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

Alina Danciu, Erik Zolotoukhine

SciencesPo
CENTRE DE DONNÉES SOCIO-POLITIQUES

archives de données issues de la statistique publique
disp

PRO GE DO

cnrs

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

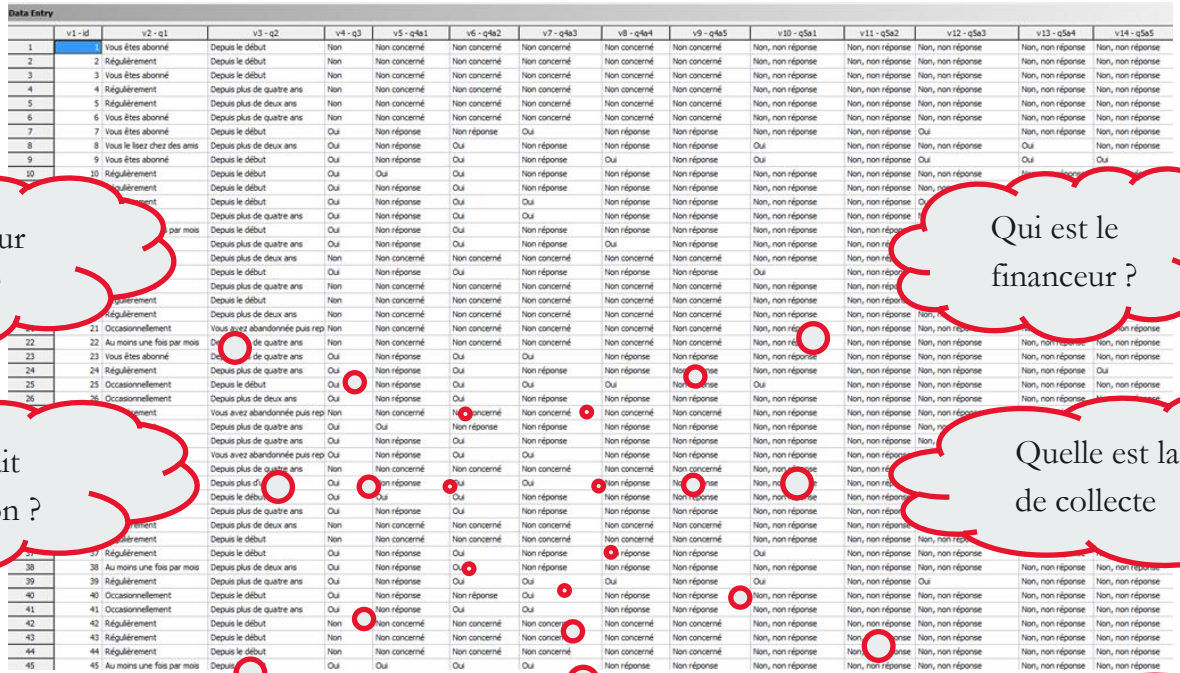
Le CDSP prend en charge les dépôts de **données et enquêtes académiques en sciences sociales**, réalisées selon des méthodes qualitatives ou quantitatives et **accompagnées d'une documentation** suffisamment riche pour en permettre la réutilisation.

Source : ***http://quetelet.progedo.fr/a-qui-sadresser/***

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 ?

Source : ***WebinarTuto@MATE, 14 mars 2019, Alina Danciu, Alexandre Mairot***



Qui est l'auteur de l'enquête ?

Qui est le financeur ?

À qui s'adressait chaque question ?

Quelle est la période de collecte

Que signifient ces codes ?

Comment citer cette enquête ?

Quelle est la question précédente ?

