



HAL
open science

**La documentation de données - Le protocole DDI.
Formation “ SHS et Sciences Ouverte ” le 16 mars 2019
EHESS**

Alina Danciu, Erik Zolotoukhine

► **To cite this version:**

Alina Danciu, Erik Zolotoukhine. La documentation de données - Le protocole DDI. Formation “ SHS et Sciences Ouverte ” le 16 mars 2019 EHESS. École thématique. Recherche SHS et Science Ouverte. Apprendre à mieux gérer et valoriser ses données, EHESS Paris, France. 2019. hal-03921172

HAL Id: hal-03921172

<https://sciencespo.hal.science/hal-03921172>

Submitted on 3 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

Atelier 3 : Protocole DDI

Apprendre à rédiger son Data Management Plan (DMP)

Mieux gérer et valoriser ses données

16 avril 2019 – EHESS

Alina Danciu, Erik Zolotoukhine

-
- Introduction
 - Principes du protocole DDI
 - La norme DDI
 - DDI en pratique

Introduction

Au sein de Quetelet PROGEDO Diffusion, l'ADISP et le CDSP accompagnent les institutions et chercheurs qui souhaitent mettre à disposition leurs données et enquêtes en vue de leur réutilisation par la communauté scientifique et académique.

L'archivage à long terme des jeux de données relève de la responsabilité des producteurs mais certains diffuseurs, comme le CDSP, proposent des prestations d'archivage à long terme des jeux de données qui leur sont confiés.

Les diffuseurs ne deviennent pas propriétaires des données mises à disposition : ils détiennent et transmettent des droits d'usage de ces données en vue de leur réutilisation à des fins scientifiques.

Les droits de propriété intellectuelle restent acquis aux personnes ou institutions ayant déposé les jeux de données.

À L'ADISP

L'ADISP est en charge de collecter les données issues **de la statistique publique**, c'est-à-dire de l'INSEE, des services statistiques ministériels et autres organismes publics.

AU CDSP

Le CDSP prend en charge les dépôts de **données et enquêtes académiques en sciences sociales**, réalisées selon des méthodes qualitatives ou quantitatives et **accompagnées d'une documentation** suffisamment riche pour en permettre la réutilisation.

Quelles sont les informations qui sont indispensables pour l'utilisation d'un fichier de données ?

Quelles sont les informations supplémentaires que vous aimeriez avoir, au cas où elles sont disponibles ?

Principes du protocole DDI

Définition

La Data Documentation Initiative (DDI) est une norme internationale permettant de décrire les données issues d'enquêtes et d'autres méthodes d'observation en sciences sociales, comportementales, économiques et de la santé. DDI peut documenter et gérer différentes étapes du cycle de vie des données de recherche, telles que la conception, la collecte, le traitement, la diffusion, la découverte et l'archivage.

Source : DDI Alliance

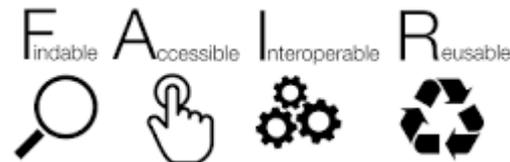


Dans quel but ?

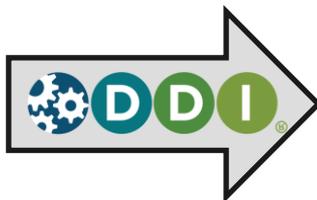
- Documentation ;
- Exploration ;
- Interopérabilité ;
- Réutilisation.

