

Providing Data in a User-Friendly Manner at the Center for Socio-Political Data Paris

Alina Danciu, Alexandre Mairot

▶ **To cite this version:**

Alina Danciu, Alexandre Mairot. Providing Data in a User-Friendly Manner at the Center for Socio-Political Data Paris. 10th Annual European DDI User Conference (EDDI18), Dec 2018, Berlin, Germany. hal-02874091

HAL Id: hal-02874091

https://sciencespo.hal.science/hal-02874091

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 4.0 International License

SciencesPo
CENTRE DE DONNÉES SOCIO-POLITIQUES

Providing Data in a User-Friendly Manner at the Center for Socio-Political Data Paris

EDDI18 – 10th Annual European DDI User Conference,
December, 4 2018, Berlin

Alina Danciu, Alexandre Mairot

Creative Commons Attribution 4.0 International (CC BY 4.0)

“Documents are malleable, mutable, and mobile”*

“Data are even more malleable, mutable, and mobile than documents.”**

* LATOUR Bruno, Sciences in Action : How to Follow Scientists and Engineers through Society, Cambridge, Havard University Press

** WALLIS Jillian, ROLANDO Elizabeth and BORGMAN Christine (2013), “If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology”, PLoS One, vol. 8, n° 7 (DOI :10.1371/journal.pone.0067332)

Making data discoverable and reusable*

- Data tend to exist in **small units**, are **linked** to many other related units,
- They are **context-dependent**: difficult to interpret without considerable documentation and context,
- **"Big science"** (large teams, long-term projects, extensive instrumentation): great in volume and consistent in structure,
- **"Small science"** (individual or small teams, data collected for specific projects): small in volume, local in character, intended for use only by these teams, and are less likely to be structured in ways that allow data to be transferred easily between teams or individuals,
- **Making data from the long tail discoverable and reusable is emerging as a major challenge.**

* WALLIS Jillian, ROLANDO Elizabeth and BORGMAN Christine (2013), "If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology", PLoS One, vol. 8, n° 7 (DOI :10.1371/journal.pone.0067332)

Data archives: putting the pieces together





Meaningful metadata

- Defining rules and procedures internally
- DDI Controlled Vocabularies Group (DDI-CVG)
- CESSDA recommandations (ex: testing CESSDA Metadata Portfolio)

One variable, One documentation, Multiple datasets

Dataset: Enquête annuelle - vague 5 (2017)

EA, ELIPSS Mars 2017

Variable ea17_E1A: Personnes dans le ménage percevant : salaires, traitements et primes

PREQUESTION TEXT

Les questions qui suivent portent sur vos revenus et les prestations sociales que vous touchez éventuellement. Elles sont détaillées afin d'établir une estimation de votre niveau de vie et celui de votre ménage.

LITERAL QUESTION

Y a-t-il actuellement dans votre ménage (y compris vous-même), une ou plusieurs personnes qui perçoi(ven)t les ressources suivantes·: - salaires, traitements et primes

Values	Categories	N	
1	Oui	1760	64.6%
2	Non	891	32.7%
3	Je ne sais pas	74	2.7%

SUMMARY STATISTICS

Valid cases	2725
Missing cases	0
This variable is numeric	

NOTES

Batterie de questions dans le questionnaire E1 (Y-a-t-il actuellement dans votre ménage (y compris vous), une ou plusieurs personnes qui perçoi(ven)t les ressources suivantes ?), comprenant plusieurs sous-questions divisées dans la base en différentes variables (E1A à E1J)

Dataset: Dynamiques de mobilisation - vague 18 (2017)

DYNAMOB - vague 18, ELIPSS Décembre 2017

Variable ea17_E1A: Personnes dans le ménage percevant : salaires, traitements et primes

PREQUESTION TEXT

Les questions qui suivent portent sur vos revenus et les prestations sociales que vous touchez éventuellement. Elles sont détaillées afin d'établir une estimation de votre niveau de vie et celui de votre ménage.

