

KonsortSWD



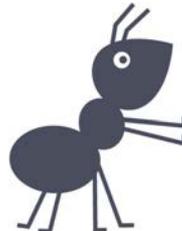
Konsortium für die
Sozial-, Verhaltens-, Bildungs- und
Wirtschaftswissenschaften

Xiaoyao Han,
Claudia Saalbach &
Knut Wenzig
DIW, Berlin

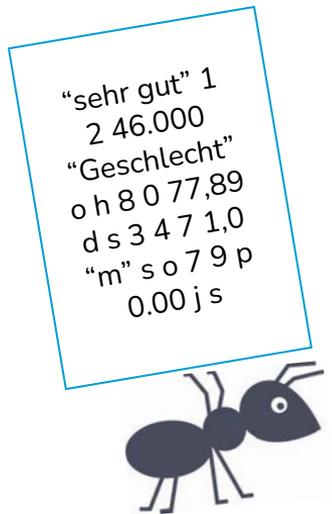


Open Data Format TA.3-M.5

KonsortSWD Community
Conference, 07.04.2022



Inhalt



- Metadaten
- Erwartungen
- Problem Statement
- Ziel
- Features
- Spezifikation

Metadaten

Studie

DZHW

Erhebungszeitraum: 01.03.2014 - 08.01.2019
 Wellen: 2
 Erhebungsdatenart: Quantitative Daten
 Daten verfügbar auf: Deutsch
 DOI: 10.21249/DZHW-gra20133.0.0
 Veröffentlicht am: 04.10.2021
 Version: 3.0.0 (aktuell)
 Zugangsweg:

In den Warenkorb legen Zum Warenkorb

Zielen...

Wenn Sie Probleme mit dem wenden Sie sich bitte an unser

DDI 3.1
 DDI 3.2
 DataCite (XML)
 DataCite (JSON)
 Schema.org

Messinstrument

Wie oft haben Sie sich ...

- ärgerlich gefühlt?	selten	mal	oft
- ängstlich gefühlt?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
- glücklich gefühlt?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
- traurig gefühlt?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

3. Haben Sie das Gefühl, dass das, was Sie in Ihrem Leben machen, wertvoll und nützlich ist?
(*) Antworten Sie bitte anhand der folgenden Skala, der Wert 0 bedeutet: überhaupt nicht wertvoll und nützlich. Mit den Werten dazwischen können Sie Ihre Einschätzung abstimmen.

überhaupt nicht wertvoll und nützlich vollkommen wertvoll und nützlich

4. Wie schätzen Sie sich persönlich ein?
(*) Antworten Sie bitte anhand der folgenden Skala, wobei der Wert 0 bedeutet: gar nicht risikobereit und der Wert 10 sehr risikobereit. Mit den Werten dazwischen können Sie Ihre Einschätzung abstimmen.

gar nicht risikobereit sehr risikobereit

5. Die folgenden Aussagen kennzeichnen verschiedene Einstellungen zum Leben und zur Zukunft.
(*) Antworten Sie bitte anwieser anhand einer Skala. Der Wert 1 bedeutet: stimme überhaupt nicht zu. Der Wert 7 bedeutet: stimme voll zu.

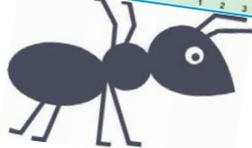
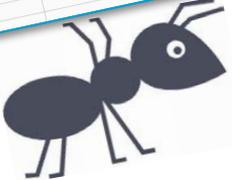
Stimme voll zu

Datensatz

bakj	jahr	geschlecht
1	1969	[1] Maennlich
2	1969	[1] Maennlich
3	1969	[1] Maennlich
4	1969	[2] Weiblich
5	1969	[1] Maennlich
6	1969	[2] Weiblich
7	1969	[2] Weiblich
8	1969	[1] Maennlich
9	1969	[2] Weiblich
10	1969	[2] Weiblich
11	1969	[1] Maennlich
12	1969	[2] Weiblich
13	1969	[2] Weiblich
14	1969	[2] Weiblich
15	1969	[2] Weiblich
16	1969	[1] Maennlich
17	1969	[2] Weiblich
18	1969	[1] Maennlich
19	1969	[1] Maennlich
20	1969	[1] Maennlich

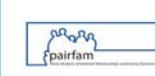
Software

Package: mypackage
 Title: What The Package Does (one line, title case required)
 Version: 0.1
 Authors: R
 Description: What the package does (one paragraph)
 Depends: R (>= 3.1.0)
 License: What license is it under?