DDI n'est pas un logiciel, mais un standard

Sélectionnez les éléments DDI qui vous correspondent



Un fichier documenté de données



Qui utilise DDI ? Quelques exemples



Centralising and Integrating Metadata from European Statistics



Produits DDI

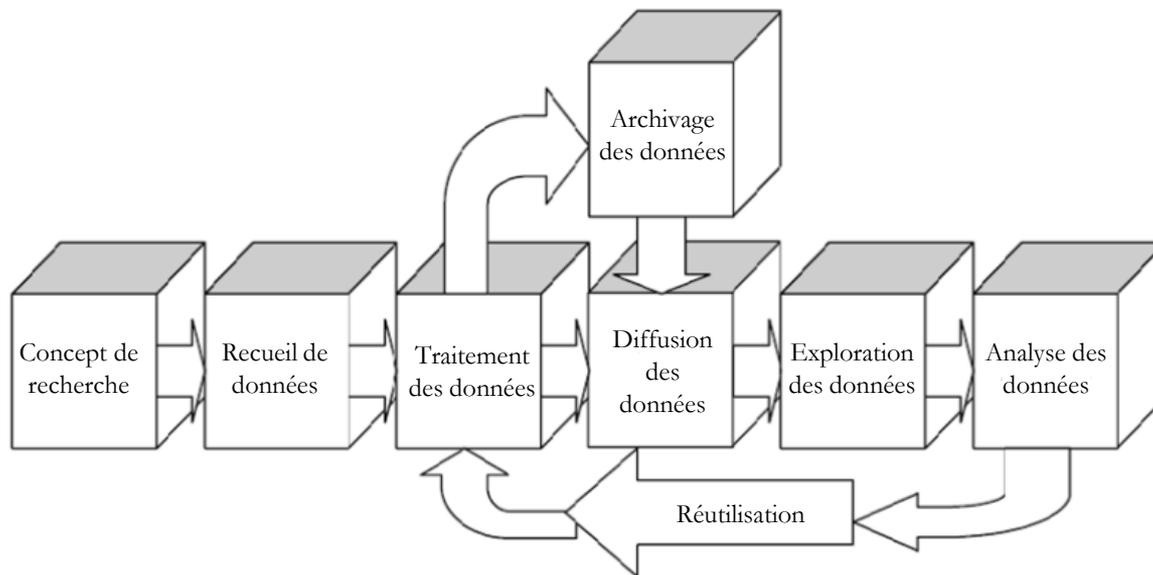
DDI 2

- Version initiale publiée en 2000 (version courante 2.5)
- Le but initial est la **documentation** des données pour permettre leur **exploration** et leur **préservation**

DDI 3

- Version initiale publiée en 2008 (version courante 3.2, version 3.3 à venir)
- Le but principal est d'ajouter du contenu pour permettre :
 - Traitement des fichiers de **données complexes** (ex : données longitudinales...)
 - **Réutilisation** des métadonnées
 - Rendre **compte du cycle de vie** complet des données et des métadonnées dès la conception, au traitement et à la documentation des données, jusqu'à leur diffusion, leur préservation et leur réutilisation

DDI 3 : le cycle de vie des données



Les points communs entre DDI 2 et DDI 3

- L'utilisation de vocabulaires contrôlés ; 
- La structuration des métadonnées ;
- L'échange d'informations ;
- La documentation de variables dans un ensemble de données ;
- La création d'un dictionnaire de codes (*codebook*) pour les utilisateurs finaux ;
- Documenter l'univers des variables (à qui les questions ont été posées ...)
- Créer des groupes de variables ;
- Documenter des jeux de données produits dans différents formats ;
- Distinguer les valeurs manquantes.

Les spécificités de DDI 2 et DDI 3

Les spécifications DDI 2 sont plus adaptées pour :

- La documentation des enquêtes simples et uniques ;
- Où il s'agit de fournir uniquement des informations concernant l'étude.

Les spécifications DDI 3 sont plus adaptées pour :

- La réutilisation, l'harmonisation des questions, des variables ;
- La gestion cohérente des métadonnées, du versionning ;
- La documentation du déroulement du questionnaire ;
- L'échange d'informations avec d'autres normes ;
- La gestion et la documentation des données selon les producteurs et à travers le temps ;
- La documentation des données en différentes langues.

La norme DDI

DDI 2 : deux niveaux de documentation



Source : Pixabay (CC-0)

Copyright © : GESIS Leibniz Institute for the Social Sciences, 2016

Structure générale de DDI (version 2)

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

Soit **270 champs / sous-champs...**

... *proposés*

... et précisément définis
(*objet, format, ...*)

→ **ADISP** : sélection d'environ **80 champs**
renseignés ou non sous Nesstar
dans un "*Template*"

Interopérabilité de DDI avec Dublin Core

1995

Création d'un **format simple** et générique
de **15 éléments**

Objectifs :

- *fournir un socle commun d'éléments descriptif*
- *améliorer la recherche de ressources*

Les 15 éléments décrivent ...

