SOEP Das Sozio-oekonomische Panel

DDI – more than just an XML-metadata-standard

Marcel Hebing (DIW Berlin)

Vancouver, April 2014

Two separate thoughts

1. DDI is more than the XML implementation.

2. DDI is more than an standard for metadata.

Unique selling proposition: **the community.**

Agenda

Introduction

Part 1: Alternatives to the XML Implementation

Part 2: More than metadata

Conclusion

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`

DDI on Rails

Understanding Data

Vision. 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 SOEP data, to compile personalized datasets, and to publish results on the publication database.

`http://www.ddionrails.org`

DDI on Rails – characteristics

- study-independent and open-source
- longitudinal data and multiple studies
- metadata search and comparison
- basket and script generator

SOEPinfo (studies) - Mozilla Firefox

SOEPinfo (studies)

hewing.soep.de

Google

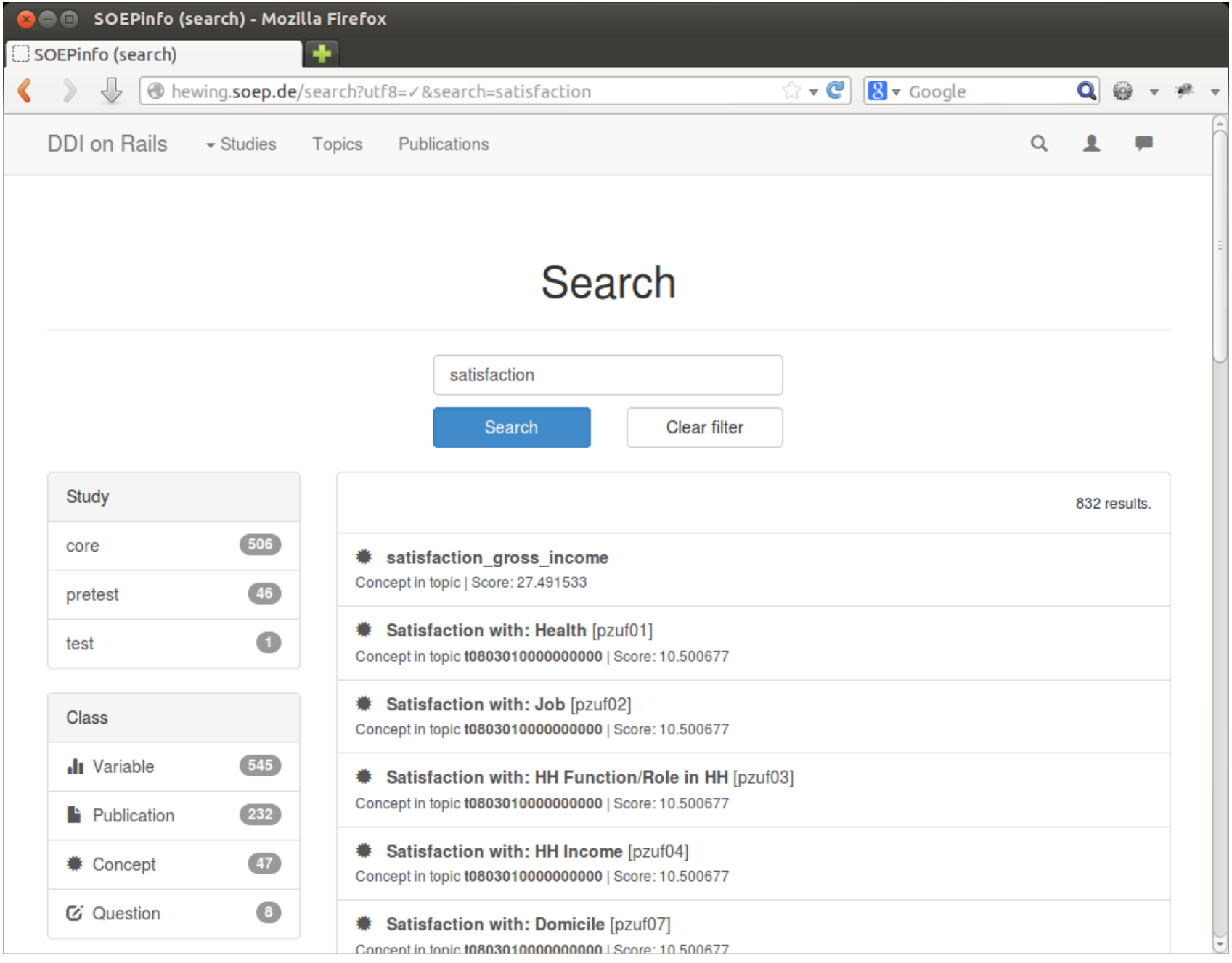
DDI on Rails Studies Topics Publications