LITERAL QUESTION

Y a-t-il actuellement dans votre ménage (y compris vous-même), une ou plusieurs personnes qui perçoi(ven)t les ressources suivantes·: - salaires, traitements et primes

Values	Categories	N	
1	Oui	1447	64.1%
2	Non	755	33.4%
3	Je ne sais pas	57	2.5%
96	Non enquêté	67	

SUMMARY STATISTICS

Valid cases	2259
Missing cases	67
This variable is numeric	

NOTES

Batterie de questions dans le questionnaire E1 (Y-a-t-il actuellement dans votre ménage (y compris vous), une ou plusieurs personnes qui perçoi(ven)t les ressources suivantes ?), comprenant plusieurs sous-questions divisées dans la base en différentes variables (E1A à E1J)

Metadata variables standardisation

UNIVERSE:

Model : “If interviewer” + [LITTERAL CONDITION] + “(meaning (“ + [ALGORITHMIC CONDITION] +”))”

Si l'interviewé est inscrit sur les listes électorales (c'est-à-dire (INSCRIPTION in 1:2))
Si l'interviewé apporte une aide à domicile à un tiers (c'est-à-dire (Q1A==1))

NOTES:

Model : “Grid question in the questionnaire” + [GRID QUESTION NAME] + “, meaning several sub-questions divided in the base in different variables (” + [VARIABLE1] + “to” + [VARIABLE2] + “).”

Batterie de questions dans le questionnaire RAD_STAT, comprenant plusieurs sous-question divisées dans la base en différentes variables (RAD_STAT1 à RAD_STAT23).
Batterie de questions dans le questionnaire MUS_GENR, comprenant plusieurs sous-questions divisées dans la base en différentes variables (MUS_GENR1 à MUS_GENR13).

dataKind: Kind of Data

DDI Alliance example*	fr.cdsp.ddi.AGORA1978	fr.cdsp.ddi.elipss.2013.04.ea	fr.cdsp.ddi.PostElect1997
survey data	Individual survey data	Data from individual surveys and geographical data taken from the census.	survey data

* http://www.ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/field_level_documentation.html

collSitu: Characteristics of Data Collection Situation

DDI Alliance example*	fr.cdsp.ddi.elipss.2012.12.pn	fr.cdsp.ddi.FES2007	fr.cdsp.ddi.LMSEP1986
There were 1,194 respondents who answered questions in face-to-face interviews lasting approximately 75 minutes each.	Estimated time to complete questionnaire: 15 minutes. Panellists who had begun the questionnaire before the end of the main fieldwork were able to access and complete it up to 31 May 2014.	The total number of households in the sample is 10,469, with 5849 valid households, making a total of 2000 interviews.	Since the questionnaires were self-administered and in paper format, it is difficult to assess how fully they were completed. It is therefore not possible to distinguish between real non-responses, individuals for whom a question was "not applicable" (i.e. filtered by a previous question) or anomalous values. In this survey, all these situations are treated as non-responses (coded 0).

* http://www.ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/field_level_documentation.html

resInstru: Type of Research Instrument

DDI Alliance example*	fr.cdsp.ddi.AGORA1981	fr.cdsp.ddi.elipss.2014.02.soligene	fr.cdsp.ddi.elipss.2014.12.ess
structured	A large part of the questionnaire consists of closed questions.	Questionnaire consisting of closed questions. Methodological and technical specificities: Presentation of some questions in the form of scenarios. Random rotation of items in certain batteries of questions. Randomisation in the order of two blocks in the questionnaire (family mutual aid and role-playing).	The questionnaire consists of closed questions and a few open questions.

* http://www.ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/field_level_documentation.html

timeMeth: Time Method

DDI Alliance example*	fr.cdsp.ddi.AGORA1986	fr.cdsp.ddi.BPF2007-R4	fr.cdsp.ddi.elipss.2013.07.fecond
panel survey cross-section trend study time-series	The Agoramétrie surveys were conducted annually from 1977 to 2005.	4 survey waves 4th wave: February 5-19, 2007	The questionnaire consists of closed questions and a few open questions.