... le contenu

- Title
- Subject
- Description
- Source
- Language
- Relation
- Coverage

... la propriété intellectuelle

- Creator
- Contributor
- Publisher
- Rights

... l'instanciation

- Date
- Type
- Format
- Identifier

DDI a intégré les **15 champs** du format **Dublin Core**

DDI en pratique

La documentation « minimum »



Le dictionnaire des codes / variables

OFFRE

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

Vide	Sans objet (non inscrit auprès de Pôle Emploi ou d'un opérateur 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 redressement consiste, en réinterrogation, à forcer la variable OFFRE à 1 lorsqu'il a reçu d'offre d'emploi depuis la précédente réalisation de l'enquête tout en ayant r

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 remontée du questionnaire d'une précédente interrogation, cette variable n'est 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 v partiellement renseignée aux T1, T2, T3 et T4 2013 pour les individus des ménages d'inactifs de 65 ans et plus en réinterrogation intermédiaire (réparables par la conditi

- 54 -

CODE DE : Département de Résidence

- Remises RI-PL et RI-PM (schéma géographique)
 - Sources de l'Information : col. 2-3 de la carte matresse commune
 - 31 postes possibles : 01 à 99.
- Le poste 99 est un chiffrage particulier pour les militaires d'Afrique du Nord sans résidence personnelle en France dont les RI ont été exploités à Montpellier.
- Correspondance Département Région de Programme

Code du département	Département	N° de code de la RP correspond	Code du département	Département	N° de code de la RP correspond	Code du département	Département	N° de code de la RP correspond
01	Ain	02	31	Haute Garonne	73	61	Orne	51
02	Alsace	22	32	Isère	73	62	Pas de Calais	21
03	Allier	03	33	Loirech	32	63	Puy-de-Dôme	02
04	Alpes Alpes	32	34	Nord	31	54	Savoie Pyrénées	72
05	Hautes Alpes	32	35	Ille et Vilaine	33	65	Hautes Pyrénées	73
06	Alpes Maritimes	32	36	Indre	24	66	Pyrénées Orientales	91
07	Ardeche	32	37	Indre et Loire	24	67	Bas Rhin	42
08	Ardenne	21	38	Isère	32	68	Haute Rhin	42
09	Artois	73	39	Jura	32	69	Rhône	02
10	Aube	21	40	Landes	72	70	Haute Saône	43
11	Aude	31	41	Loir et Cher	24	71	Saône et Loire	81
12	Avignon	73	42	Loire	62	72	Sarthe	52
13	Bouches-du-Rhône	62	43	Haute Loire	62	73	Savoie	02
14	Calvados	51	44	Loire Atlantique	62	74	Savoie Savoie	02
15	Central	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	Seine et Marne	11
18	Cher	24	48	Lozère	31	78	Seine et Oise	11
19	Corbise	61	49	Mayenne	62	79	Seine Oise	11
20	Corse	33	50	Meuse	51	80	Somme	22
21	Côte d'Or	31	51	Meurthe	21	81	Tarn	73
22	Côte du Nord	53	52	Meurthe et Moselle	21	82	Tarn et Garonne	73
23	Creuse	61	53	Mayenne	52	83	Var	62
24	Dordogne	72	54	Meurthe et Moselle	41	84	Vaucluse	62
25	Drôme	43	55	Meuse	41	85	Vendée	52
26	Drôme	43	56	Morbihan	53	86	Vienne	61
27	Eure	23	57	Normandie	41	87	Haute Vienne	71
28	Eure et Loir	24	58	Normandie	41	88	Vosges	41
29	Finistère	63	59	Normandie	31	89	Yonne	61
30	Gard	91	60	Oise	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) appartient dans les départements de la Seine (ROR = 1175). Les départements de Montpellier correspondent à ROR = 9999.

Le questionnaire / formulaire

FAMILLE - MÉNAGE

1. Conjoint en institution

FACONJ

Précédemment, vous m'avez indiqué que votre conjoint ne vivait pas chez vous. Réside-t-il actuellement de manière permanente dans un établissement d'hébergement pour personnes âgées (EHPAD ou maison de retraite) ?

- Oui → FACOPR
- Non → INTRO13

FACOPR

Quel est le prénom de votre conjoint ?

(NSPRF/ non autorisés)

FACOC

Combien payez-vous, pour les frais d'hébergement de [FACOPR] (après déduction éventuelle de l'APA), par semaine, mois ou année ?