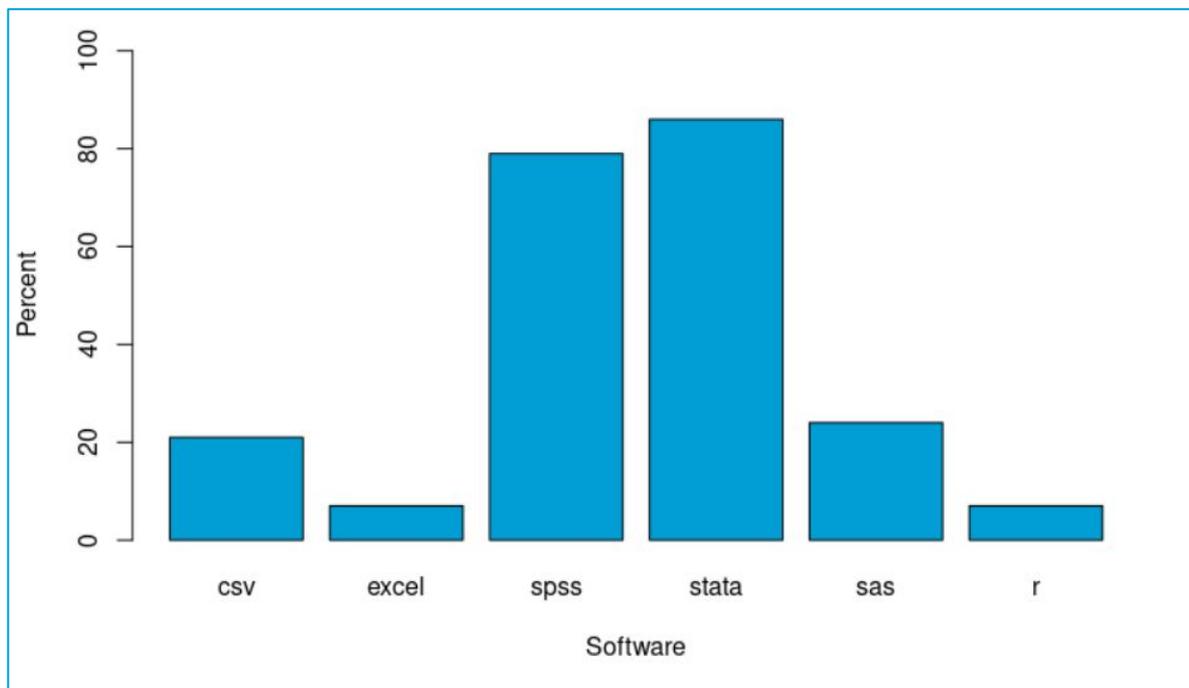

Erwartungen

Datenquellen- KonsortSWD Forschungsdatenzentren

 Federal Institute for Vocational Education and Training Research Data Centre	 ALBUS	 DEUTSCHE BUNDESBANK EUROSISTEM	 AGD	 IWH Forschungsdienstleistungen	 PIAAC	 SHARE Survey of Health, Ageing and Retirement in Europe 2007-2010/2011-2013	 SAFE Statistical Analysis for Research in Europe
 eLabour	 RESEARCH DATA CENTRE IAB of the German Federal Employment Agency IAB of the Institute for Employment Research IZA	 BZgA	 GePaRD	 IDSC INTERNATIONAL DATA SERVICE CENTER	 leibniz-psychology.org	 SOEP The Socio-Economic Panel	 International Survey Programs
 DeZIM FORSCHUNGS DATEN ZENTRUM	 Kraftfahrt-Bundesamt	 DJI Deutsches Jugendinstitut	 German Microdata Lab	 STATISTISCHE ÄMTER DES BUNDES UND DER LÄNDER FORSCHUNGSDATENZENTREN	 Qualiservice data sharing	 Elections	 ZEW FDZ ResearchDataCentre
 Economics & Business Data Center	 Forschungsdatenzentrum Betriebs- und OrganisationsDATEN	 DZA German Centre of Gerontology	 forschungsdatenzentrum bildung	 Lifbi LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES	 ROBERT KOCH INSTITUT	 FORSCHUNGS DATENZENTRUM	 Bundesamt für Migration und Flüchtlinge
 Deutsche Rentenversicherung Bund	 DZSTATIS FDZ	 fdz.DZHW German Research Data Centre for Higher Education Research and Science Studies	 IQI Institut zur Qualitätsentwicklung im Bildungswesen Forschungsdatenzentrum	 pairfam	 fdz ruhr BFI IWH	 Leibniz Institute of Ecological Urban and Regional Development	 aviDa AVIATION AND INFRASTRUCTURE RESEARCH DATA CENTER