Source : ***WebinarTuto@MATE, 14 mars 2019, Alina Danciu, Alexandre Mairot***

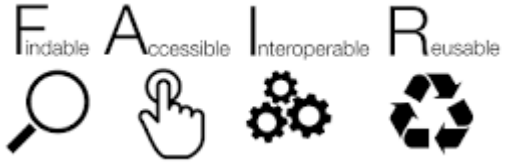
Principes du protocole DDI

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

Source : ***WebinarTuto@MATE, 14 mars 2019, Alina Danciu, Alexandre Mairot***

Un fichier documenté de données

*Source : **WebinarTuto@MATE, 14 mars 2019, Alina Danciu, Alexandre Mairot***

Qui utilise DDI ? Quelques exemples

*Source : **WebinarTuto@MATE, 14 mars 2019, Alina Danciu, Alexandre Mairot***

Produits DDI

DDI 2

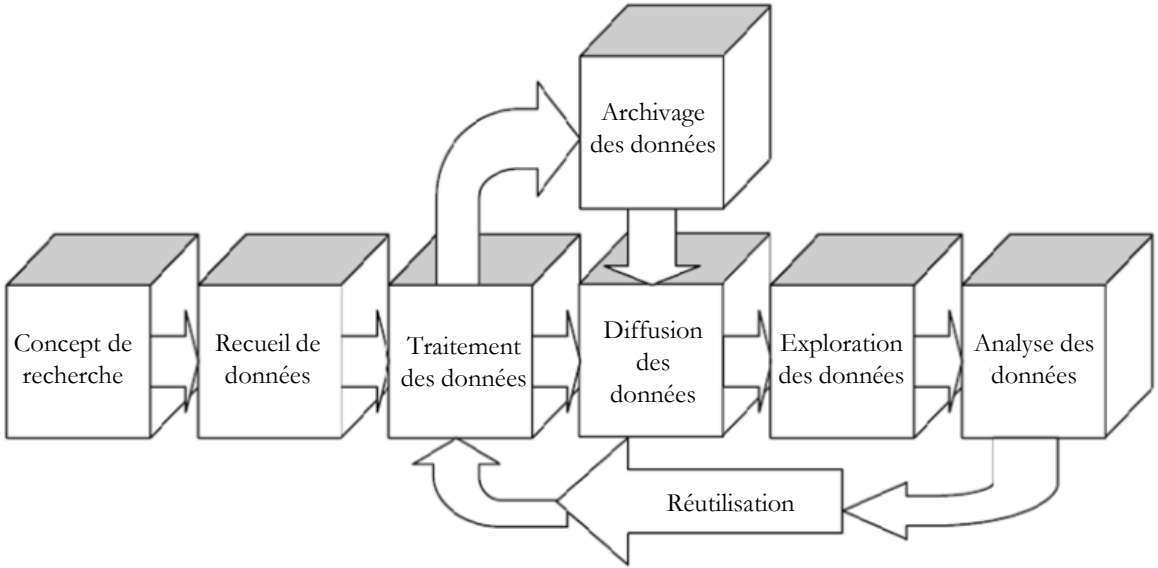
- Version initiale publiée en 2000 (version courante 2.5)
- Le but initial est la **documentation** des données pour permettre leur **exploration** et leur **préservation**

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

*Source : **WebinarTuto@MATE, 14 mars 2019, Alina Danciu, Alexandre Mairot***

DDI 3 : le cycle de vie des données



*Source : **WebinarTuto@MATE, 14 mars 2019, Alina Danciu, Alexandre Mairot***

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.

*Source : **WebinarTuto@MATE, 14 mars 2019, Alina Danciu, Alexandre Mairot***

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.