Instruction : Ne pas compter les frais remboursés (que ce soit par un organisme ou une personne de l'entourage).

_____ euros

FACOCUT

Unité de temps :

[FACOC] par...

- semaine
- mois
- année

2. Cohabitation

INTRO3

Nous allons maintenant parler des personnes qui vivent avec vous.

Pour l'enquêteur : Il y a un proxy, attention les questions s'adressent au senior interrogé [SENPRE].

FAIE

Quel est le lien entre le [FAPRE] et le senior ? [FAPRE] est...

Instruction : attention les modalités correspondent à une réponse de la forme : « c'est mon/mme... ». Si l'interrogé dit « je suis sa mère / son père », vous devez cocher 3 :

- « [FAPRE] est un enfant du senior »
- Le compagnon / la compagne (conjoint, fiancé, copain, petit-ami) du senior
 - L'enfant (fils/fille) du senior
 - Le père / la mère du senior
 - Le frère ou la sœur du senior
 - Le petit-enfant ou le grand-parent du senior
 - Le gendre, la belle-fille, le beau-père ou la belle-mère du senior
 - Le neveu, la nièce, le cousin, la cousine, l'oncle ou la tante du senior
 - Un autre membre de la famille ou de la belle-famille du senior
 - Un(e) ami(e) du senior
 - Le pensionnaire, sous-location, logeur, enfant en nourrice sans lien de parenté avec le senior
 - Un autre membre de l'entourage du senior (bénévole, voisin...)
 - Un aidant professionnel du senior (infirmière, aide-soignant(e)...)

FAJTS

Avez-vous toujours vécu avec [FAPRE] ?

- Oui → INTRO4
- Non → FAANC

Autres informations « indispensables »

Contextualisation de l'enquête

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

Le terrain

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

Les données / L'analyse

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

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
- ✓ ...

Les champs de DDI : in « document description »

Titre	
Enquête sur l'avenir du Réseau Quetelet - 2020	
N° d'identification	
lil-2020	
Personne(s) responsable(s) de la documentation	
Name	Affiliation
RF	ADISP-CMH
AR	ADISP-CMH
EZ	ADISP-CMH
Adresse documentation site ADISP	
http://www.cmh.ens.fr/greco/enquetes/XML/lil-2020.xml	
Version du document	
Version 1 (14/09/2010) : première publication sur Nesstar webview. Version 2 (11/04/2012) : élimination des variables sensibles. Version 3 (23/01/2015) : harmonisation de la documentation de la partie haute.	
Date de 1ère production du fichier Nesstar	
Year 2010	Month 9
Date de la dernière version	
Year 2015	Month 1

Titre

Indication des auteurs successifs

Lien vers le site web

Versioning des notices
(dates et modifications effectuées)

1^{ère} sous-partie "citation de l'enquête"

CHAMP	EXEMPLE
Titre	Enquête Ménages Déplacements, Nantes / Loire-Atlantique - 2015
<i>Titre alternatif</i>	EMD Nantes - 2015
N° identification	Lil-1025
Producteurs	Cerema / CG de Loire-Atlantique
Diffuseur	Adisp
<i>Nom de série</i>	Enquêtes Ménages Déplacements (EMD)
<i>Informations sur la série</i>	Les EMD sont un outil de connaissance de la mobilité quotidienne d'une population habitant dans un périmètre...
Citation bibliographique	EMD ... - 09, Cerema, ... (producteurs), Adisp (diffuseur)
...	

2^{ème} sous-partie "étendue de l'enquête"

CHAMP	EXEMPLES
Résumé de l'enquête	L'EMD de Nantes porte sur 154 communes ...
Thème(s)	Conditions de vie
Mots-clefs	Transport, véhicule, déplacement, stationnement, trajet, géolocalisation
Univers	Ménages résidant dans une commune du périmètre de l'EMD. Individus de 5 ans et plus. Déplacements effectués la veille de l'interview
Dates de collecte	Début : 2014-09-29 / Fin : 2015-03-15
Couverture géographique	154 communes (autres exemples : France métropolitaine, DOM, ...)
Unité géographique	Zonage en 123 secteurs (autres exemples : commune, département, région...)
Unité d'analyse	Ménages, individus, déplacements
...	