* http://www.ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/field_level_documentation.html

weight: Weighting

DDI Alliance example*	fr.cdsp.ddi.AGORA1982	fr.cdsp.ddi.elipss.2014.05.dynamob	fr.cdsp.ddi.PostElec2012
The 1996 NES dataset includes two final person-level analysis weights which incorporate sampling, nonresponse, and post-stratification factors. One weight (variable #4) is for longitudinal micro-level analysis using the 1996 NES Panel. The other weight (variable #3) is for analysis of the 1996 NES combined sample (Panel component cases plus Cross-section supplement cases). In addition, a Time Series Weight (variable #5) which corrects for Panel attrition was constructed. This weight should be used in analyses which compare the 1996 NES to earlier unweighted National Election Study data collections	The data from surveys conducted by Agoramétrie are not weighted.	The sampling procedure for the pilot survey had underestimated the impact that the small sample size and the process for selecting the primary units could have on the weight spread. To limit this spread, it was proposed that a uniform sampling weight should be set for the whole sample. The individual weightings were adjusted for nonresponse by the homogeneous response groups method, then by an adjustment on five criteria in the 2014 Annual Census Survey (sex, age, nationality, qualification and ZEAT (study and development zone)). The weightings of the survey respondents alone are adjusted once again on the same criteria to correct for nonresponse in the survey wave. Together, they add up to the size of the respondent sample. The weighting documentation gives more details on the adjustment procedure and the use of weightings.	The weighting was calculated on the criteria of sex, age, socio-professional category of the reference person, region, conurbation size, official results of the 1st and 2nd rounds of the 2012 presidential election, and educational qualifications. Four weighting variables are available: Weight0: Adjustment for sociodemographic criteria: Sex, age, occupation of household head, Region, category of municipality. Weight1: Weight0 adjustment + vote in 1st round of 2012 presidential election. Weight2: Weight1 adjustment + vote in 2nd round of the 2012 presidential election. Weight3: Weight2 adjustment + educational qualification. They are described in the following document:

* http://www.ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/field_level_documentation.html

DDI3 principles in a DDI2 environment*

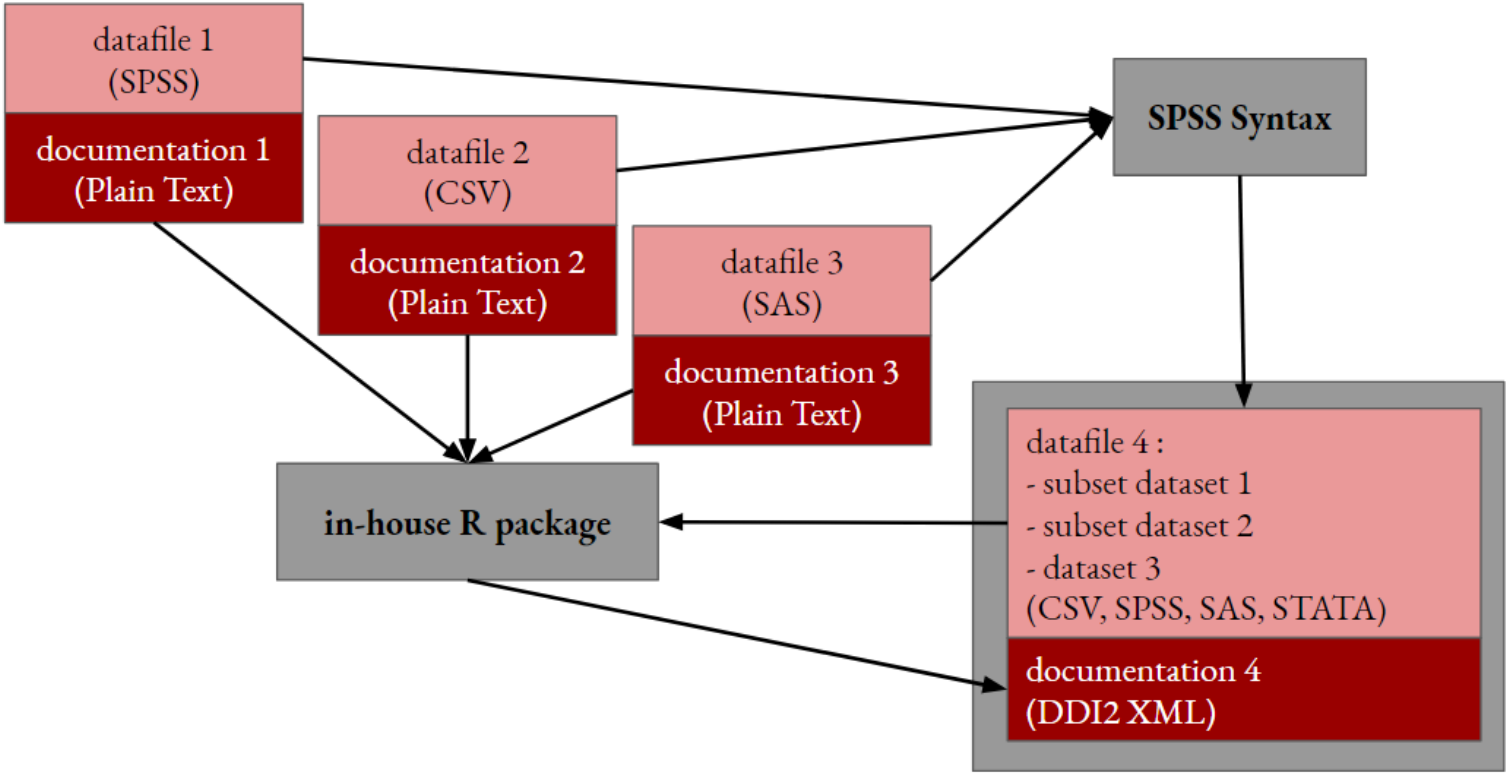
DDI3 Principles	How we manage in DDI2
All Identifiable Objects have an object source attribute that can contain a DDI URN.	Link between documentation source and documentation target
Description of the source of the data.	Documentation all the sources of data beyond the using for the documentation with the DDI2 structure.
Identification of the versions	Versionning documentation file and data file
Lifecycle pulls out the various bits so that you can manage and control them better	Standardisation universe, notes across the dataset (next step at the level collection)