*Source : **WebinarTuto@MATE, 14 mars 2019, Alina Danciu, Alexandre Mairot***

La norme DDI

DDI 2 : deux niveaux de documentation

Niveau “étude”

Niveau “variable”

Source : Pixabay (CC-0)

Structure générale de DDI (version 2)

1.0	**Document** description	✓72 champs et sous-champs ✓Description de la phase (interne) de documentation
2.0	**Study** description	✓101 champs et sous-champs ✓Eléments généraux sur l'enquête
3.0	**File(s)** description	✓30 champs et sous-champs ✓Présentation générale du (des) fichier(s)
4.0	**Variables** description	✓61 champs et sous-champs ✓Description de chaque variable du fichier
5.0	**Other materials**	✓7 champs et sous-champs ✓Intégration des documents liés à l'enquête

Soit **270 champs / sous-champs**…
… *proposés* → **ADISP** : sélection d'environ **80 champs** renseignés ou non sous Nesstar dans un *"Template"*
… et précisément définis
(objet, format, …)

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	… la **propriété intellectuelle**	… l'instanciation
○ Title	○ Creator	○ Date
○ Subject	○ Contributor	○ Type
○ Description	○ Publisher	○ Format
○ Source	○ Rights	○ Identifier
○ Language		
○ Relation		
○ Coverage		

→ **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 de 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 lorsque reçu d'offre d'emploi depuis la précédente réalisation de l'enquête tout en 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 c remontée du questionnaire d'une précédente interrogation, cette variable n'es 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)

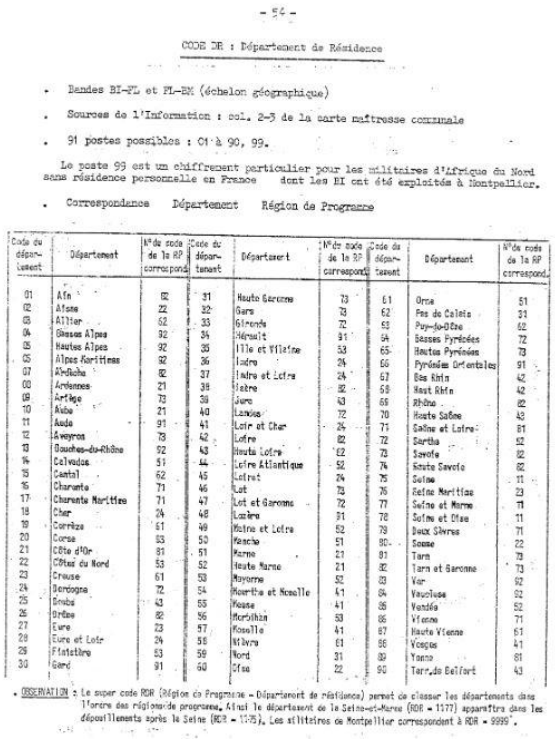
cf	voir partie Nomenclatures des variables
Vide	Sans objet (personnes non actives occupées)
0000	Non renseigné

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 ménag d'inactifs de 65 ans et plus en réinterrogation intermédiaire (repérables par la conditi



- 56 -

CODE DR : Département de Résidence

- Bandes BI-FL et FL-EM (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 EI ont été exploités à Montpellier.

- Correspondance Département Région de Programme

Code du départ-tement	Département	N°de code de la RP correspond.	Code du départ-tement	Département	N°de code de la RP correspond.	Code du départ-tement	Département	N°de code de la RP correspond.
01	Ain	82	31	Haute Garonne	73	61	Orne	51
02	Aisne	22	32	Gers	73	62	Pas de Calais	31
03	Allier	52	33	Gironde	72	63	Puy-de-Dôme	52
04	Basses Alpes	92	34	Hérault	91	64	Basses Pyrénées	72
05	Hautes Alpes	92	35	Ille et Vilaine	53	65	Hautes Pyrénées	73
06	Alpes Maritimes	92	36	Indre	24	66	Pyrénées Orientales	91
07	Ardèche	82	37	Indre et Loire	24	67	Bas Rhin	42
08	Ardennes	21	38	Isère	82	68	Haut Rhin	42
09	Ariège	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	Aveyron	73	42	Loire	82	72	Sarthe	52
13	Bouches-du-Rhône	92	43	Haute Loire	82	73	Savoie	82
14	Calvados	51	44	Loire Atlantique	52	74	Haute Savoie	82
15	Cantal	82	45	Loiret	24	75	Suisse	11
16	Charente	71	46	Lot	73	76	Seine Maritime	23
17	Charente Maritime	71	47	Lot et Garonne	72	77	Seine et Marne	13
18	Cher	24	48	Lozère	91	78	Seine et Oise	11
19	Corrèze	61	49	Maine et Loire	52	79	Deux Sèvres	71
20	Corse	93	50	Manche	51	80	Somme	22
21	Côte d'Or	81	51	Marne	21	81	Tarn	73
22	Côtes 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	Meurthe et Moselle	41	84	Vaucluse	92
25	Doubs	43	55	Meuse	41	85	Vendée	52
26	Drôme	82	56	Morbihan	53	86	Vienne	71
27	Eure	23	57	Moselle	41	87	Haute Vienne	61
28	Eure et Loir	24	58	Nièvre	81	88	Vosges	41
29	Finistère	53	59	Nord	31	89	Yonne	81
30	Gard	91	60	Oise	22	90	Ter.de Belfort	43

