



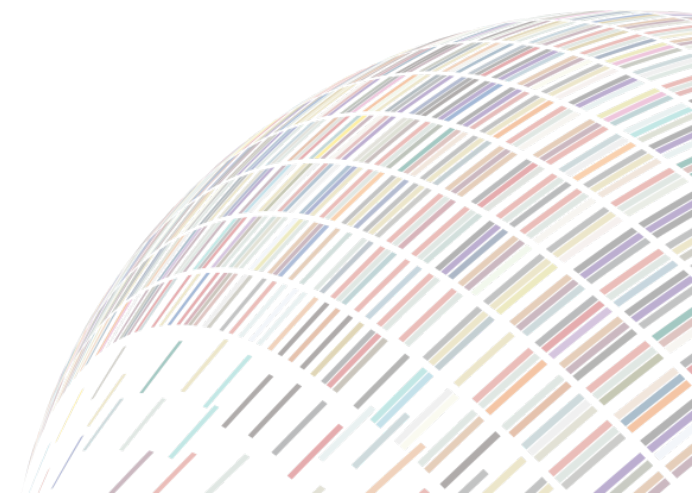
DATA  
INFRASTRUCTURE

*Développer  
la culture des données*

# Préparer ses données et métadonnées

23 novembre 2021

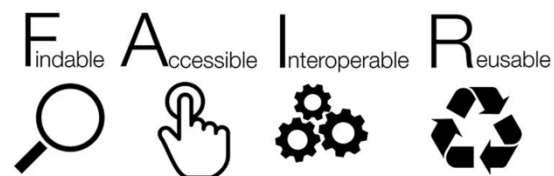
**Lorraine Adam (Ingénieure d'étude à l'ADISP - Progedo)**  
**Erik Zolotoukhine (Ingénieur d'étude à l'ADISP - Progedo)**



# Des métadonnées FAIR

**Objectif : savoir ce que contiennent les données et comment elles ont été produites**

C'est ce qui permet leur réutilisation.

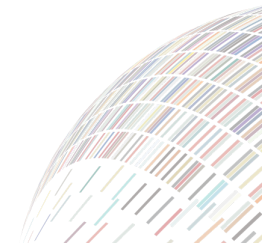


Démarche « FAIR » (**F**aciles à trouver, **A**ccessibles, **I**nteropérable et **R**éutilisables)

Utilisation d'un standard utilisé dans les SHS : DDI.

Concrètement ça veut dire avoir :

- un résumé explicatif, qui précise le producteur, le champ couvert...
- des éléments méthodologiques tels que le mode de collecte, la méthode d'échantillonnage...
- un dictionnaire des variables...
- + tout autre élément utile



# Importance de la doc. pour l'analyse secondaire

## ➤ En amont ...

Prendre connaissance  
du mode de production  
des données :

- ✓ problématique
- ✓ champ de l'enquête
- ✓ population concernée
- ✓ plan de sondage / pondérations
- ✓ ...

Lire les supports de collecte  
(questionnaires, formulaires  
administratifs ...) :

- ✓ libellé des questions / rubriques
- ✓ filtres éventuels
- ✓ consignes aux enquêteurs
- ✓ ...

Etudier les variables diffusées  
(dictionnaire des variables) :

- ✓ variables brutes / reconstruites
- ✓ variables brutes ou en tranches
- ✓ niveau de détail / d'agrégation
- ✓ nomenclatures utilisées
- ✓ ...

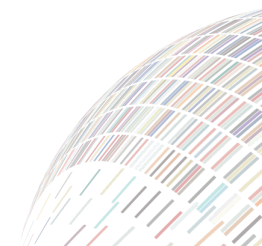
## ➤ Au cours de la recherche et à l'issue (*publications, communications, conférences...*) ...

Contextualiser les résultats de recherche :

- ✓ description de la source utilisée
- ✓ champ de l'enquête / population étudiée  
(*effectifs, redressement éventuel*)
- ✓ mode de recueil des données
- ✓ ...