* JOHNSON Jon, SMITH Dan, THOMAS Wendy, WACKEROW Joachim (2018), *Workshop :* "Data Documentation Initiative (DDI) – Train the Trainers", Dagstuhl, 23th-28th September (https://www.dagstuhl.de/18393)

DDI3 principles in a DDI2 environment*

DDI3 Principles	How we manage in DDI2
Allows reuse of question	Standardisation variable labels across the collection of dataset
Reuse and metadata	Routine to reuse documentation for producing XML Files
Separates question content from the use of the question	Quetelet Questions bank
Captures assembly into a questionnaire (question flow, instructions, informational text)	Documentation of the "Interviews instructions", universe, prequestion and describe in the notes tags all the context of the question
Tracks the flow of data and deposits it into a variable	Documentation of the formula in separated field

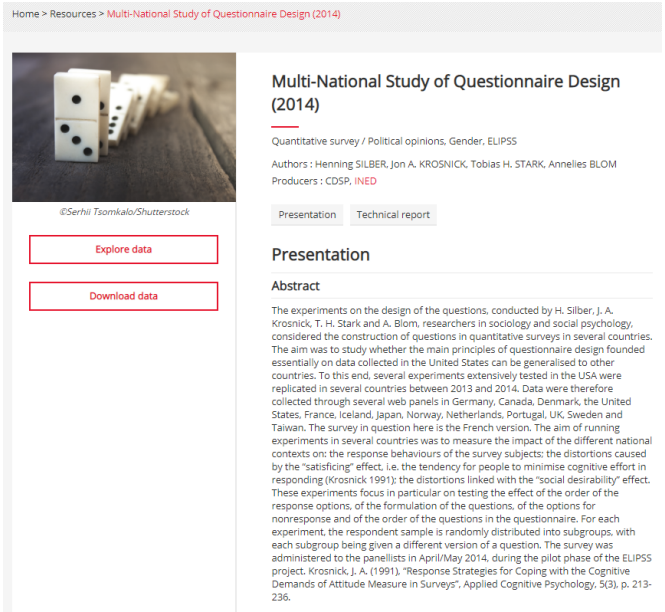
* JOHNSON Jon, SMITH Dan, THOMAS Wendy, WACKEROW Joachim (2018), *Workshop :* "Data Documentation Initiative (DDI) – Train the Trainers", Dagstuhl, 23th-28th September (https://www.dagstuhl.de/18393)



Link to in-house R package

Translating metadata to improve discoverability

- ⅓ of our data users prefer English as a working language
- Bilingual study metadata-level (work in progress)
- Users can identify easier relevant surveys for their research



How about accessibility?



alina.danciu@sciencespo.fr
alexandre.mairot@sciencespo.fr