OBSERVATION : Le super code RDR (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 (RDR = 1177) apparaîtra dans les dépouillements après la Seine (RDR = 1175). Les militaires de Montpellier correspondent à RDR = 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) ?
1. Oui → FACOPR
2. Non → INTRO13

FACOPR
Quel est le prénom de votre conjoint ?
|_____|
(NSP/RF 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…
1. semaine
2. mois
3. 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].

FALIE
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/ma… ». Si l'interrogé dit « je suis sa mère / son père », vous devez cocher 3 : « [REPRE] est un enfant du senior. »
1. Le compagnon / la compagne (conjoint, fiancé, copain, petit-ami) du senior
2. L'enfant (fils/fille) du senior
3. Le père / la mère du senior
4. Le frère ou la sœur du senior
5. Le petit-enfant ou le grand-parent du senior
6. Le gendre, la belle-fille, le beau-père ou la belle-mère du senior
7. Le neveu, la nièce, le cousin, la cousine, l'oncle ou la tante du senior
8. Un autre membre de la famille ou de la belle-famille du senior
9. Un(e) ami(e) du senior
10. Le pensionnaire, sous-locataire, logeur, enfant en nourrice sans lien de parenté avec le senior
11. Un autre membre de l'entourage du senior (bénévole, voisin…)
12. Un aidant professionnel du senior (infirmièr(e), aide-soignant(e)…)

FATJS
Avez-vous toujours vécu avec [FAPRE] ?
1. Oui → INTRO4
2. 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…)* …**

