Documenting Panel Data

Marcel Hebing and Knut Wenzig | EDDI 2015

Updated Version, doi:10.5281/zenodo.55613,

Abstract

The key characteristics of panel studies include repeated measures for a more or less stable sample over time. The core challenge in documenting panel studies is the documentation of these repeated measures (usually questions) and the resulting variables because various reasons can require modifications of measures over time—resulting in comparable but not identical data structures.

The DDI standard provides not one but multiple options for the documentation of panel data. In this workshop we like to present various options and discuss their feasibility for common use cases. The German Socio-Economic Panel (SOEP) will provide the primary use case, but participants are also invited to introduce and discuss their own use cases.

The workshop starts with a short introduction of both panel studies and the DDI standard. Therefore, no previous knowledge of the DDI standard is required to participate in the workshop. The goal for the workshop is to gain a deeper understanding of possible documentation strategies for panel studies.

Agenda

1. Introduction
2. Metadata basics
3. Linking data
4. Questionnaire documentation

   --- break ---
5. Working session
6. Implementation and design choices
7. Use case: SOEP
8. Wrap up

Introduction

Introduction

- What is today's topic?

- Course instructors
- Participants

- What are the specific challenges of documenting a panel study?
- Participants: what are your challenges?

I am...

- Name
- Institution
- Background
- Do you actually work with metadata?
- What do you expect from the workshop / would like to learn?

The **German Socio-Economic Panel (SOEP)** is a wide-ranging representative longitudinal study of private households, located at the German Institute for Economic Research, DIW Berlin. Every year, there were nearly 11,000 households, and more than 20,000 persons sampled by the fieldwork organization TNS Infratest Sozialforschung.

The data provide information on all household members, consisting of Germans living in the Old and New German States, Foreigners, and recent Immigrants to Germany. The Panel was started in 1984.

Some of the many topics include household composition, occupational biographies, employment, earnings, health and satisfaction indicators.

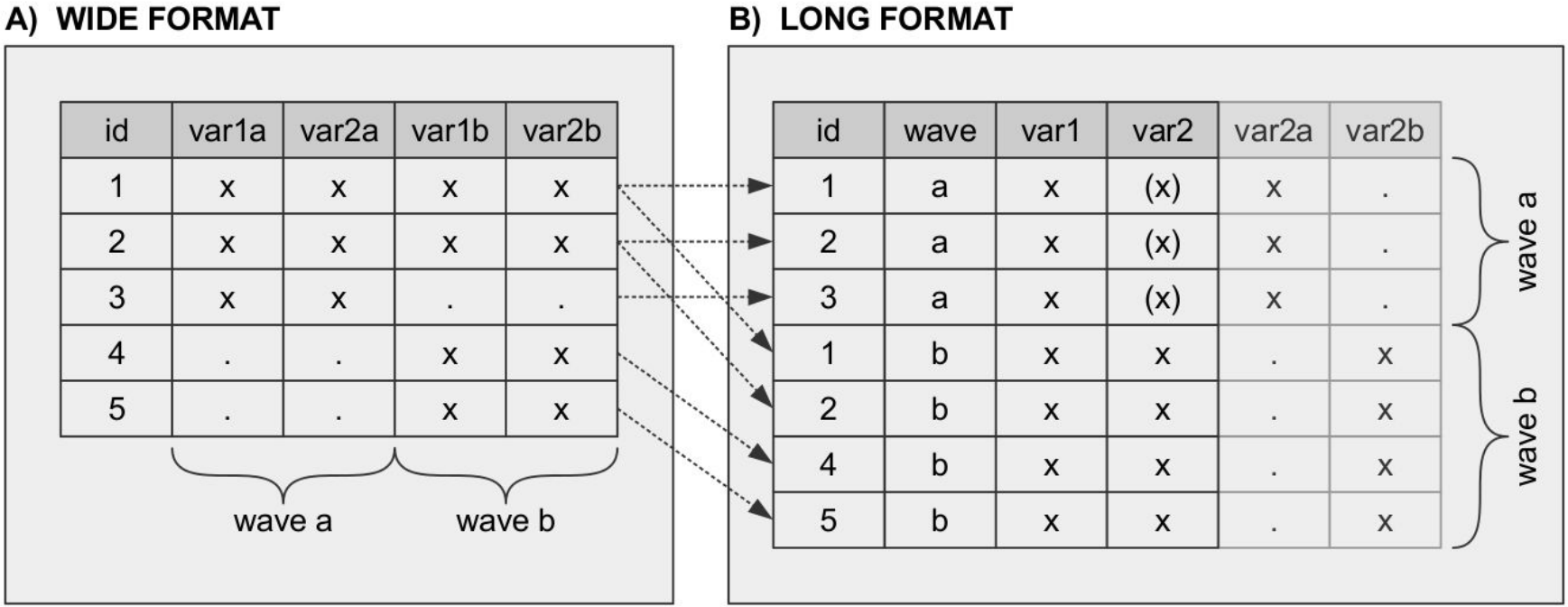
http://www.diw.de/soep

Challenges, specific to panel studies

- Finding repeated measures
- Understanding repeated measures
- Finding the corresponding variables

Advanced:

- Measures change over time
- Finding generated / transformed variables



Metadata basics

XML

```
<table>
    <width>100</width>
    <height>70</height>
</table>
```

XML

```
<table id="2" class="kitchen-table">
    <width unit="cm">100</width>
    <height unit="cm">70</height>
</table>
```

XML

```
<table>
    <tr>
        <td>Element 1</td>
        <td>Element 2</td>
    </tr>
</table>
```

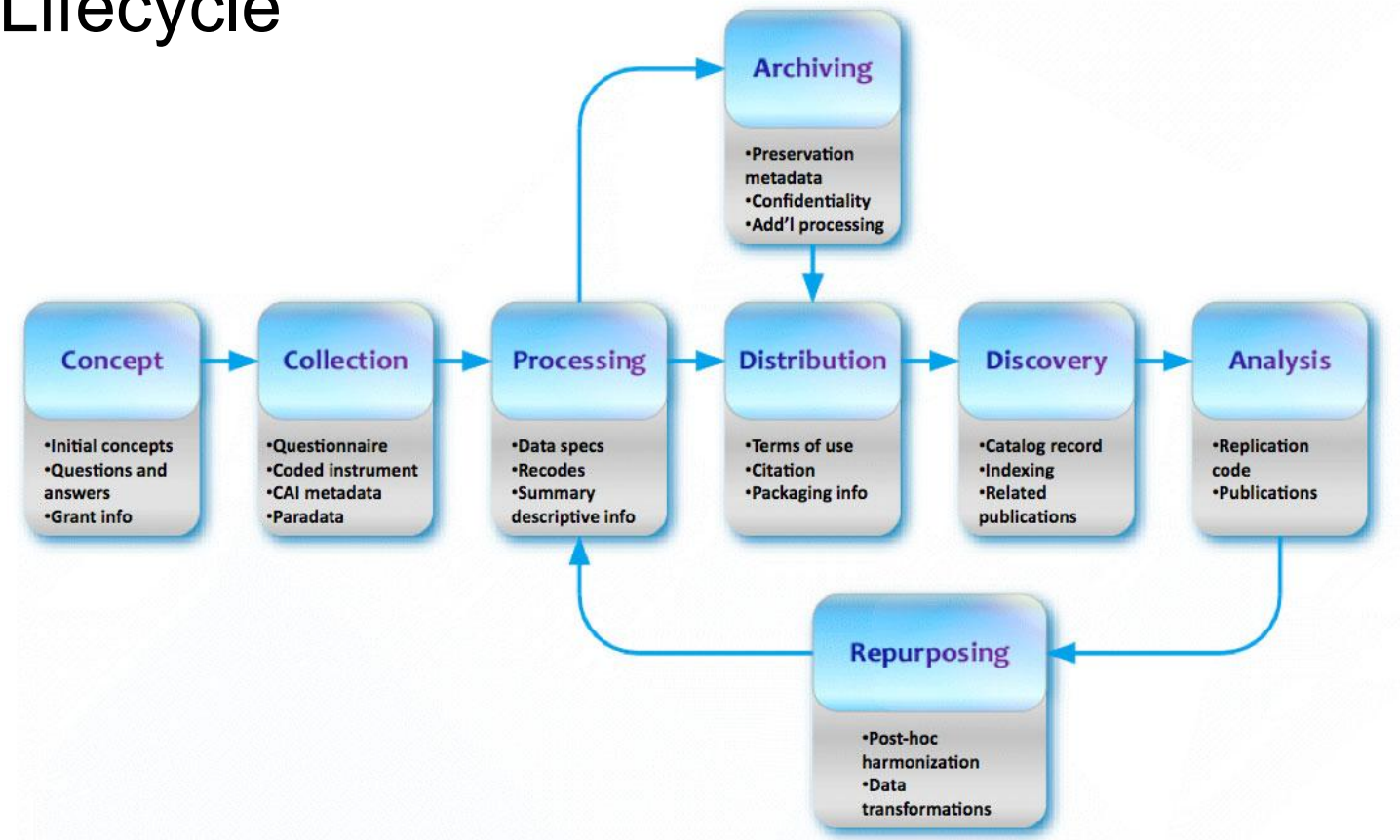
XML

```
<table id="2" class="kitchen-table">
    <width unit="cm">100</width>
    <height unit="cm">70</height>
</table>

<table>
    <tr>
        <td>Element 1</td>
        <td>Element 2</td>
    </tr>
</table>
```