Afficher les éventuelles limites de l'étude :

- ✓ qualité des données  
(*corrections, valeurs manquantes*)
- ✓ effectifs faibles ou insuffisants
- ✓ représentativité géographique
- ✓ ...



# Partager ses données de recherche

## Les données

Réfléchir aux données partageables / que l'on souhaite partager ...

- variables brutes seules vs. ajout de variables construites a posteriori ?
- suppression de variables de « mauvaise » qualité / quasiment inexploitable ?
- appariement avec des données externes de cadrage ?

... dans le respect des réglementations (RGPD, secret statistique...)

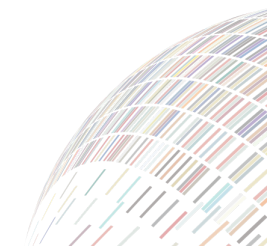
Déterminer les formats des fichiers de données

## Les métadonnées associées

- Données en production : veiller à documenter pendant le projet (à l'aide du PGD)
- Données déjà produites : reconstituer / reformaliser les métadonnées

Constituer les documents annexes utiles

Si possible, respecter la norme internationale DDI (de documentation en SHS)



# Préparation des données

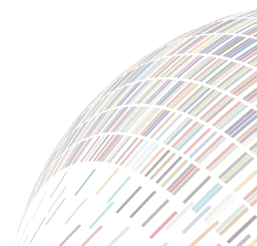
On ne peut pas partager le fichier d'origine, il y a un travail à faire en amont pour partager des données de « qualité » et qui respectent la réglementation.

- **Apurement** des fichiers (nettoyage, valeurs aberrantes, données manquantes, ...)
- Attention portée à la **possible ré-identification** par croisement, appariement... d'où nécessité d'agrégation de certaines variables, des valeurs extrêmes, ...
- Choix entre production d'un fichier anonyme ou pseudonymisation (« DPR » et « FPR »)

L'anonymisation : rendre impossible toute identification de la personne par quelque moyen que ce soit et de manière irréversible.

La pseudonymisation : faire qu'on ne puisse plus attribuer les données relatives à une personne physique sans information supplémentaire.

<https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles>



# Pseudonymisation

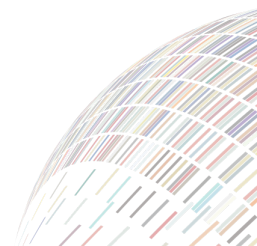
Objectif : éviter qu'on puisse reconnaître quelqu'un et apprendre quelque chose sur lui.

L'important est de préserver la vie privée des enquêtés : peut-on apprendre quelque chose sur un individu qu'on ne pourrait pas savoir sans cette base de données ?

Cela demande de **supprimer les éléments d'identification directe** (nom, adresse, code insee...) mais aussi les **valeurs rares** qui pourraient permettre une réidentification des personnes (il ne faut pas pouvoir isoler un individu dans la base).

Attention aux questions « ouvertes » et aux données « sensibles ».

Il n'y a pas « 1 méthode ».  
Il s'agit de faire des choix sur ce qui est le plus important à conserver  
et ce qui peut être moins précis pour respecter le RGPD.



# Les recommandations des institutions

De l'Insee pour la production de fichiers pseudonymisés

## ➤ **Suppressions...**

- ... des variables de gestion (*adressage, enquêteurs, terrain ...*)
- ... des identifiants individuels directs / indirects (*NIR / SIRENE, noms, prénoms,...*)
- ... du jour de naissance (*conservation de l'année, et du mois si nécessaire*)
- ... (de la plupart) des libellés en clair (*profession / nationalité détaillée, autres à préciser, ...*) ;  
prévoir le recodage de ces questions en nomenclatures plus agrégées (*pays en 5 à 10 codes...*)

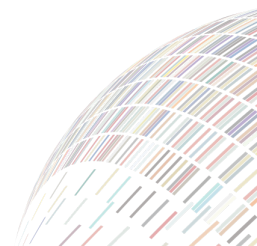
## ➤ **Agrégations** de variables fines telles que géographie, profession, diplôme ...