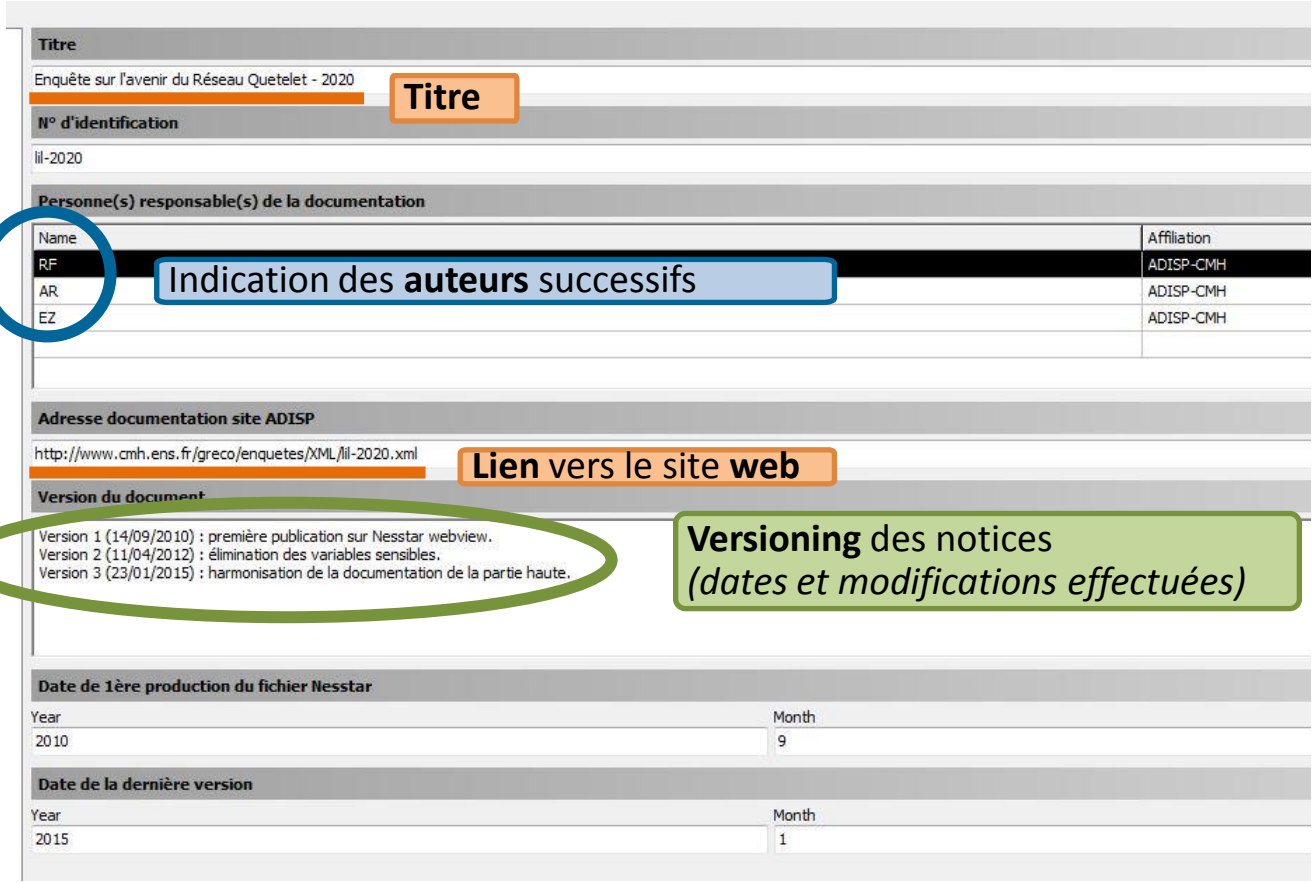
Contextualiser les résultats de recherche :

- ✓ 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 »



Les champs de DDI : in « study description » (1/5)

1ère sous-partie "citation de l'enquête"

CHAMP	*EXEMPLE*
Titre	Enquête Ménages Déplacements, Nantes / Loire-Atlantique - 2015
Titre alternatif	EMD Nantes - 2015
N° identification	Lil-1025
Producteurs	Cerema / CG de Loire-Atlantique
Diffuseur	Adisp
Nom de série	Enquêtes Ménages Déplacements (EMD)
Informations sur la série	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)
…	

Les champs de DDI : in « study description » (2/5)

2ème 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
…	

Les champs de DDI : in « study description » (3/5)