Erwartungen

Bereitgestellte Datenformate von FDZs des KonsortSWD

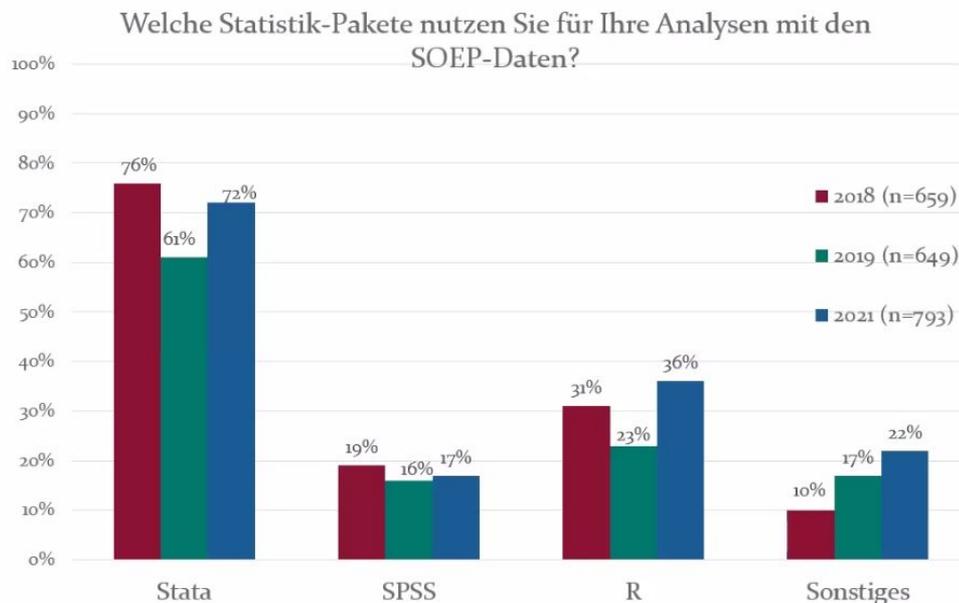


Quelle

Eigene Analyse der FDZ Webseiten,
Januar 2022

Erwartungen

Genutzte Datenformate – SOEP User Survey



Problem Statement

Datenformate im Kontext der FAIR Prinzipien

Findability

Accessibility

Interoperability

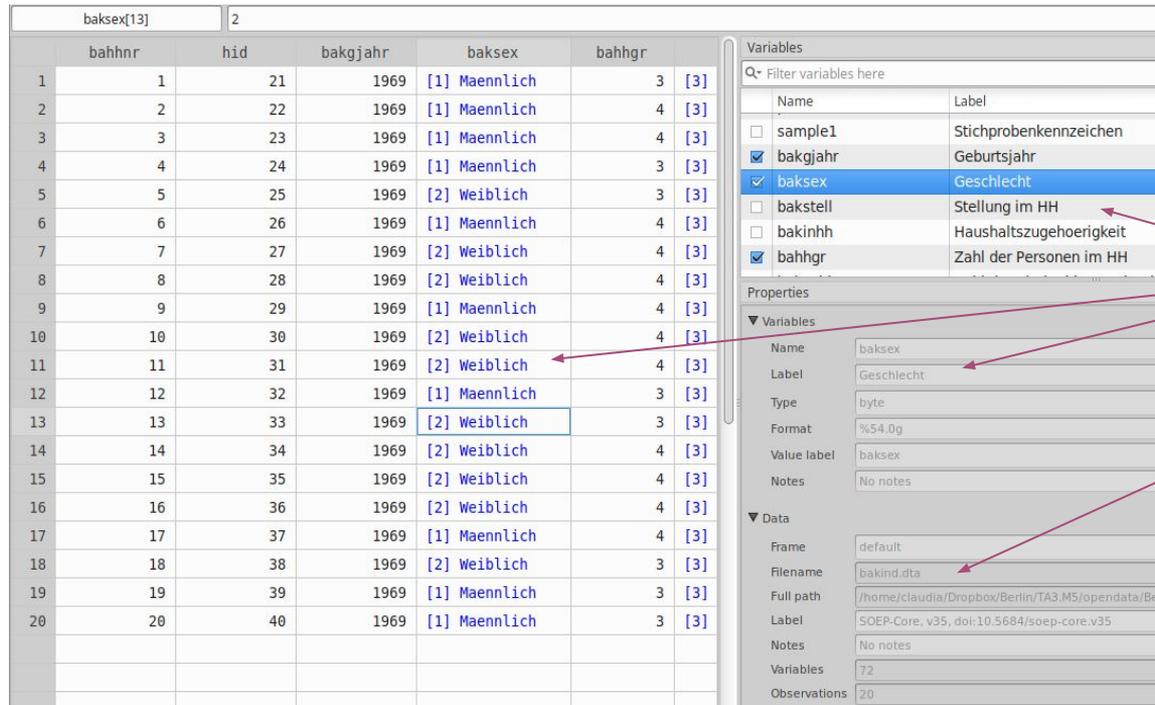
- (Proprietäre) Datenformate sind nicht interoperabel zwischen verschiedener Software oder Software Versionen
 - FDZs produzieren verschiedene Datenformate > kostenintensiv
- Software und Services zur Konvertierung von Datenformaten existieren
 - konzeptionell inkompatible Spezifikationen
 - schwer nachvollziehbare Prozesse
 - Risiko Informationsverlust