Notamment, privilégier le département à la commune et la PCS à 2 chiffres, sauf nécessité contraire ; possibilité d'en conserver une si l'autre est agrégée (*exemple : si la PCS est une variable d'intérêt (cf enquête emploi et RP), conservation de la PCS à 4 chiffre vs. géographie à la région*)

## ➤ Au delà, les producteurs adaptent ces règles à leur source, et ajoutent éventuellement des restrictions supplémentaires ...

De l'InSHS : guide pour la recherche

*Les sciences humaines et sociales et la protection des données à caractère personnel dans le contexte de la science ouverte :* [https://www.inshs.cnrs.fr/sites/institut\\_inshs/files/pdf/Guide\\_rgpd\\_2021.pdf](https://www.inshs.cnrs.fr/sites/institut_inshs/files/pdf/Guide_rgpd_2021.pdf)





# Préparation de la documentation

## Les "indispensables" ...

## Le dictionnaire des codes / des variables

### OFFRE

A reçu une offre d'emploi ou de formation de Pôle Emploi ou d'un opérateur de placement

Vide	Sans objet (non inscrit auprès de Pôle Emploi ou d'un opérateur de placement, OFFICC distinct de 1)
1	Oui
2	Non

Nouvelle variable en 2013.

(Cette variable est issue d'un redressement de la réponse à la question OFFRE (A51) du questionnaire ; le redressement consiste, en réinterrogation, à forcer la variable OFFRE à 1 lorsque l'enquêté déclare n'avoir pas reçu d'offre d'emploi depuis la précédente réalisation de l'enquête tout en ayant reçu une antérieurement).

### OPA

Orientation des productions agricoles, pour les exploitants agricoles

Vide	Sans objet (EXPLAG distinct de 1)
1	Polyculture (culture des terres labourables)
2	Maraîchage ou horticulture
3	Vigne ou arbres fruitiers
4	Élevage d'herbivores (bovins, ovins, ...)
5	Élevage de granivores (volailles, porcins, ...)
6	Polyculture - élevage
7	Élevage d'herbivores et de granivores
8	Autres

Nouvelle variable en 2013.

Attention, comme cette variable est nouvelle en 2013 et qu'elle est partiellement construite à partir d'une donnée remontée du questionnaire d'une précédente interrogation, cette variable n'est pleinement renseignée sur l'ensemble des six vagues d'enquête (RGA = 1,...6) qu'à partir du T2 2014.

### P

Profession principale des actifs occupés (PCS 4 chiffres)

cf	voir partie Nomenclatures des variables
Vide	Sans objet (personnes non actives occupées)
0000	Non renseigné

Variable codée à partir du libellé de profession déclaré lors de l'enquête.

### PACS

Individu pacsé

Vide	Sans objet (AGE < 18 ou MATRI = 2) ou non renseigné
1	Oui
2	Non

Nouvelle variable en 2013.

Nota Bene : Du fait de la nouvelle enquête en 2013 et du protocole de collecte en vigueur, cette variable n'est que partiellement renseignée aux T1, T2, T3 et T4 2013 pour les individus des ménages composés exclusivement d'inactifs de 65 ans et plus en réinterrogation intermédiaire (repérables par la condition RES = 16).