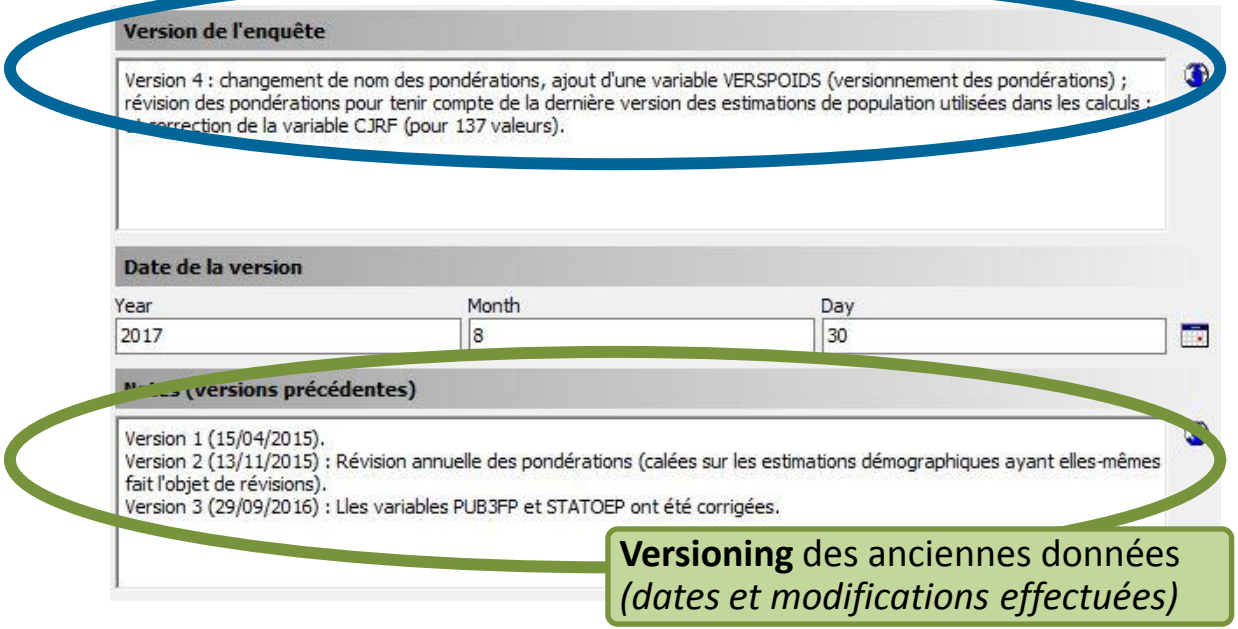
3ème sous-partie "méthodologie et traitement"

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)

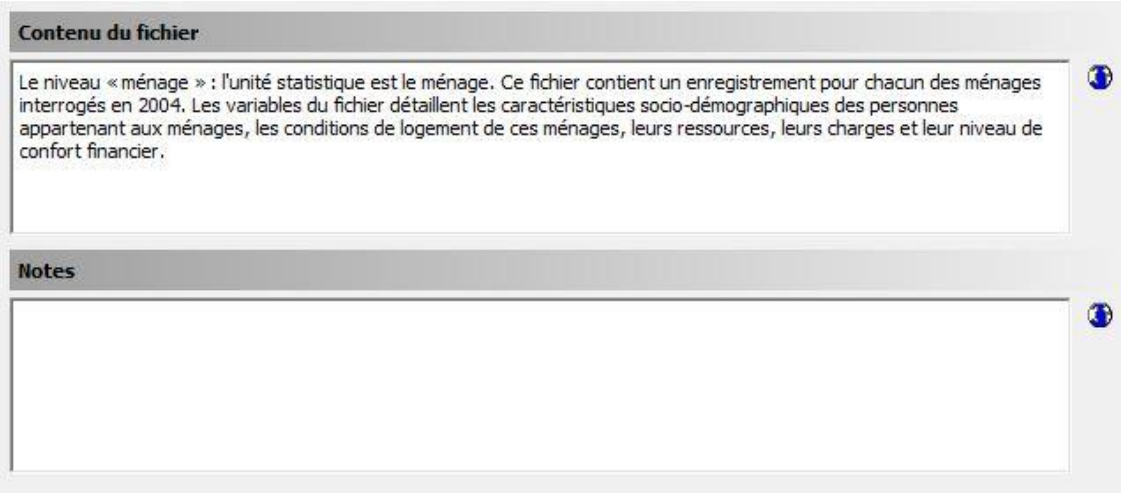
4ème sous-partie "versions de l'enquête"