3^{ème} sous-partie "méthodologie et traitement"

CHAMP	EXEMPLE
Méthode d'échantillonnage	Echantillon de ménages tiré de manière aléatoire et stratifié géographiquement
Mode(s) de collecte	Enquête en face à face avec un enquêteur
Instrument(s) de collecte	Questionnaires directifs (3 fiches : ménage, personnes, déplacements)
Pondération	
Opérations de contrôle	
Nettoyage des données	
Taux de réponse	
...	

Les champs de DDI : in « study description » (4/5)

4^{ème} sous-partie "versions de l'enquête"

Version "actuelle" des données

Version de l'enquête

Version 4 : changement de nom des pondérations, ajout d'une variable VERSPOIDS (versionnement des pondérations) ; révision des pondérations pour tenir compte de la dernière version des estimations de population utilisées dans les calculs ; correction de la variable CJRF (pour 137 valeurs).

Date de la version

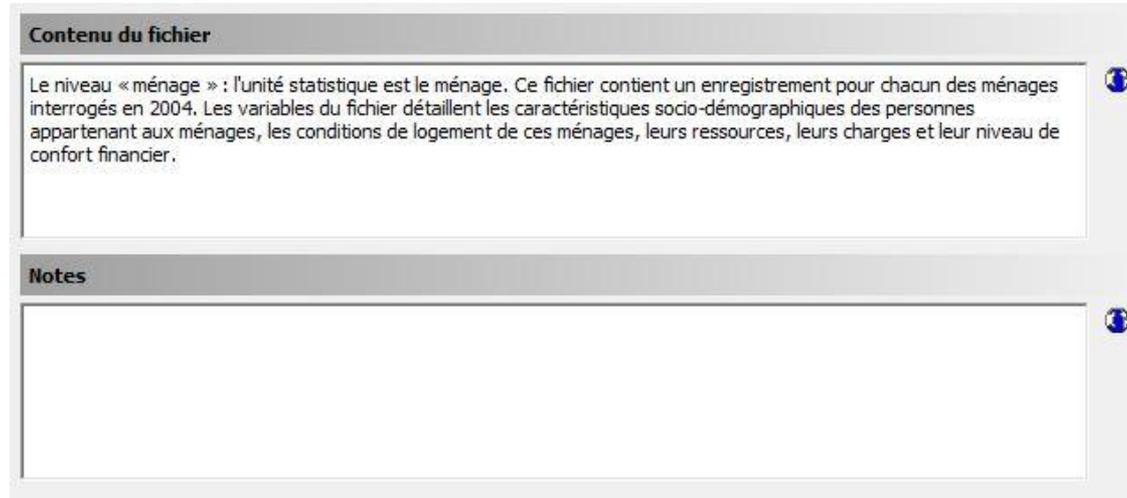
Year	Month	Day
2017	8	30

Notes (versions précédentes)

Version 1 (15/04/2015).
Version 2 (13/11/2015) : Révision annuelle des pondérations (calées sur les estimations démographiques ayant elles-mêmes fait l'objet de révisions).
Version 3 (29/09/2016) : Les variables PUB3FP et STATOEP ont été corrigées.

Versioning des anciennes données
(dates et modifications effectuées)

Les champs de DDI : in « file description »



Contenu du fichier

Le niveau « ménage » : l'unité statistique est le ménage. Ce fichier contient un enregistrement pour chacun des ménages interrogés en 2004. Les variables du fichier détaillent les caractéristiques socio-démographiques des personnes appartenant aux ménages, les conditions de logement de ces ménages, leurs ressources, leurs charges et leur niveau de confort financier.

Notes

- **2 seuls champs conservés pour décrire le fichier**

Les champs de DDI : in « variable description »

The screenshot displays the DDI software interface. On the left, a table lists variables with columns for Number, Name, and Label. The variable 'v18' (P7) is selected, and its label 'Possession du permis de conduire voiture' is highlighted with an orange box labeled 'Nom et label'. Below the table, the 'Documentation' tab is active, showing a tree view of documentation elements for the selected variable. A green circle highlights the 'Documentation' tab, with an arrow pointing to a green box on the right. This box contains a list of documentation elements: 'Texte de la question (d'origine)', 'Univers', 'Consignes aux enquêteurs', 'Introduction à la question', and 'Texte consécutif à la question'. On the right side of the interface, a 'Modalités de la variable' panel shows a category hierarchy for the variable, with three categories: '1 - Oui', '2 - Non', and '3 - Conduite accompagnée et leçons de conduite'. A blue circle highlights this hierarchy.