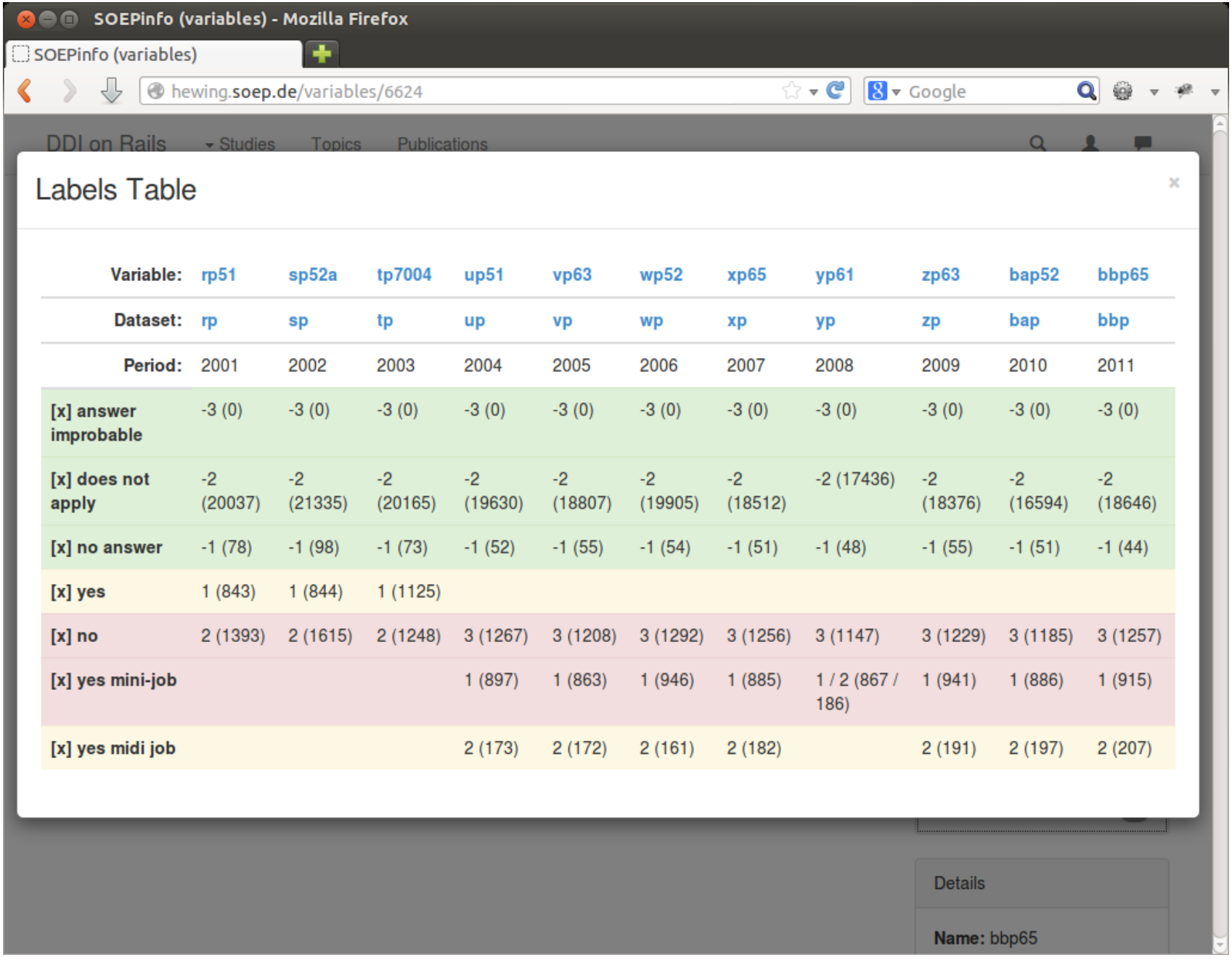
DDI on Rails

Understanding Data

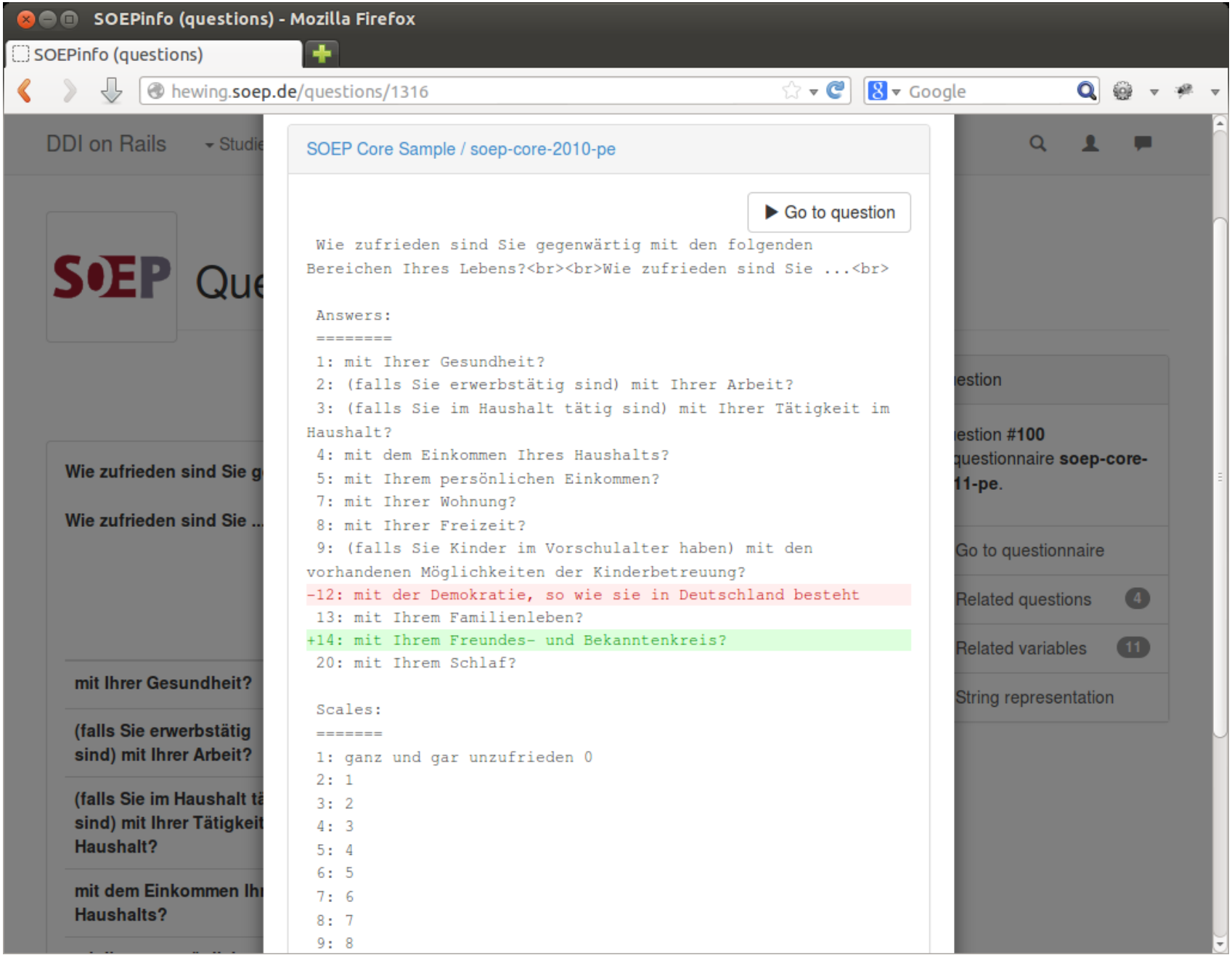
Search

Studies	Topics	News from the website
SOEP Long Format	not assigned!	Jobverlust der Mutter kann Entwicklung


