  
l'Enseignement Supérieur



**MEOS**  
MINISTÈRE  
DES SPORTS

**SOeS**  
Ministère de la Transition  
écologique et solidaire



**CEPREMAP**  
CENTRE POUR LA RECHERCHE ÉCONOMIQUE ET SES APPLICATIONS



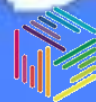
+ les données issues de la recherche



cessda

Coordonner et harmoniser  
les pratiques au niveau européen

APIS



UK Data Service



Soda

gesis

Leibniz-Institut  
für Sozialwissenschaften



ČSDA

SASD

slovenský  
archív  
sociálnych  
údají



TÁRKI



FORS  
explore.understand.share.



UNIDATA  
Bicocca Data Archive



SOCIAL  
DATA  
NETWORK



 **eosc** | **FAIRCORE4EOSC**  
Core Components Supporting a FAIR EOSC



**FAIR-IMPACT**  
Expanding FAIR solutions across EOSC







DATA  
INFRASTRUCTURE

*Développer  
la culture des données*



Produire, gérer, **informer**, accompagner







DATA  
INFRASTRUCTURE

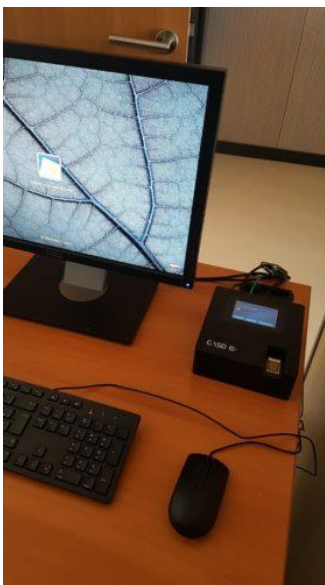
*Développer  
la culture des données*



Produire, gérer, informer, **accompagner**

# Les Plateformes Universitaires de Données

Développer la culture des données au plus près des acteurs de la recherche



## • Accompagnement personnalisé de projets

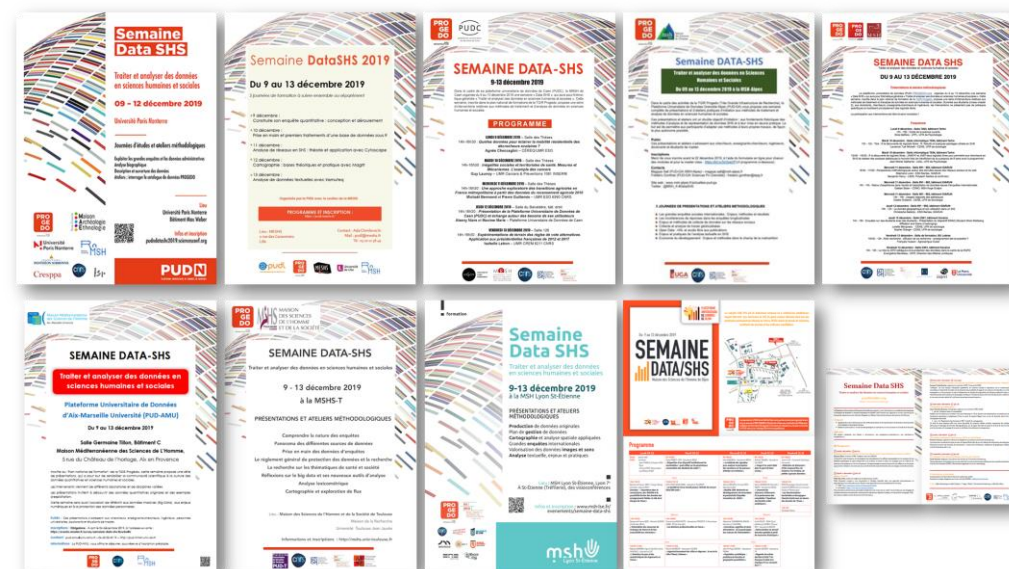
- Recherche de données
- Conseils méthodologiques
- Aide au traitement

## • Formations

- Conseils et organisations
- Formations sur mesure
- Intervention dans les cursus L, M, D