Version "actuelle" des données



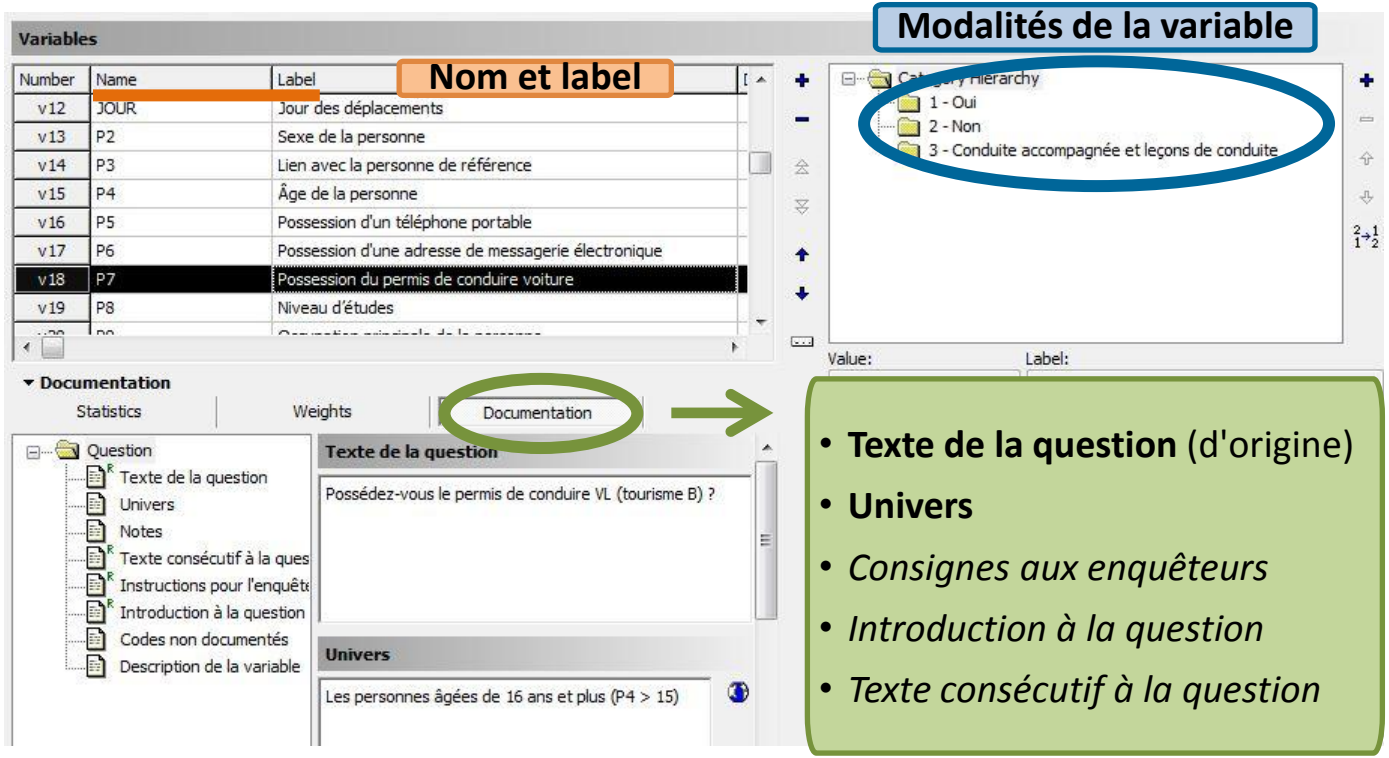
Versioning des anciennes données
(dates et modifications effectuées)

Les champs de DDI : in « file description »