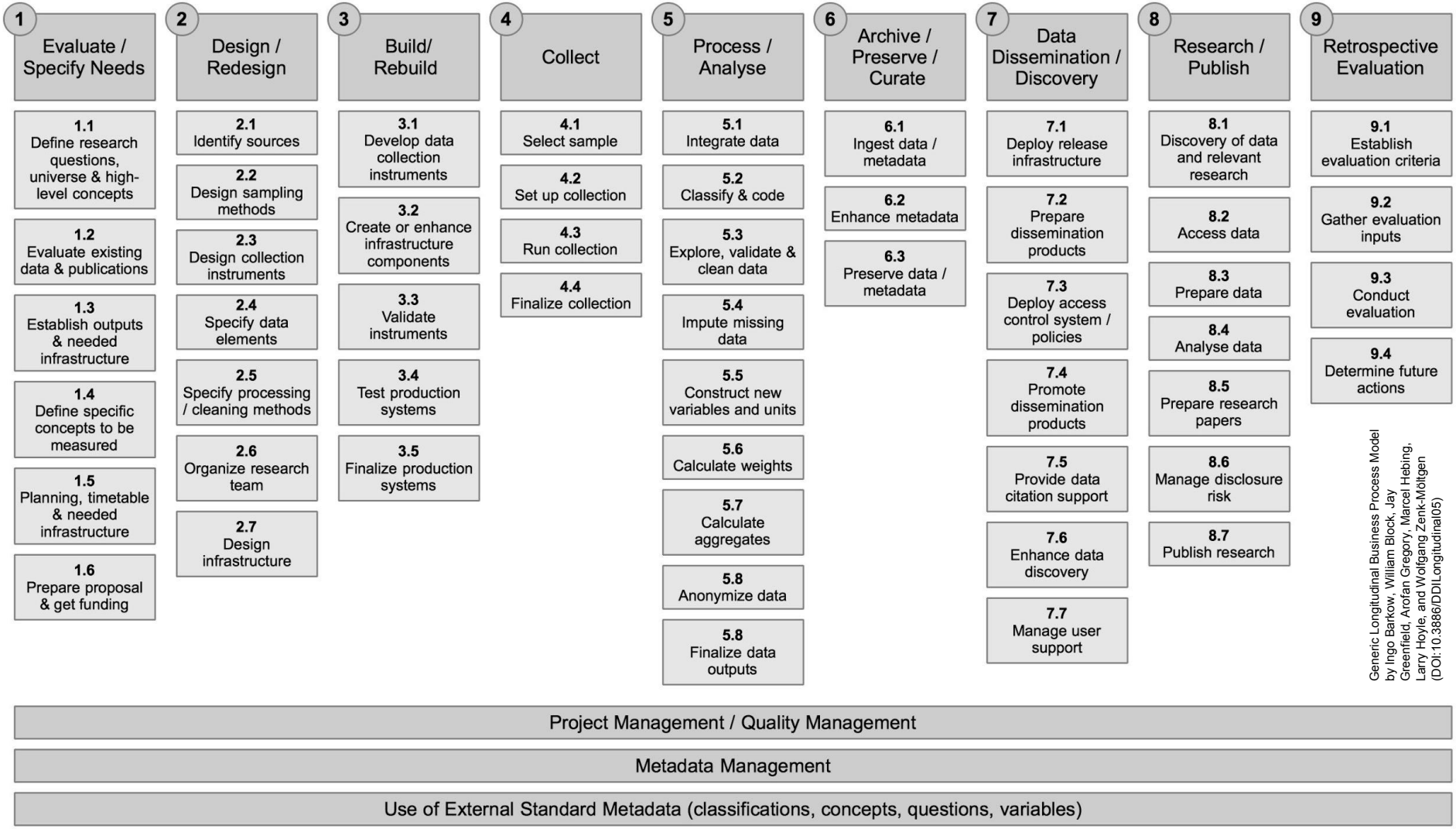
XML

```
<a:table id="2" class="kitchen-table" xmlns:a="https://...">
    <a:width unit="cm">100</a:width>
    <a:height unit="cm">70</a:height>
</a:table>

<b:table xmlns:b="https://...">
    <b:tr>
        <b:td>Element 1</b:td>
        <b:td>Element 2</b:td>
    </b:tr>
</b:table>
```

DDI Lifecycle



Source: http://www.ddialliance.org/sites/default/files/what-is-ddi-diagram.jpg



Generic Longitudinal Business Process Model by Ingo Barkow, William Block, Jay Greenfield, Arofan Gregory, Marcel Hebing, Larry Hoyle, and Wolfgang Zenk-Möltgen (DOI:10.3886/DDILongitudinal05)

Versions of DDI

DDI Codebook

- Version 1.x (Nesstar)
- Version 2.x

DDI Lifecycle

- Version 3.x
- Version 4.x = Moving Forward (under development)

DDI Codebook tree (incl. DDI Lite)

http://www.ddialliance.org/sites/default/files/ddi-lite.html

Linking data

What do we want to achieve?

Within one panel study:

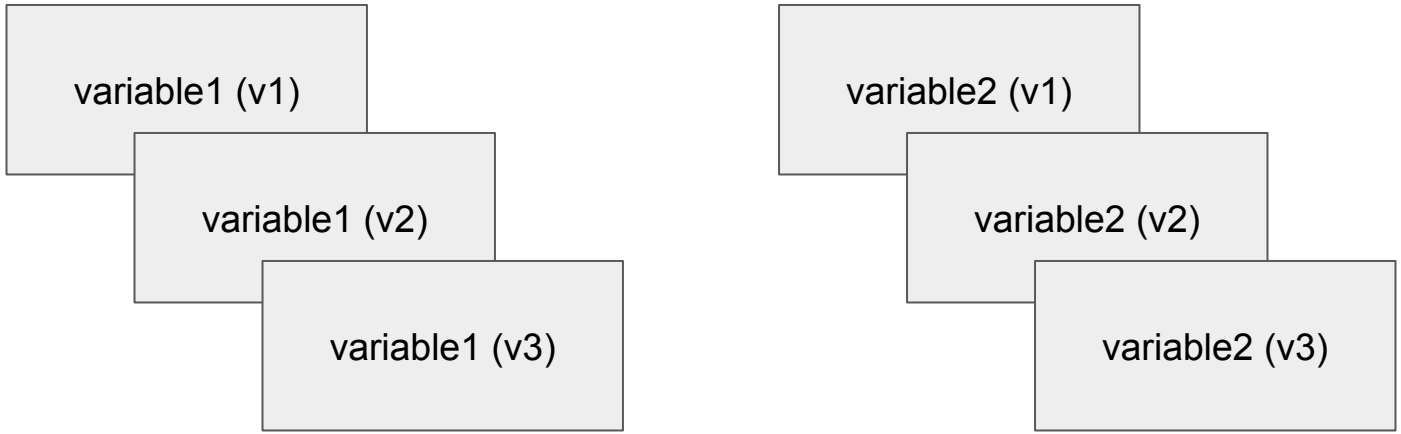
- Find related variables over time
- Document changes over time

For multiple (panel) studies:

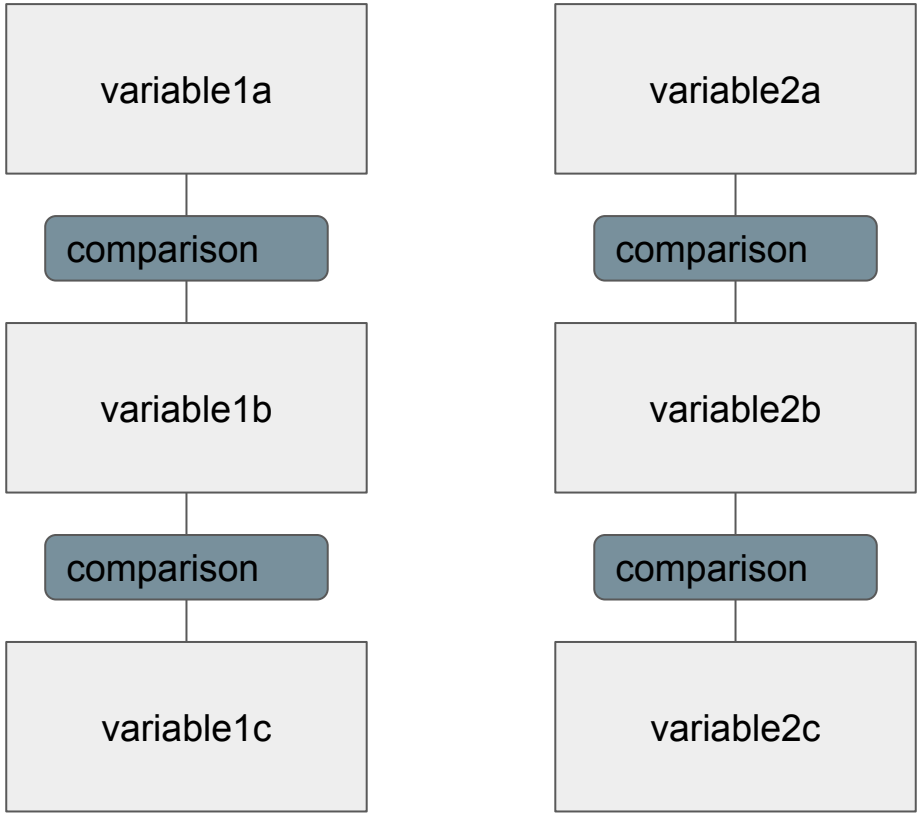
- Find related variables across studies

Linking variables: design options

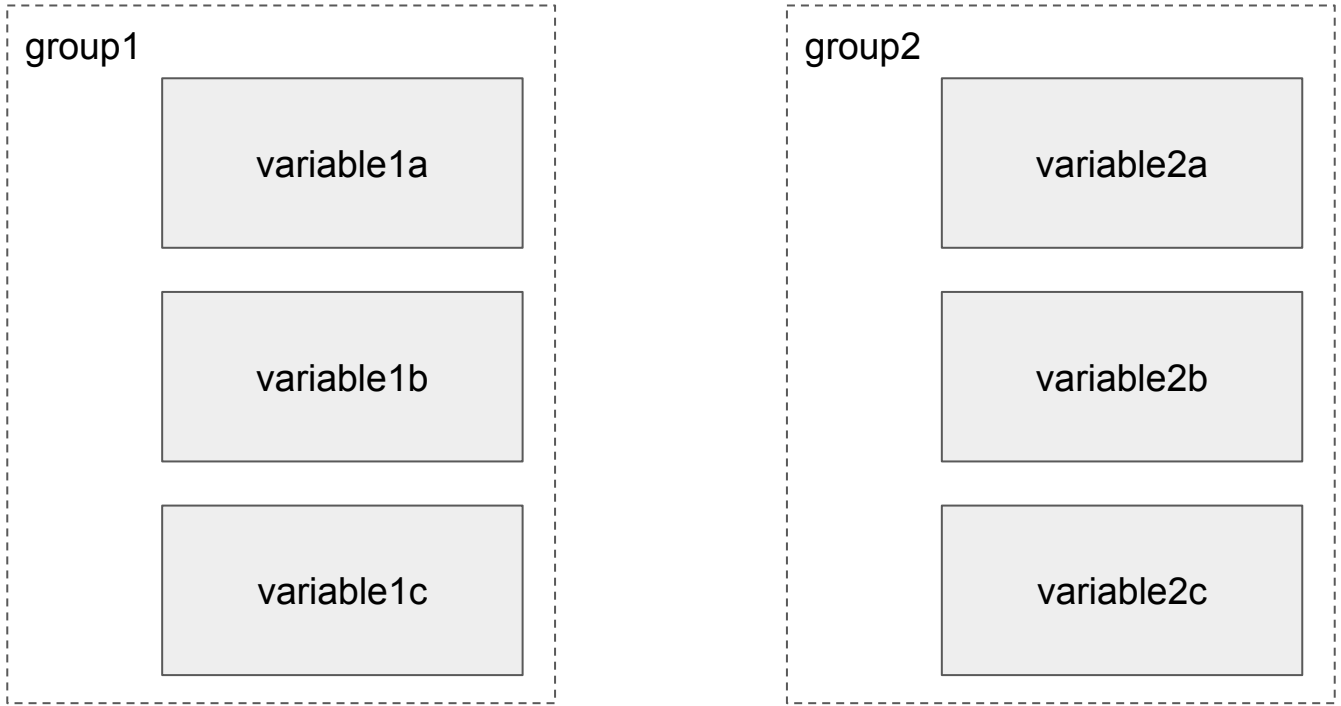
Versioning


variable1 (v1)
variable1 (v2)
variable1 (v3)
variable2 (v1)
variable2 (v2)
variable2 (v3)

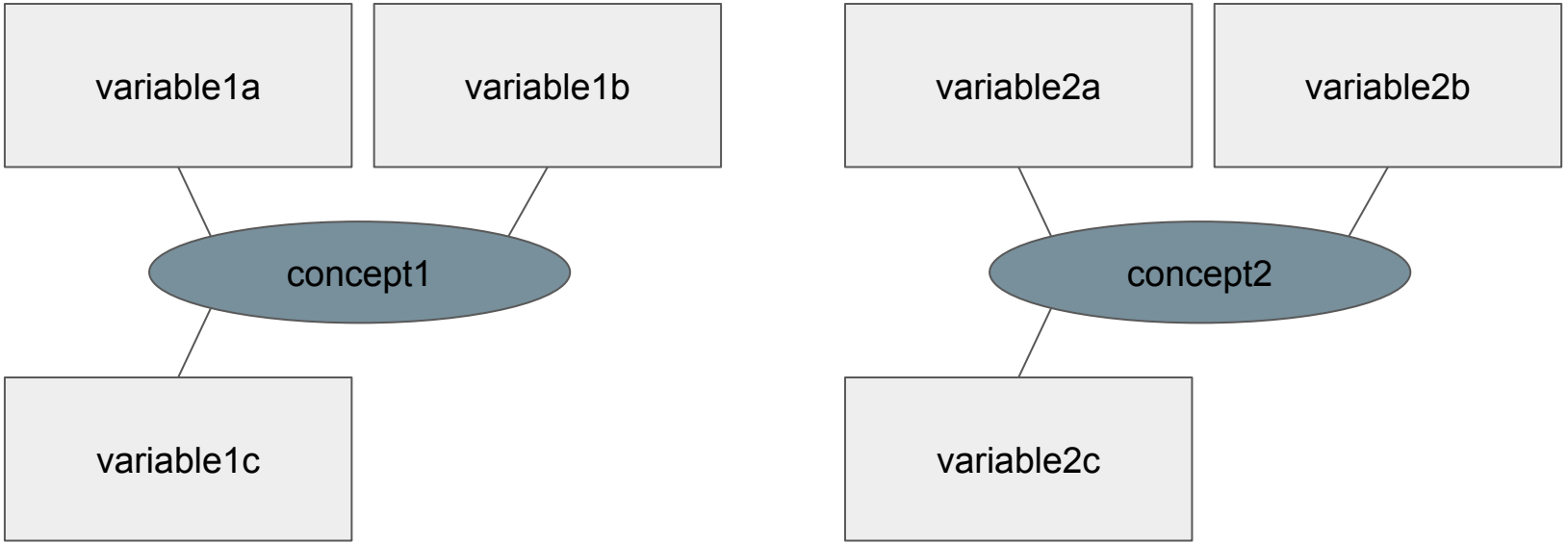
Direct links



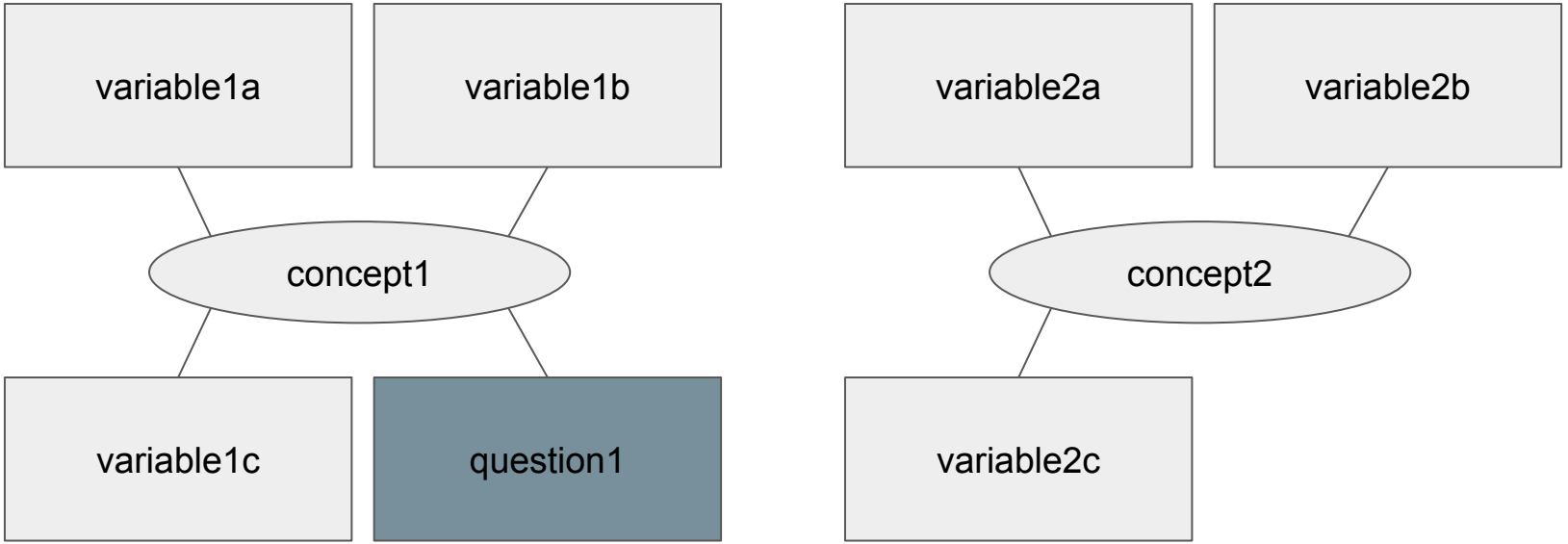
Groups


group1
variable1a
variable1b
variable1c
group2
variable2a
variable2b
variable2c

Concepts


variable1a
variable1b
concept1
variable1c
variable2a
variable2b
concept2
variable2c

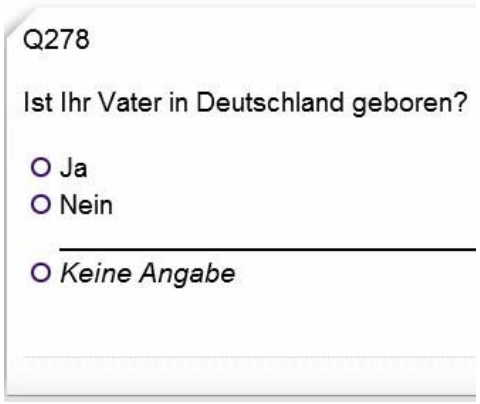
Concepts


variable1a
variable1b
concept1
variable1c
question1
variable2a
variable2b
concept2
variable2c