## • Organisation d'évènements scientifiques

- Séminaires
- Semaine DATA SHS 2<sup>ème</sup> semaine de décembre

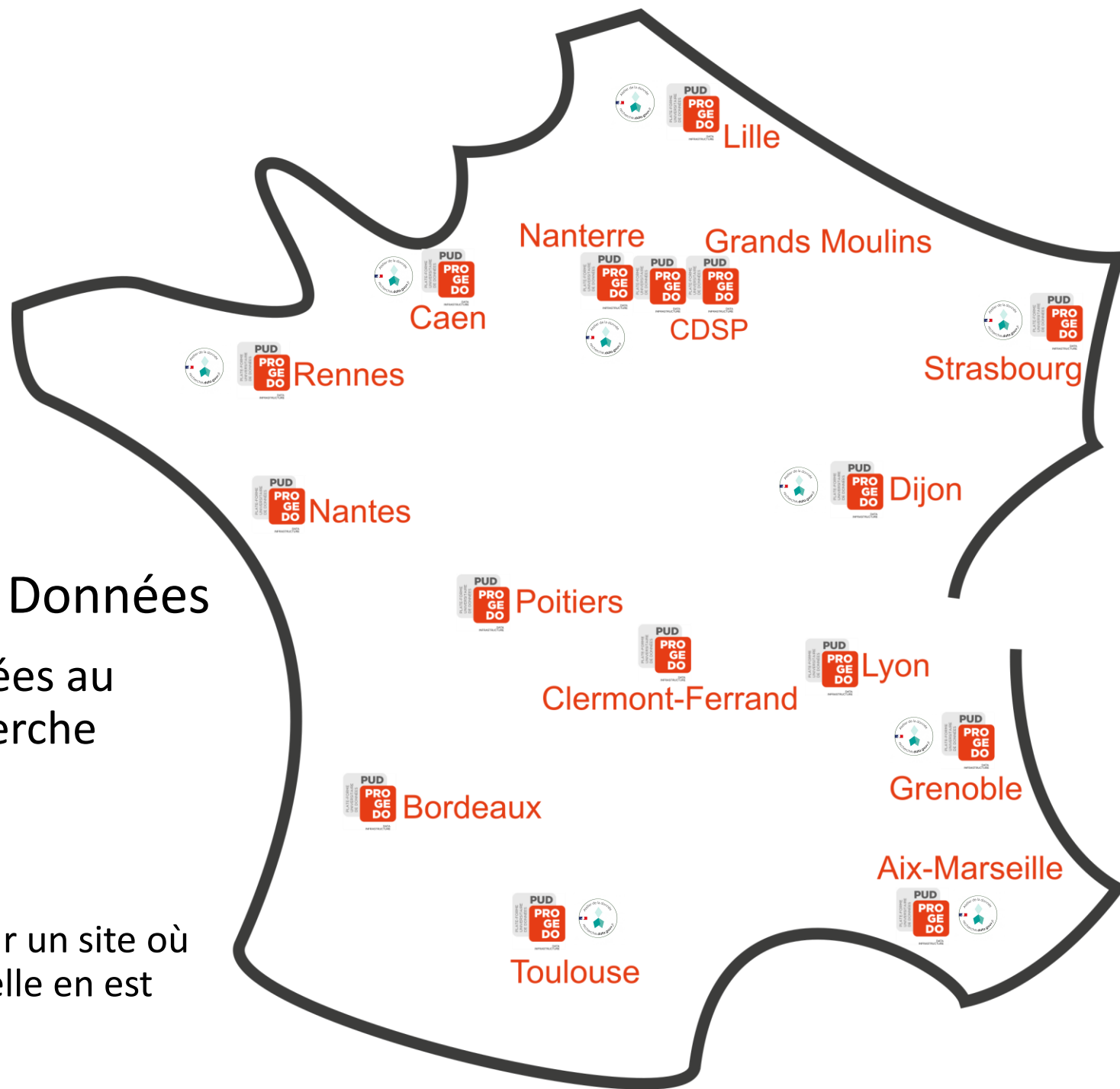


## Les Plateformes Universitaires de Données

Développer la culture des données au  
plus près des acteurs de la recherche



Lorsqu'une PUD est présente sur un site où  
existe un atelier de la donnée, elle en est  
partie prenante





DATA  
INFRASTRUCTURE

# Politique Française des données en SHS

Méthodes  
&  
Statistiques  
Publiques

Opinion et  
enquêtes  
sociopolitiques

Santé

Données  
historiques  
économiques  
et financières

Population  
et parcours  
de vie

Coordination des représentants des  
chercheurs au CCS et au comité du label

QUETELET

PRO  
GE  
DO

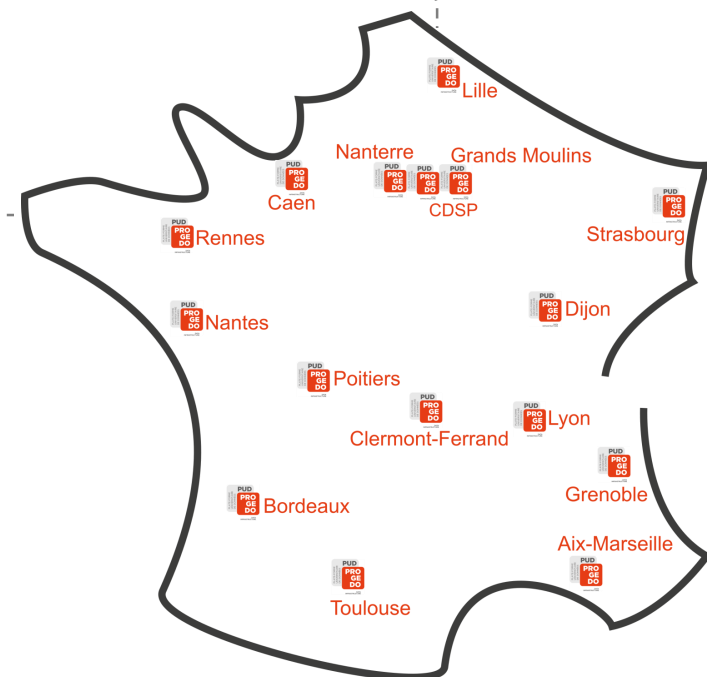
DIFFUSION



## Plateformes Universitaires de Données



et Universités partenaires







**PRO  
GE  
DO**

DATA  
INFRASTRUCTURE

*Développer  
la culture des données*



Merci à vous  
<http://progedo.fr>



13 - 16 mars 2023

**Love Data Week 2023**

Université de Lorraine