Fichier Accueil Insertion Mise en page Formules Données Révision						
Couper Copier Reproduire la mise en forme Presse-papiers Police						
J60						
A	B	C	D	E	F	
1	NOM DE LA VARIABLE	FICHE MENAGE				
2	MP4	JOUR DES DEPLACEMENTS				
3		1 LUNDI				
4		2 MARDI				
5		3 MERCREDI				
6		4 JEUDI				
7		5 VENDREDI				
8	P3_PREF	OCCUPATION PRINCIPALE DE LA PERSONNE DE REFERENCE				
9	MP17	OCCUPATION PRINCIPALE DE LA PERSONNE INTERROGEE POUR				
10		1 TRAVAIL A PLEIN TEMPS				
11		2 TRAVAIL A TEMPS PARTIEL				
12		3 APPRENTISSAGE, FORMATION, STAGE				
13		4 ETUDIANT				
14		5 SCOLAIRE, JUSQU'AU BAC				
15		6 CHOMEUR, CHERCHEUR D'EMPLOI				
16		7 RETRAITE				
17		8 RESTE AU FOYER				
18		9 AUTRE (et non réponse)				
19	MP16	SEXE DE LA PERSONNE INTERROGEE POUR LE QUESTIONNAIRE				
20		1 MASCULIN				
21		2 FEMININ				
22	M1	TYPE D'HABITAT				
23		1 INDIVIDUEL ISOLE				
24		2 INDIVIDUEL ACCOLE				
25		3 PETIT COLLECTIF (jusqu'à 3 étages)				
26		4 GRAND COLLECTIF (plus de 3 étages)				
27		5 AUTRES				
28	M2	TYPE D'OCCUPATION DU LOGEMENT				
29		1 PROPRIETAIRE				
30		2 ACCEDANT A LA PROPRIETE				
31		3 LOCATAIRE HLM				
32		4 AUTRE LOCATAIRE				
33		5 LOGE GRATUITEMENT				

- 54 -

CODE DR : Département de Résidence

- Bandes BI-FL et FL-RM (échelon géographique)
- Sources de l'Information : col. 2-3 de la carte maîtresse communale
- 91 postes possibles : 01 à 90, 99.

Le poste 99 est un chiffrement particulier pour les militaires d'Afrique du Nord sans résidence personnelle en France dont les BI ont été exploités à Montpellier.

Correspondance Département Région de Programme

Code du départ-lement	Département	N° de code de la RP correspondant	Code du départ-lement	Département	N° de code de la RP correspondant	Code du départ-lement	Département	N° de code de la RP correspondant
01	Ain	02	31	Haute Garonne	73	61	Orne	51
02	Alsace	22	32	Gers	73	62	Pas de Calais	31
03	Allier	52	33	Gironde	72	63	Puy-de-Dôme	62
04	Alpes Alpes	32	34	Hérault	91	64	Saône Pyrénées	72
05	Hautes Alpes	92	35	Ille et Vilaine	53	65	Hautes Pyrénées	73
06	Alpes Maritimes	92	36	Isère	24	66	Pyrénées Orientales	91
07	Ardèche	82	37	Isère et Loire	24	67	Saône	42
08	Ardennes	21	38	Isère	32	68	Haute Rhin	42
09	Artois	73	39	Jura	43	69	Rhône	82
10	Aube	21	40	Landes	72	70	Haute Saône	43
11	Aude	91	41	Loir et Cher	24	71	Saône et Loire	81
12	Avignon	73	42	Loire	82	72	Sarthe	52
13	Bouches-du-Rhône	92	43	Haute Loire	62	73	Savoie	82
14	Celvedos	51	44	Loire Atlantique	52	74	Scute Savoie	82
15	Cantal	62	45	Loiret	24	75	Saône	11
16	Charente	71	46	Lot	73	76	Seine Maritime	23
17	Charente Maritime	71	47	Lot et Garonne	72	77	Saône et Marne	11
18	Cher	24	48	Lozère	91	78	Saône et Oise	11
19	Corrèze	51	49	Maine et Loire	52	79	Deux Sèvres	22
20	Corse	53	50	Mayenne	51	80	Somme	73
21	Côte d'Or	81	51	Mayenne	21	81	Tarn	73
22	Côte du Nord	53	52	Haute Marne	21	82	Tarn et Garonne	73
23	Creuse	61	53	Mayenne	52	83	Var	92
24	Dordogne	72	54	Mayenne et Mayenne	41	84	Vaucluse	92
25	Doubs	43	55	Mayenne	41	85	Vendée	52
26	Doubs	82	56	Mayenne	53	86	Vienne	71
27	Eure	23	57	Mayenne	41	87	Haute Vienne	61
28	Eure et Loir	24	58	Mayenne	61	88	Vosges	41
29	Finistère	91	59	Mayenne	31	89	Yonne	81
30	Gard	53	60	Mayenne	22	90	Terr. de Belfort	43

• **OBSERVATION** : Le super code ROR (Région de Programme - Département de résidence) permet de classer les départements dans l'ordre des régions de programme. Ainsi le département de la Seine-et-Marne (ROR = 1177) apparaîtra dans les dénombrements après la Seine (ROR = 1175). Les militaires de Montpellier correspondent à ROR = 9999.