Questionnaire documentation

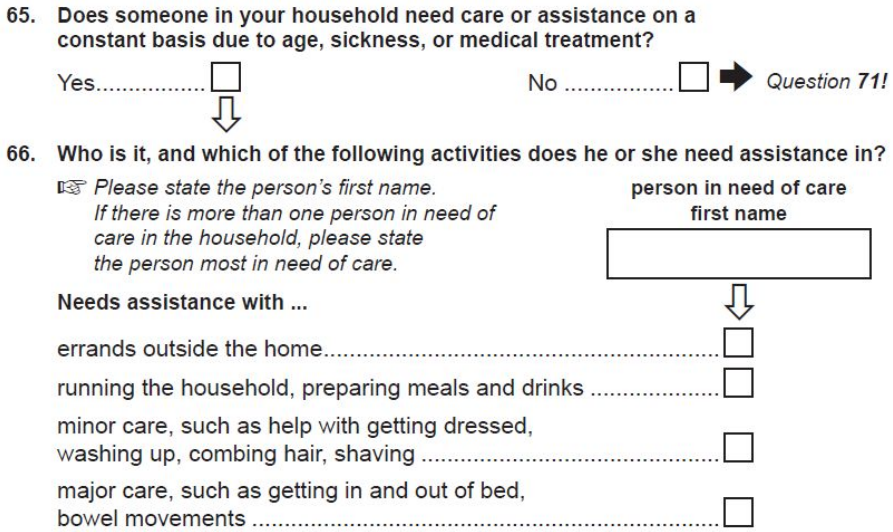
Questionnaire metadata

- Reference material
- Some information is essential (has to be preserved, depends on use case)
- Some information can be ignored (not captured by metadata)
- i18n: multilingual infrastructure (fieldwork and/or documentation)
- Re-use of information (next wave)



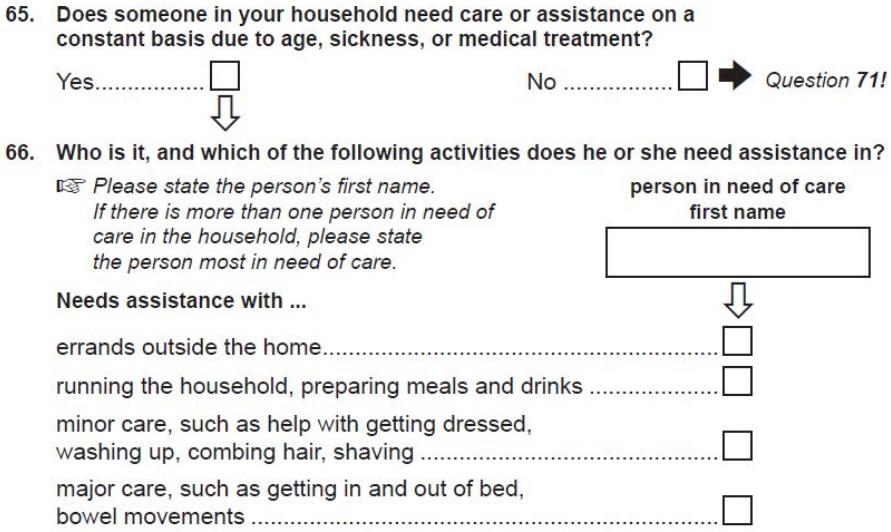
<< CAPI Screen

vv PAPI



(Not) preserved information

Source Material (Paper)



65. **Does someone in your household need care or assistance on a constant basis due to age, sickness, or medical treatment?**

Yes..... ☐ ⇩ No ..... ☐ ➡ *Question 71!*

66. **Who is it, and which of the following activities does he or she need assistance in?**

☞ *Please state the person's first name. If there is more than one person in need of care in the household, please state the person most in need of care.*

person in need of care first name

Needs assistance with ...

errands outside the home..... ☐

running the household, preparing meals and drinks ..... ☐

minor care, such as help with getting dressed, washing up, combing hair, shaving ..... ☐

major care, such as getting in and out of bed, bowel movements ..... ☐

Produced with Metadata

65 Does someone in your household need care or assistance on a constant basis due to age, sickness, or medical treatment?

Yes 1
No 2

65;hpfleg hpfleg 2 @ 71

65;hpfleg=1

66 Who is it, and which of the following activities does he or she need assistance in?

Please state the person's first name. If there is more than one person in need of care in the household, please state the person most in need of care.

person in need of care first name

66:hpnam hpnam

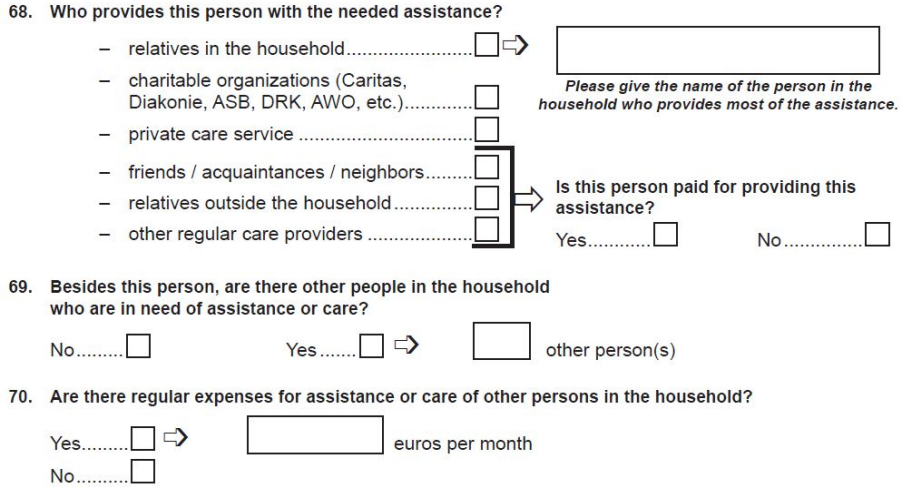
Needs assistance with ...

errands outside the home 1
running the household, preparing meals and drinks 1
minor care, such as help with getting dressed, washing up, combing hair, shaving 1
major care, such as getting in and out of bed, bowel movements 1

66:hhil1 hhil1
66:hhil2 hhil2
66:hhil3 hhil3
66:hhil4 hhil4

(Not) preserved information

Source Material (Paper)



68. Who provides this person with the needed assistance?

- relatives in the household ☐ ⇨ [ ]
- charitable organizations (Caritas, Diakonie, ASB, DRK, AWO, etc.) ☐
- private care service ☐
- friends / acquaintances / neighbors ☐
- relatives outside the household ☐
- other regular care providers ☐

Please give the name of the person in the household who provides most of the assistance.

⇨ **Is this person paid for providing this assistance?**

Yes ☐ No ☐

69. Besides this person, are there other people in the household who are in need of assistance or care?

No ☐ Yes ☐ ⇨ [ ] other person(s)

70. Are there regular expenses for assistance or care of other persons in the household?

Yes ☐ ⇨ [ ] euros per month

No ☐

68 Who provides this person with the assistance he / she needs?

relatives in the household	1
charitable organizations (Caritas, Diakonie, ASB, DRK, AWO, etc.)	1
private care service	1
friends / acquaintances / neighbors	1
relatives outside the household	1
other regular care providers	1

68:hhvon1 hhvon1
68:hhvon7 hhvon7
68:hhvon3 hhvon3
68:hhvon9 hhvon9
68:hhvon6 hhvon6
68:hhvon8 hhvon8

[relatives in the household] Please give us the name of the person in the household who is the main caregiver.

68:hhnam hhnam 68:hhvon1=1

[friends / acquaintances / neighbors, relatives outside the household,other regular care providers]Is this person paid for providing this assistance?

Yes 1
No 2

68:hhbez hhbez 68:hhvon9=1 | 68:hhvon6=1 | 68:hhvon8=1

69 Besides this person, are there other people in the household who are in need of assistance or care?

No 2
Yes 1

69;hpfleg2 hpfleg2

[Yes] ... other person(s)

69:hpfleg3 hpfleg3 69;hpfleg2=1 hpfleg2=1

70 Are there regular expenses for assistance or care of other persons in the household?

Yes 1
No 2

70;hpfleg4 hpfleg4

Yes ... euros per month

70:hpfleg5 hpfleg5 70;hpfleg4=1

Produced with Metadata >>

Example: What DDIonRails preserves and adds

Preserved:

- Question numbers
- Textual information (question texts, instructions, answers)
- Routing (logical: filter, goto)

Added:

- Values for answers
- Concepts
- Links to variables
- Translations

Not preserved:

- Layout (horizontal/vertical arrangement, text prior/after open ended questions)
- Typography (bold, underlined)
- Graphical information
- Routing (textual)

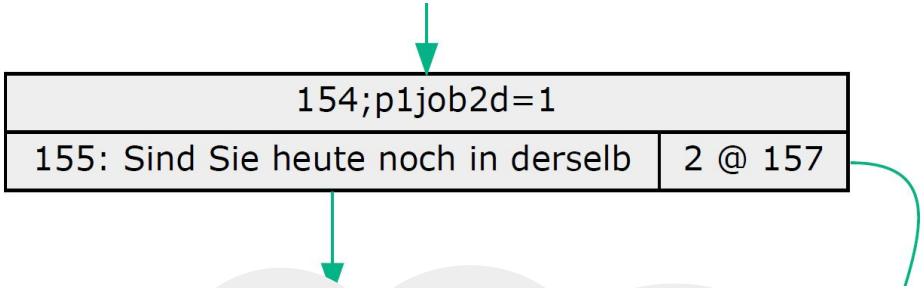
What information do you want to preserve?