Number	Name	Label
v12	JOUR	Jour des déplacements
v13	P2	Sexe de la personne
v14	P3	Lien avec la personne de référence
v15	P4	Âge de la personne
v16	P5	Possession d'un téléphone portable
v17	P6	Possession d'une adresse de messagerie électronique
v18	P7	Possession du permis de conduire voiture
v19	P8	Niveau d'études
v20	P9	Composition familiale de la personne

Modalités de la variable

- 1 - Oui
- 2 - Non
- 3 - Conduite accompagnée et leçons de conduite

Documentation

- Texte de la question (d'origine)
- Univers
- Consignes aux enquêteurs
- Introduction à la question
- Texte consécutif à la question

Texte de la question

Possédez-vous le permis de conduire VL (tourisme B) ?

Univers

Les personnes âgées de 16 ans et plus (P4 > 15)

➤ Structuration des variables dans la rubrique "**variable groups**"

Les champs de DDI : in « other materials »

Les documents
associés et proposés

Other Materials

[Questionnaire auprès des représentants de la direction](#)

Liste des variables "Direction"

Questionnaire auprès des représentants du personnel

Liste des variables "Représentants du personnel"

Questionnaire auprès des salariés

Liste des variables "Salariés"

+ Title:

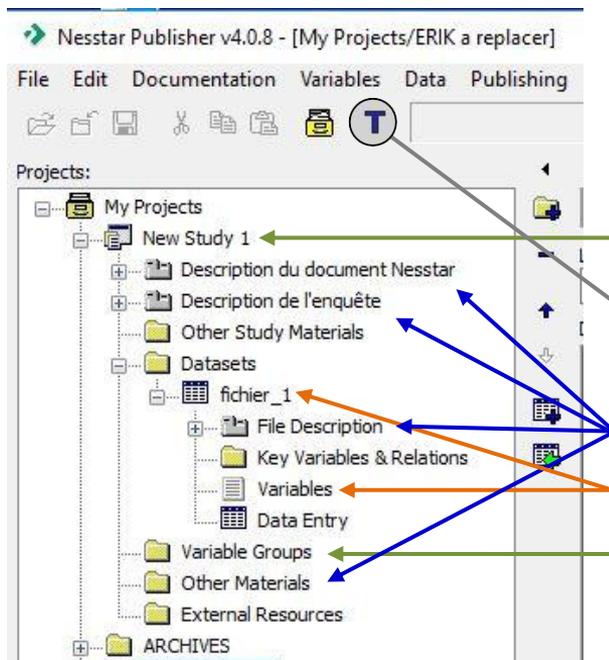
Questionnaire auprès des représentants de la direction

- Location (URL):

http://www.progedo-adisp.fr/documents/il-1211/REPONSE_2017_RD1_Questionnaire_RD2017.pdf

↓ Description:

Lien vers la consultation / téléchargement
du document (sur les sites ADISP et Nesstar)



Déroulé de présentation :

- ❖ Création d'une "nouvelle étude"
- ❖ Le principe du "Template" :
i.e. les 80 champs DDI conservés par l'ADISP
- ❖ Description "générale" de l'enquête
- ❖ Import d'un fichier et documentation des variables
- ❖ Définition des "variable groups" :
structuration des questions
(au choix *par fichiers, thèmes, ...*)
- ❖ Sauvegarde / création d'un fichier *fic.nesstar*
(et export du DDI dans un **fichier XML**)

Utilisations de DDI (et Nesstar) à l'étranger

ICPSR

*Inter-university Consortium
for Political and Social Research*

Les archives de données européennes



L'aide proposée aux producteurs de données

Conseils et supports disponibles ...

... sur le site de l'ADISP

... et sur celui du CDSP



Comment déposer des données ?

Dépôt de données

- Quelques explications
- Une notice, avec les principales informations à conserver / fournir



- Les principes FAIR
- Le guide du déposant (quantitatif)
- Le guide beQuali (qualitatif)

Merci !

alina.danciu@sciencespo.fr

erik.zolotoukhine@cnrs.fr