SOEPinfo (search) - Mozilla Firefox
SOEPinfo (search)
hewing.soep.de/search?utf8=✓&search=satisfaction
Google
DDI on Rails
Studies
Topics
Publications
Search
satisfaction
Search
Clear filter
Study
core 506
pretest 46
test 1
Class
Variable 545
Publication 232
Concept 47
Question 8
832 results.
satisfaction_gross_income
Concept in topic | Score: 27.491533
Satisfaction with: Health [pzuf01]
Concept in topic t0803010000000000 | Score: 10.500677
Satisfaction with: Job [pzuf02]
Concept in topic t0803010000000000 | Score: 10.500677
Satisfaction with: HH Function/Role in HH [pzuf03]
Concept in topic t0803010000000000 | Score: 10.500677
Satisfaction with: HH Income [pzuf04]
Concept in topic t0803010000000000 | Score: 10.500677
Satisfaction with: Domicile [pzuf07]



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


SOEPinfo (questions) - Mozilla Firefox
SOEPinfo (questions)
hewing.soep.de/questions/1316
Google
DDI on Rails
Studie
SOEP Que
Wie zufrieden sind Sie g
Wie zufrieden sind Sie ...
mit Ihrer Gesundheit?
(falls Sie erwerbstätig sind) mit Ihrer Arbeit?
(falls Sie im Haushalt tä sind) mit Ihrer Tätigkeit Haushalt?
mit dem Einkommen Ihr Haushalts?
SOEP Core Sample / soep-core-2010-pe
▶ Go to question
Wie zufrieden sind Sie gegenwärtig mit den folgenden Bereichen Ihres Lebens?

Wie zufrieden sind Sie ...
Answers:
========
1: mit Ihrer Gesundheit?
2: (falls Sie erwerbstätig sind) mit Ihrer Arbeit?
3: (falls Sie im Haushalt tätig sind) mit Ihrer Tätigkeit im Haushalt?
4: mit dem Einkommen Ihres Haushalts?
5: mit Ihrem persönlichen Einkommen?
7: mit Ihrer Wohnung?
8: mit Ihrer Freizeit?
9: (falls Sie Kinder im Vorschulalter haben) mit den vorhandenen Möglichkeiten der Kinderbetreuung?
-12: mit der Demokratie, so wie sie in Deutschland besteht
13: mit Ihrem Familienleben?
+14: mit Ihrem Freundes- und Bekanntenkreis?
20: mit Ihrem Schlaf?
Scales:
=======
1: ganz und gar unzufrieden 0
2: 1
3: 2
4: 3
5: 4
6: 5
7: 6
8: 7
9: 8
estion
estion #100
questionnaire soep-core-11-pe.
Go to questionnaire
Related questions 4
Related variables 11
String representation

Check it out!

https://data.soep.de

Lessons learned #1: When it comes to using a metadata standard. . .

humans are the most expensive and very limited resource.

Problems when using the DDI-XML-standard

- Reseachers work with tables (Stata, SPSS, R, Excel).
- DDI-L is too complicated.
- Researchers don't care about metadata, unless there are benefits.
- Many editors (in particular students), short training periods.

Part 1: Alternatives to the XML Implementation

Common data types in most programming languages

- boolean
- integer / numeric / float
- character / string

- collection / array
- key:value / hash / list / object

Problem with XML

- It mixes arrays and lists.
- "keys" are no longer unique:
  1. Attributes and elements might have the same name.
  2. Multiple elements with the same name are valid.