Some notes on routing

- Common default: go to next question
  - No more specification needed
  - Exceptions needed

Two different approaches in instruments:

- Question's gatekeeper ("filter")
  - Defines the universe of this particular question
  - Condition which has to be true
- After a question ("goto")
  - Defines the way to the next question depending on the answer (and perhaps other information)



Which approach is used in your institution? What are your experiences?

What do data users like, what survey designers – and why?

Which approach is more, which is less parsimonious?

What about visualization?

Will it convert?

Routing in DDI

ControlConstruct:

Extensible structure for control elements used in describing flow logic within the instrument: IfThenElse, RepeatUntil, RepeatWhile, Loop, Sequence, ComputationItem, StatementItem, and QuestionConstruct. (from DDI 3.2 XML Schema Documentation)

Shortcoming:

Some kind of code needed to specify conditions and or calculations, but language not defined within DDI.

- Operators
- References

```
<d:IfThenElse>
    <d:IfCondition>
        <r:Code programmingLanguage="Neutral">Counter != 1</r:Code>
    </d:IfCondition>
    <d:ThenConstructReference>
        <r:ID>333ae135-784d-4435-9e54-...</r:ID>
    </d:ThenConstructReference>
</d:IfThenElse>
source: http://www.colectica.com/census2010-ddi-metadata (shortened, DDI 3.1)
```

Example: Routing in DDIonRails

- Each item (one item is related to one variable) in a question can have a filter and a goto.
- A filter can have references to one or more (prior) items in the conditions.

- Gotos only evaluate the answer of this item and direct to the appropriate next answer.
- Room for improvement (e.g. loops), but works!

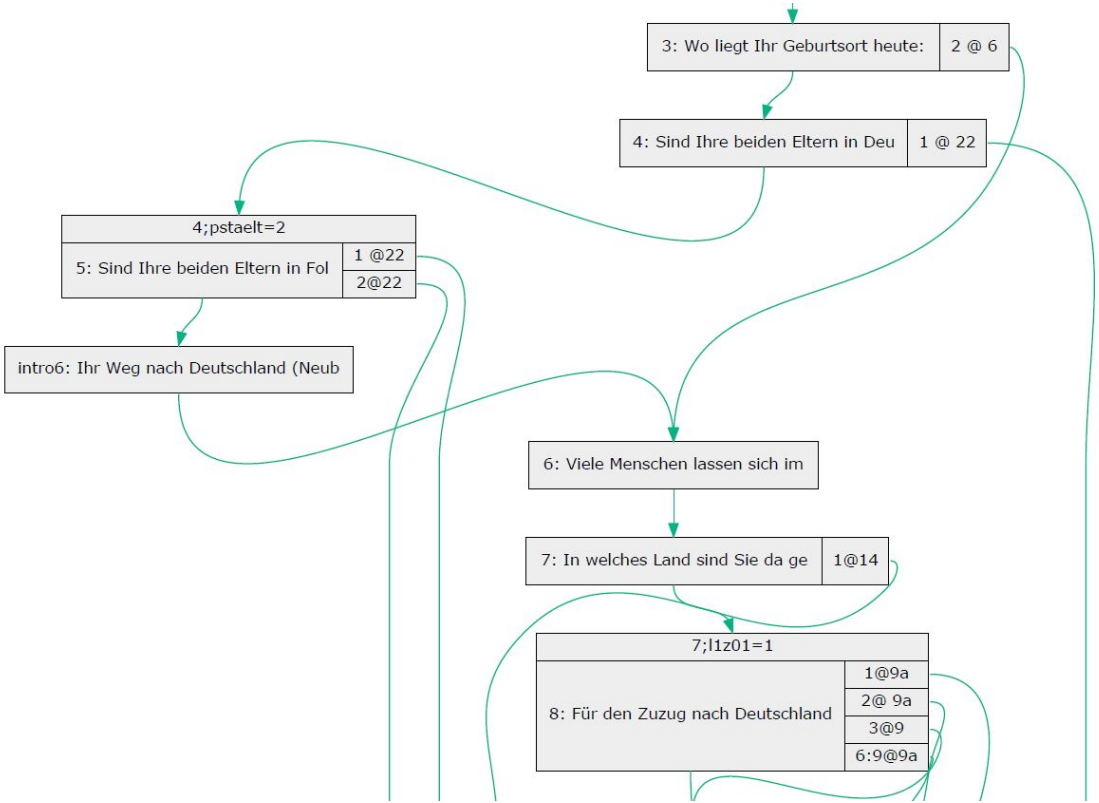
<< Screenshot:
http://ddionrails.org/imports/questions_csv.html

Rules for filter and goto

Filter and goto definitions consist of question names and symbols only, no keywords (e.g. “goto”) are used.

- Symboles `( ) = < > @ | & : != <= >=`
- Filter `(AGE > 20) & (SEX = 1)` means: this question is asked if “age” is greater than 20 and “sex” is 1
- Goto `(2 @ TARGET)` means: if the answer to the current question is 2 then go to question “target”
- Refer to items using the colon as a seperator, e.g. `(PSOR:2 = 3)`.
- Value lists and ranges: `(x = 1:3)` is equal to `(x = 1,2,3)` is equal to `(x = 1) | (x = 2) | (x = 3)`

Example: Visualise routing



- Flow chart, algorithmic derived from DDIonRails metadata
- Filters displayed
- Gotos parsed
- Layout/rendering by Graphviz

How is filter/goto-approach connected with visualisation?

Make information re-usable and deal with changes

Re-use:

- Means: Combine parts of a question and give them an identificator, which has to be used if the question appears again.
- Tracks permanence.
- Helps to limit amount of information, which has to be managed (entered, translated).
- Makes things more complicated: one more relation.
- Agency needed: assign IDs, ensure integrity, supervise corrections (internal question bank)

Link over time:

- Same methods like those presented for variables
- Comparison seems to be more appropriate

Which parts of a question do you/ would you make re-useable?

How many resources do you have to track and describe changes?

Working session

Working session

Go into smaller groups

Questions:

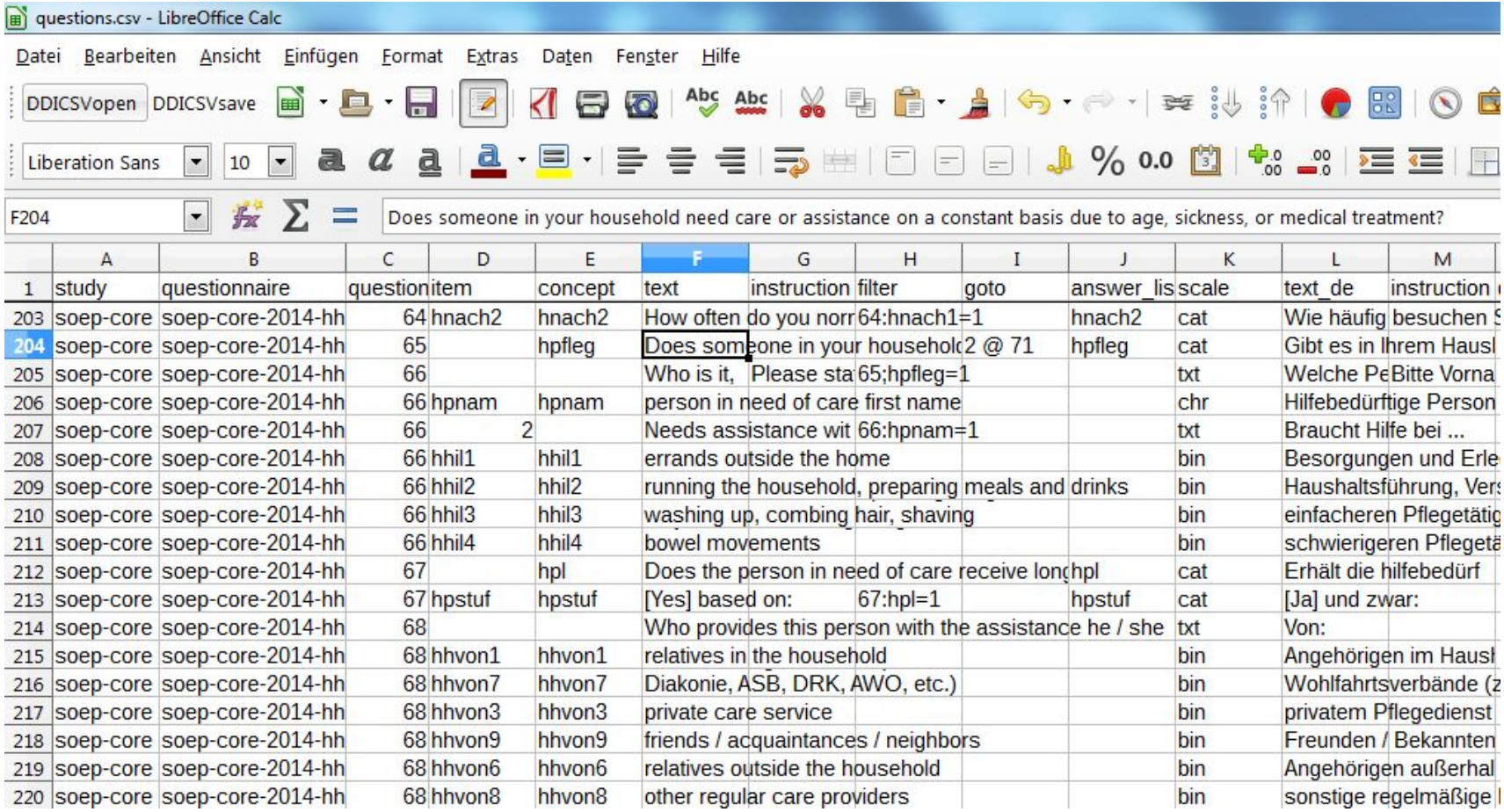
1. What is your use case?
2. How do you manage variables and questions?
3. Are your solutions interoperable?
4. Which of the solutions (regarding variables) would work for you?