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

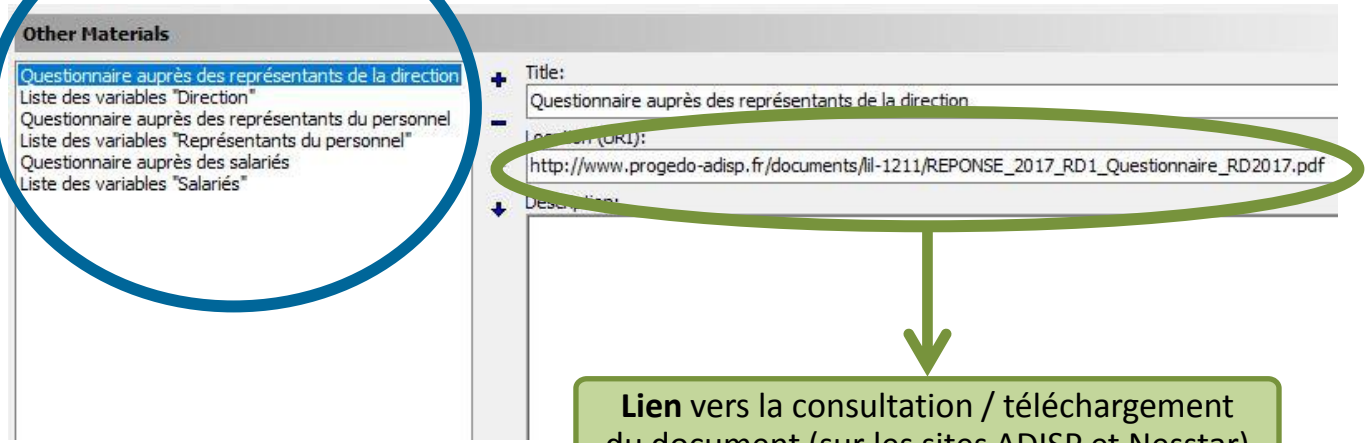
Les champs de DDI : in « variable description »



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

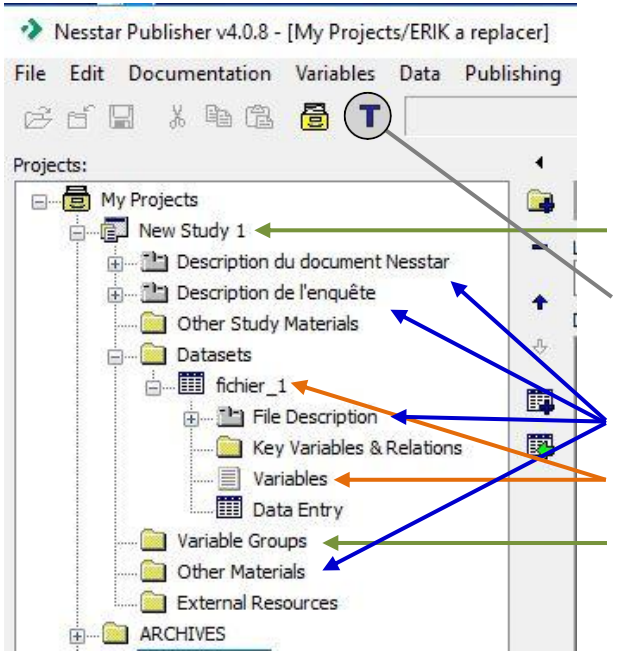
Les champs de DDI : in « other materials »

Les **documents** associés et proposés



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

Présentation de l'outil Nesstar



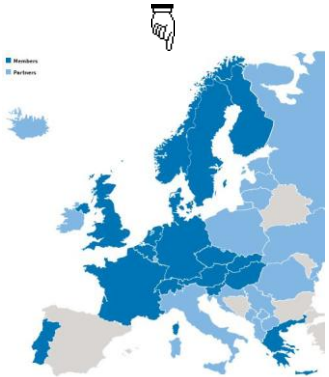
<u>Déroulé de présentation</u> :

- 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**

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

… et sur celui du **CDSP**

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

Merci !

alina.danciu@sciencespo.fr

erik.zolotoukhine@cnrs.fr