- Most programming languages have arrays and lists as native data structures, but they don't have a structure like XML.

This is perfectly valid XML

```
<individual name="Peter">
  <name>Max</name>
  <name>David</name>
</individual>
```

$\rightarrow$ XML requires a lot of mapping.

Small example (XML)

```
<dataset name="dta">
  <variable>var1</variable>
  <variable>var2</variable>
  <variable>var3</variable>
</dataset>
```

Small example (YAML)

```
dataset:
    name: dta
    variables:
        - var1
        - var2
        - var3
```

Small example (JSON)

```
"dataset":{"name:"dta","variables":["var1", "var2", "var3"]}
```

Parse XML

```
# Load package
require 'nokogiri'

# Read XML
xml_file = '<dataset name="dta"><variable>var1</variable>' +
  '<variable>var2</variable><variable>var3</variable></dataset>'

# Parse XML
xml = Nokogiri::XML.parse(xml_file)
obj_2 = {}
obj_2["name"] = xml.css("dataset").first.attr("name")
obj_2["variables"] = []
xml.css("variable").each do |variable|
  obj_2["variables"] << variable.text.strip
end
```

Parse JSON

```
# Load package
require 'json'

# Read JSON
json = '{"name":"dta","variables":["var1", "var2", "var3"]}'

# Parse JSON
obj_1 = JSON.parse(json)
```

Parsing XML and JSON

- XML: 2 + 7 lines of code (Ruby)
- JSON: 2 + 1 line of code (Ruby)

Small example (CSV)

```
dataset,variable
dta,var1
dta,var2
dta,var3
```

Parse

```
dta <- read.csv("variables.csv")
```

Parsing XML, JSON, and CSV

- XML: 2 + 7 lines of code (Ruby)
- JSON: 2 + 1 line of code (Ruby)
- CSV: 1 line of code (R)

In the case that size matters

- JSON: 60 % of XML
- CSV: 40 % of XML

CSV

- Very efficient: editing metadata is up to 40 times faster than using other technologies.
- Good tools (Excel, LibreOffice, R, Stata, SPSS, . . .).
- Easy to use (researchers and students know these tools).
- Very good data quality (editors understand the structure).
- Easy to validate (using unit tests or statistical packages).
- Adding new fields if necessary.
- The structure can correspond to a relational database.
- It becomes easy to analyse metadata.

Alternatives

- YAML and JSON,
- CSV and relational databases,
- and many others.

Part 2: More than metadata

Lessons learned #2:

Metadata are worthless without the research data they describe.

Metadata and proprietary data formats

- Stata and SPSS include some metadata (like labels).
- Proprietary formats might change at any time,
  not caring about interoperability.
- It's a weird combination of data and metadata.

Do we like to depend on proprietary formats?

»A **Simple Data Format** package contains:

- Data files in CSV
- (Minimal) dataset information in JSON (including a schema for the CSV)«

`http://dataprotocols.org/simple-data-format/`

(Open Knowledge Foundation)

What I like about this

1. Open standards, easy to implement.
2. Plain text, good for archiving.
3. Clear: separating and complementing.
4. Can be used for non-relational data (e.g. pictures).

data.csv

```
var1,var2,var3
A,1,2
B,3,4
```

datapackage.json

```
{ "name": "my-dataset",
  "resources": [
    { "path": "data.csv",
      "schema": {
        "fields": [
          { "name": "var1",
            "type": "string" },
          { "name": "var2",
            "type": "integer" },
          { "name": "var3",
            "type": "number" }
]}}]}
```

Idea: DDI Data Format

- Keep the CSV file.
- Use DDI-C in JSON format.
- Live happily ever after.

Finally. . .

My DDI Top 10

1. DDI-Community
2. Developers Group
3. “Concepts”
4. GLBPM
5. Workshops and conferences
6. The idea of the data lifecycle and the reuse of metadata
7. Metadata-driven data processing
8. DDI 1.2 (Nesstar Publisher subset)
9. Data management
10. DDI-C and DDI-L

Thank you.