Implementation and design choices

Implementation and design choices

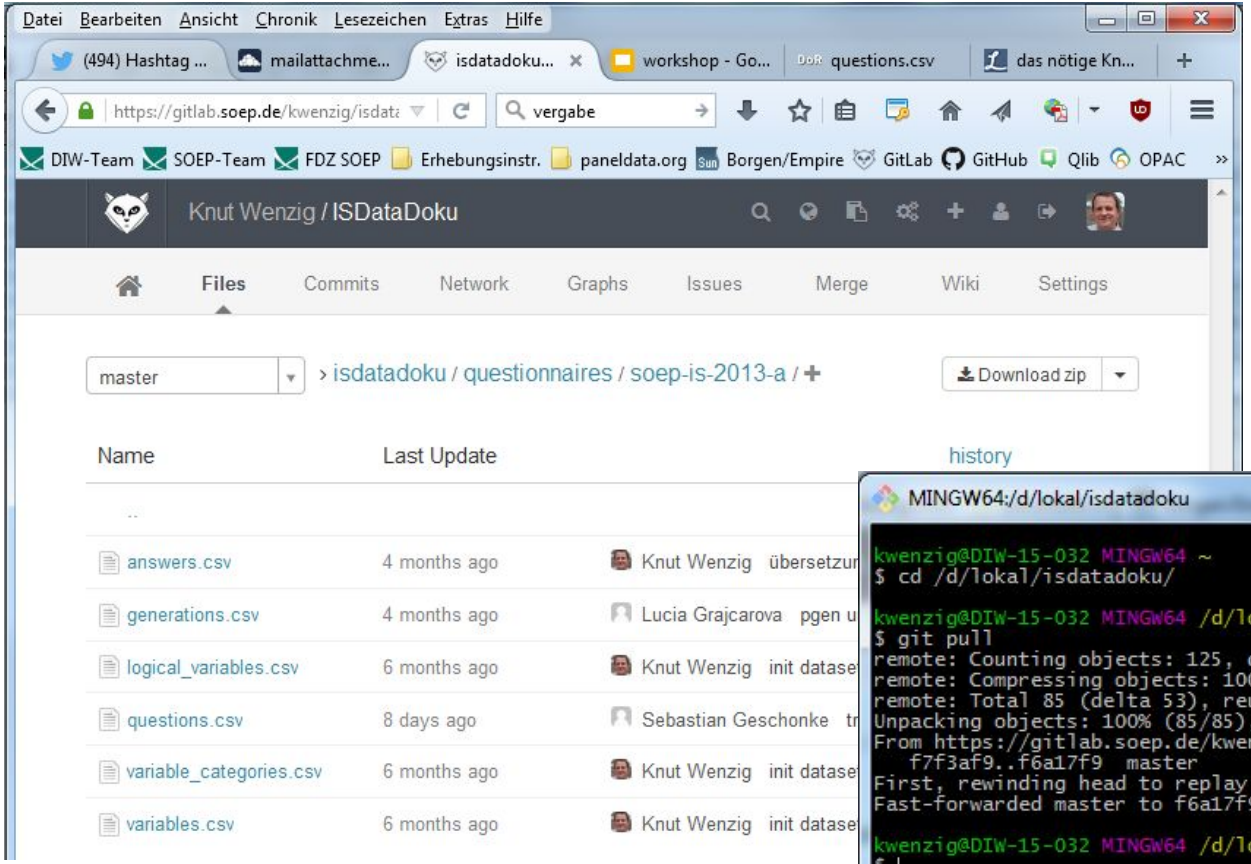
- XML vs CSV vs other formats
- Relational databases
- Do you have software developers
- Who is managing your metadata
- How many studies / collaboration
- Git

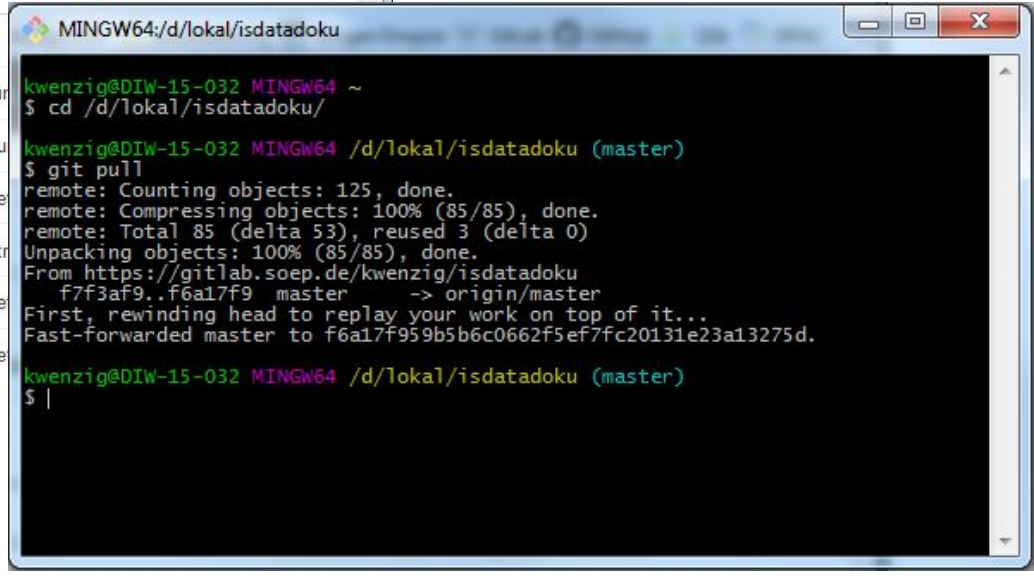
Use case: SOEP



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	study	questionnaire	question	item	concept	text	instruction	filter	goto	answer_lis	scale	text_de	instruction
203	soep-core	soep-core-2014-hh	64	hnach2	hnach2	How often do you norr		64:hnach1=1		hnach2	cat	Wie häufig	besuchen S
204	soep-core	soep-core-2014-hh	65		hpfleg	Does someone in your household			2 @ 71	hpfleg	cat	Gibt es in Ihrem Haush	
205	soep-core	soep-core-2014-hh	66			Who is it,	Please sta	65;hpfleg=1			txt	Welche Pe	Bitte Vorna
206	soep-core	soep-core-2014-hh	66	hpnam	hpnam	person in need of care	first name				chr	Hilfebedürftige Person	
207	soep-core	soep-core-2014-hh	66	2		Needs assistance wit		66:hpnam=1			txt	Braucht Hilfe bei ...	
208	soep-core	soep-core-2014-hh	66	hhil1	hhil1	errands outside the home					bin	Besorgungen und Erle	
209	soep-core	soep-core-2014-hh	66	hhil2	hhil2	running the household, preparing meals and				drinks	bin	Haushaltsführung, Vers	
210	soep-core	soep-core-2014-hh	66	hhil3	hhil3	washing up, combing hair, shaving					bin	einfacheren Pflegetätig	
211	soep-core	soep-core-2014-hh	66	hhil4	hhil4	bowel movements					bin	schwierigeren Pflegetä	
212	soep-core	soep-core-2014-hh	67		hpl	Does the person in need of care receive long				hpl	cat	Erhält die hilfebedürf	
213	soep-core	soep-core-2014-hh	67	hpstuf	hpstuf	[Yes] based on:		67:hpl=1		hpstuf	cat	[Ja] und zwar:	
214	soep-core	soep-core-2014-hh	68			Who provides this person with the assistance				he / she	txt	Von:	
215	soep-core	soep-core-2014-hh	68	hhvon1	hhvon1	relatives in the household					bin	Angehörigen im Haush	
216	soep-core	soep-core-2014-hh	68	hhvon7	hhvon7	Diakonie, ASB, DRK, AWO, etc.)					bin	Wohlfahrtsverbände (z	
217	soep-core	soep-core-2014-hh	68	hhvon3	hhvon3	private care service					bin	privatem Pflegedienst	
218	soep-core	soep-core-2014-hh	68	hhvon9	hhvon9	friends / acquaintances / neighbors					bin	Freunden / Bekannten	
219	soep-core	soep-core-2014-hh	68	hhvon6	hhvon6	relatives outside the household					bin	Angehörigen außerhal	
220	soep-core	soep-core-2014-hh	68	hhvon8	hhvon8	other regular care providers					bin	sonstige regelmäßige	

Screenshot: Questionnaire in LibreOffice Calc with two new buttons





Screenshots: Git Bash and Gitlab

CSV files on Git

Pro:

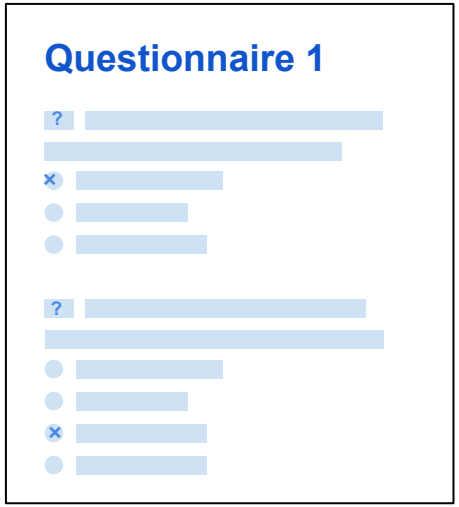
- No server only software on clients needed (but Gitlab or similar make things easier)
- No special frontend (“editor”) needed: lean development
- Version control helps to track changes and reset to previous version in case of errors
- Metadata easy accessible for programming (Ruby, R, Stata)
- Establish version control know-how