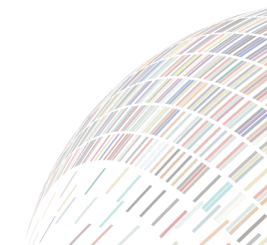


# Préparation de la documentation

*Les autres documents qu'il peut être utile de partager*

- Présentation de l'enquête
- Note méthodologique :
  - pour les variables calculées
  - pour la pondération ou l'échantillonnage
- Résultats publiés
- Etc...

*Qu'est-ce qui peut aider l'utilisateur à bien comprendre les données ?*



# Préparation de la documentation

*Les métadonnées "essentielles" ... \**

## Contextualisation de l'enquête

- **Objet de l'enquête** : résumé du sujet, des objectifs...
- **Univers** : Qui est concerné, enquêté ?
- **Couverture géographique** : région ? métropole ? DOM ? autres... ?
- ...

## Le terrain

- **Echantillonnage** ?
- **Date** de collecte ?
- **Mode** de collecte : face à face, téléphone, auto-administré ... ?
- ...

## Les traitements / Les analyses

- **Pondérations** ? Redressements ?
- **Unité d'analyse** / statistique des données : individu, ménage, ... ?
- **Unité géographique** des données : la commune, le département..., la France ?
- ...

*\* Exemples de champs DDI à renseigner*



# Historique de la norme DDI : **Data Documentation Initiative**

1996



Création d'un **standard** de documentation technique pour **décrire et conserver les informations et données d'enquêtes en sciences humaines et sociales**

**Objectifs :** - *structuration des informations*  
- *harmonisation des pratiques de doc.*

Années 2000



- Mise en place du **Réseau Quetelet**
- **Choix de DDI** pour la documentation des données (*traduction de la description de DDI en français...*)
- Développement de **Nesstar** : outil de diffusion de données et métadonnées sur le web (export de fiches XML), selon le standard DDI

Années 2010



- **Evolution de DDI** :  
de DDI-C (codebook) à DDI-L (lifecycle = DDI 3)
- Logiciel Nesstar : pas d'évolution vers DDI3
- ADISP / CDSP / INED : utilisation du **logiciel Nesstar** (sous DDI 2.1)

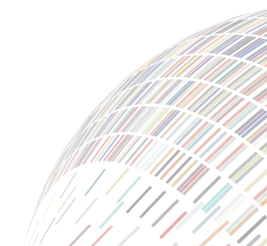


# Structure générale de DDI (version 2.1)

1.0	<b>Document</b> description	<ul style="list-style-type: none"> <li>✓ 72 champs et sous-champs</li> <li>✓ Description de la phase (interne) de documentation</li> </ul>
2.0	<b>Study</b> description	<ul style="list-style-type: none"> <li>✓ 101 champs et sous-champs</li> <li>✓ Éléments généraux sur l'enquête</li> </ul>
3.0	<b>File(s)</b> description	<ul style="list-style-type: none"> <li>✓ 30 champs et sous-champs</li> <li>✓ Présentation générale du (des) fichier(s)</li> </ul>
4.0	<b>Variables</b> description	<ul style="list-style-type: none"> <li>✓ 61 champs et sous-champs</li> <li>✓ Description de chaque variable du fichier</li> </ul>
5.0	<b>Other materials</b>	<ul style="list-style-type: none"> <li>✓ 7 champs et sous-champs</li> <li>✓ Intégration des documents liés à l'enquête</li> </ul>

Soit **270 champs / sous-champs** *proposés* et précisément définis (*objet, format, ...*)

**ADISP** : sélection d'environ **80 champs**



Merci pour votre attention

Contact : [diffusion.adisp@cnrs.fr](mailto:diffusion.adisp@cnrs.fr)