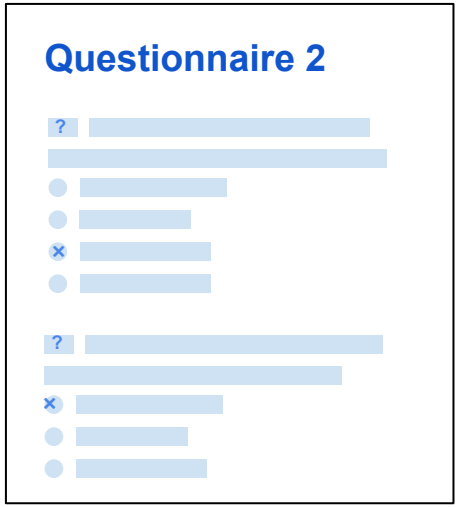
Con:

- Integrity of metadata not enforced
- Annoying issues with seperators, encoding, quotes (LibO Calc and a macro helps)
- Transfer to database (for web-use)

Do you already use version control?

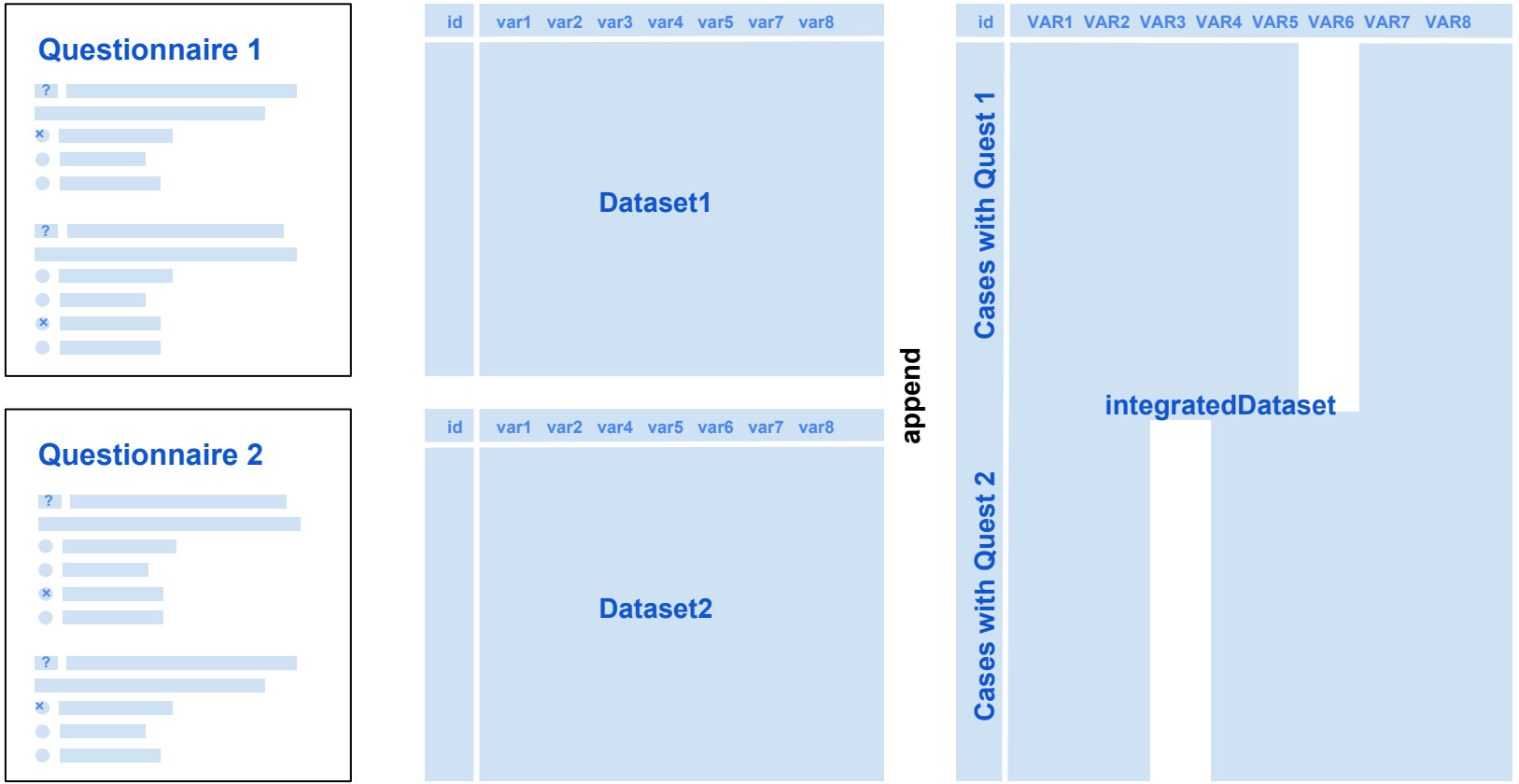
Use case: FiD integration





- pool information of very similar studies which were carried out in the same year
  - SOEP
  - Families in germany
- very similar: integration of datasets from different waves
- integration reduces burden of data users dramatically
  - identification of similar questions/variables
  - harmonisation of information is standardised

Use case: FiD integration



Use case: FiD integration – necessary steps

- Identify corresponding questions/variables
- Correct, (harmonise)
- Rename variables:
  - Dataset1, var1 > integratedDatset, VAR1
  - See table
- Compare corresponding variables
  - Prevent errors
  - Variable labels
  - Value labels
  - Accept differences or make corrections
- Append datasets
  - Fill sparse areas with missing code
- Evaluate work
- (Harmonise)

i_dataset	i_variable	o_dataset	o_variable
Dataset1	var1	iDataset	VAR1
Dataset1	var2	iDataset	VAR2
Dataset2	var1	iDataset	VAR1
Dataset2	var2	iDataset	VAR2

Use case: FiD integration – dorevaluate

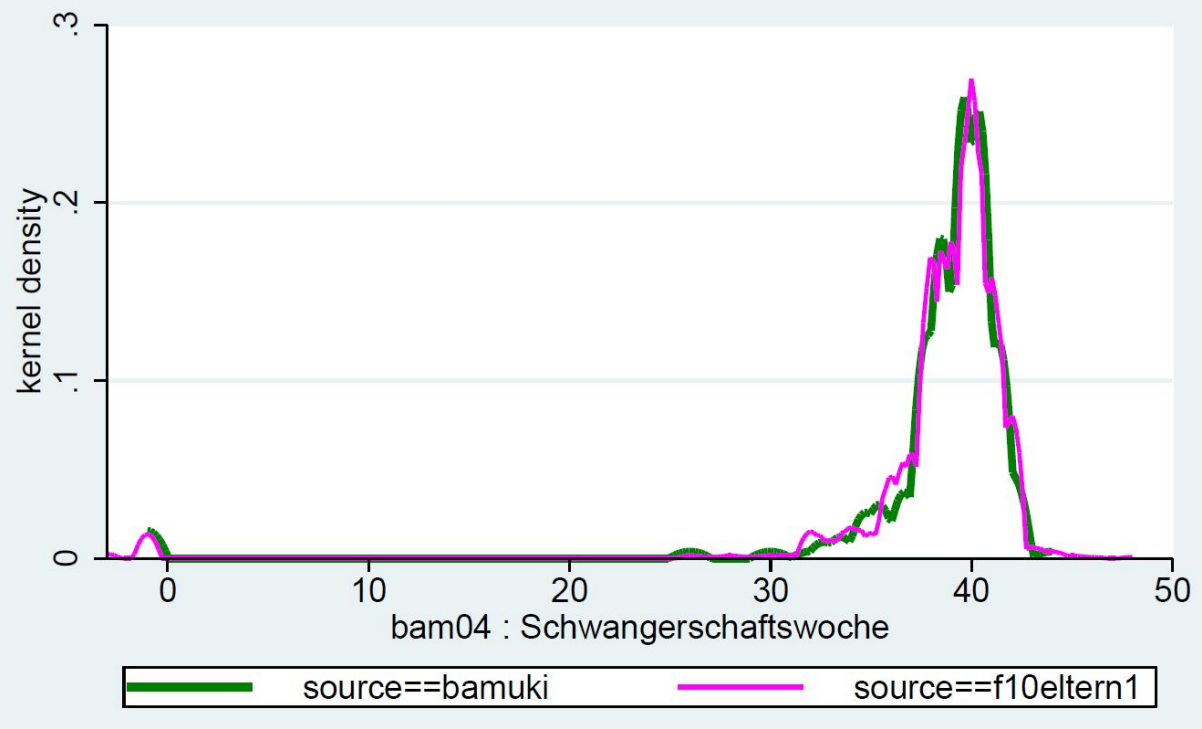
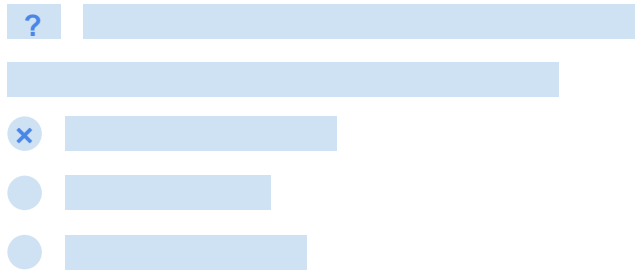


Diagramm shows successful integration of a metric variable (week of pregnancy)

Use case: FiD integration – documentation

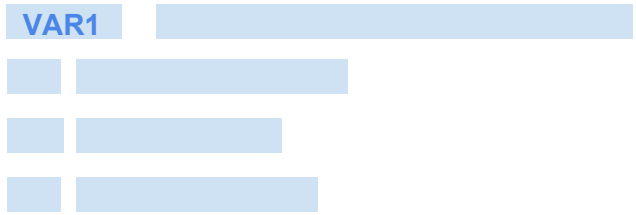
Questionnaire 1



See:
Dataset1, var1
integratedDatatset, VAR1

original renaming information – nothing more

Codebook: integratedDataset



Sources:
Questionnaire 1, Q12
Questionnaire 2, Q14

Use case: FiD integration with active metadata

Result:

- 62 Stata files with integrated information
- 305 lines of code (without corrections)
- 21915 (non-)renaming of variables
- 61464 differences in variable labels and value labels were accepted

- Stata ados which rely on DDIonRails metadata: http://ddionrails.org/stata/
  - dorrename, dorcomparedta, dorcomparexls, dorappend, dorevaluate
  - dororder, dorlabeldta
  - https://github.com/ddionrails/stata

Metadata driven data processing

- code written for data preparation more structured and better to maintain
- metadata (and documentation) more accurate
- documentation ready when data are ready

paneldata.org

- Successor for SOEPinfo
- Multiple studies
- Multiple releases / distributions
- Linking across studies
- Panel-specific functionality

DDI on Rails

Understanding Data

DDI on Rails

Understanding Data

→ The data portal DDI on Rails accompanies researchers throughout the entire course of their research projects from conception to publication/citation.

→ The system offers researchers the possibility to explore the data, to compile personalized datasets, and to publish results on the publication database.

→ In contrast to similar products, DDI on Rails is study-independent and open-source, is able to document data with multiple versions/distributions and the specific characteristics of a longitudinal study, and is easy to use.

paneldata.org beta ...

https://paneldata.org

Search

paneldata.org Studies Topics Publications

Search

paneldata.org is based on DDI on Rails, designed and built by Marcel Hebing.

Search

satisfaction life

Search | Clear search

All | Concepts | Variables | Questions | Publications

Study

- SOEP Core study 120
- Pairfam 40
- SOEP Pretest 32
- SOEPlong 11
- IAB-SOEP Migration Sample 10
- SOEP Innovation Sample 7

Class

- Publication 261
- Variable 206
- Concept 26
- Question 14

507 results.

Virtual Life Satisfaction - KYKLOS
Publication by **Castronova, Edward J. and Gert G. Wagner (2011)**

In conclusion, we would like to ask you about your satisfaction with your life in general. [M19]
Question in study: **soep-pretest** | questionnaire: **pre_2008_mukl_b** | period: **2008** | analysis unit: **p**

In conclusion, we would like to ask you about your satisfaction with your life in general. [M19]
Question in study: **soep-pretest** | questionnaire: **pre_2008_mukl_c** | period: **2008** | analysis unit: **p**

In conclusion, we would like to ask you about your satisfaction with your life in general. [110]
Question in study: **soep-pretest** | questionnaire: **pre_2010** | period: **2010** | analysis unit: **p**

In conclusion, we would like to ask you about your satisfaction with your life in general. [199]
Question in study: **soep-pretest** | questionnaire: **pre_2011** | period: **2011** | analysis unit: **p**

Satisfaction [item_5974]
Concept in topic: **satisfaction**

Unemployment, recall error, and life satisfaction
Publication by **Jürges, Hendrik (2004)**

Artists' Job and Life Satisfaction (Master-Thesis)

Search

satisfaction life

Search | Clear search

All | Concepts | Variables | Questions | Publications

Study
SOEP Core study (x)

Analysis unit
individual level 5

Conceptual dataset
Original data (net) (x)

Period
2011 (x)

5 results.

Satisfaction With Family Life [bbp0110] bbp0110
Variable in study: **soep-core** | dataset: **bbp** | period: **2011** | analysis unit: **p** | concept: **pzuf13**

Satisfaction With Social Life [bbp0111] bbp0111
Variable in study: **soep-core** | dataset: **bbp** | period: **2011** | analysis unit: **p** | concept: **pzuf14**

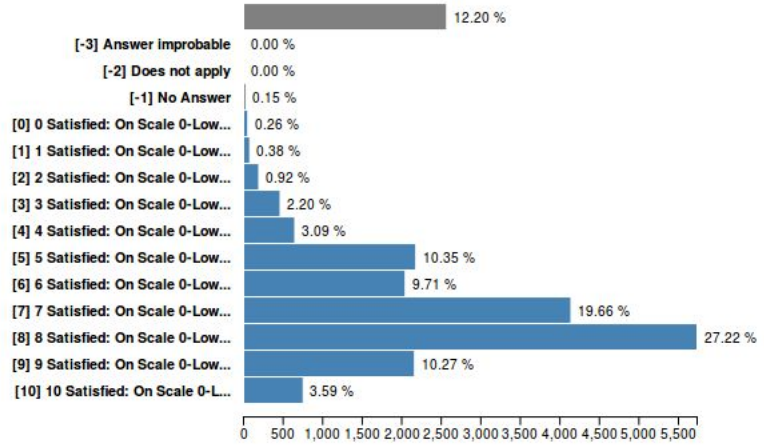
Current Life Satisfaction [bbj99] bbj99
Variable in study: **soep-core** | dataset: **bbpage17** | period: **2011** | analysis unit: **p** | concept: **item_5423**

Satisfaction With Life At Today [bbp15201] bbp15201
Variable in study: **soep-core** | dataset: **bbp** | period: **2011** | analysis unit: **p** | concept: **pzule1**

Satisfaction With Life In 5 Years [bbp15202] bbp15202
Variable in study: **soep-core** | dataset: **bbp** | period: **2011** | analysis unit: **p** | concept: **pzule2**

Satisfaction With Life At Today [variable: bbp15201]

Categories



Statistics

invalid	2602.0
valid	18467.0

Related variables

Current Life Satisfaction [bcp151] bcp151

Variable in study: **soep-core** | dataset: **bcp** | period: **2012** | analysis unit: **p** | concept: **pzule1**

Variable

Study:
SOEP Core study [soep-core]

Dataset:
Personal questionnaire [bbp]

Concept:
Current Life Satisfaction [pzule1]

Number of Categories:
15

→ Go to dataset
→ Go to concept
Frequencies
Question
Generated variables
Related variables 29

SOEP Mini- Job, Midi- Job [variable: bcp52]

Compare category labels

Variable:	bcp52	bbp65	bap52	zp63	yp61	xp65	wp52	vp63	up51	tp7004	sp52a	rp51
Dataset:	bcp	bbp	bap	zp	yp	xp	wp	vp	up	tp	sp	rp
Period:	2012	2011	2010	2009	2008	2007	2006	2005	2004	2003	2002	2001
[x] answer improbable	-3 (0)	-3 (0)	-3 (0)	-3 (0)	-3 (0)	-3 (0)	-3 (0)	-3 (0)	-3 (0)	-3 (0)	-3 (0)	-3 (0)
[x] does not apply	-2 (18310)	-2 (18646)	-2 (16594)	-2 (18376)	-2 (17436)	-2 (18512)	-2 (19905)	-2 (18807)	-2 (19630)	-2 (20165)	-2 (21335)	-2 (20037)
[x] no answer	-1 (33)	-1 (44)	-1 (51)	-1 (55)	-1 (48)	-1 (51)	-1 (54)	-1 (55)	-1 (52)	-1 (73)	-1 (98)	-1 (78)
[x] yes mini-job	1 (966)	1 (915)	1 (886)	1 (941)	1 (867)	1 (885)	1 (946)	1 (863)	1 (897)			
[x] yes midi job	2 (215)	2 (207)	2 (197)	2 (191)	2 (186)	2 (182)	2 (161)	2 (172)	2 (173)			
[x] no	3 (1282)	3 (1257)	3 (1185)	3 (1229)	3 (1147)	3 (1256)	3 (1292)	3 (1208)	3 (1267)	2 (1248)	2 (1615)	2 (1393)
[x] yes										1 (1125)	1 (844)	1 (843)

Label: Mini-Job, Midi-Job

Categories:

- [-3] nicht valide
- [-2] trifft nicht zu
- [-1] keine Angabe
- [1] Ja, Mini-Job
- [2] Ja, Midi-Job
- [3] Nein

Mini- Job, Midi- Job [zp63] zp63
Variable in study: **soep-core** | dataset: **zp** | period: **2009** | analysis unit: **p** | concept: **paz11**

Mini- Job, Midi- Job [yp61] yp61
Variable in study: **soep-core** | dataset: **yp** | period: **2008** | analysis unit: **p** | concept: **paz11**

Mini- Job, Midi- Job [xp65] xp65

Generated variables
Related variables 11
Compare category labels

Details

Current Life Satisfaction [concept: pzule1]

All topics
attitudes, values, and personality
personality
satisfaction

Concept

Name:
pzule1

Title:
Current Life Satisfaction

Topic:
satisfaction

- SOEPlong 1
- Base II 3
- SOEP Core study 35
- SOEP Pretest 24

paneldata.org

- hosted service to document panel data
- study-specific domains (e.g., paneldata.org/soep-core)
- about.paneldata.org for additional information

Wrap up

License