Reproducibility

- Nicht-Interoperabilität erschwert Reproduzierbarkeit
- Ein verbesserter Zugang zu Metadaten ist die Grundlage für Reproduzierbarkeit
- Nutzer müssen ihre gewohnte Softwareumgebung verlassen, um auf Metadaten wie Fragebögen, methodische oder technische Berichte zuzugreifen

Erwartungen

Stata Data Editor



The screenshot displays the Stata Data Editor interface. The main window shows a dataset with 20 observations and 6 variables: bahnhr, hid, bakgjahr, baksex, and bahngr. The 'Variables' panel on the right provides metadata for the selected variables. The 'Properties' panel shows details for the 'baksex' variable.

Observation	bahnhr	hid	bakgjahr	baksex	bahngr
1	1	21	1969	[1] Maennlich	3 [3]
2	2	22	1969	[1] Maennlich	4 [3]
3	3	23	1969	[1] Maennlich	4 [3]
4	4	24	1969	[1] Maennlich	3 [3]
5	5	25	1969	[2] Weiblich	3 [3]
6	6	26	1969	[1] Maennlich	4 [3]
7	7	27	1969	[2] Weiblich	4 [3]
8	8	28	1969	[2] Weiblich	4 [3]
9	9	29	1969	[1] Maennlich	4 [3]
10	10	30	1969	[2] Weiblich	4 [3]
11	11	31	1969	[2] Weiblich	4 [3]
12	12	32	1969	[1] Maennlich	3 [3]
13	13	33	1969	[2] Weiblich	3 [3]
14	14	34	1969	[2] Weiblich	4 [3]
15	15	35	1969	[2] Weiblich	4 [3]
16	16	36	1969	[2] Weiblich	4 [3]
17	17	37	1969	[1] Maennlich	4 [3]
18	18	38	1969	[2] Weiblich	3 [3]
19	19	39	1969	[1] Maennlich	3 [3]
20	20	40	1969	[1] Maennlich	3 [3]

Variables Panel:

Name	Label
<input type="checkbox"/> sample1	Stichprobenkennzeichen
<input checked="" type="checkbox"/> bakgjahr	Geburtsjahr
<input checked="" type="checkbox"/> baksex	Geschlecht
<input type="checkbox"/> bakstell	Stellung im HH
<input type="checkbox"/> bakinhh	Haushaltszugehoerigkeit
<input checked="" type="checkbox"/> bahngr	Zahl der Personen im HH

Properties Panel (for baksex):

Property	Value
Name	baksex
Label	Geschlecht
Type	byte
Format	%54.0g
Value label	baksex
Notes	No notes

Variablen Metadaten

Datensatz Metadaten

Spezifikation

Metadaten Elemente und Attribute

Datensatz Metadaten

- Name
- Label (mehrsprachig)
- Beschreibung (mehrsprachig)
- Url

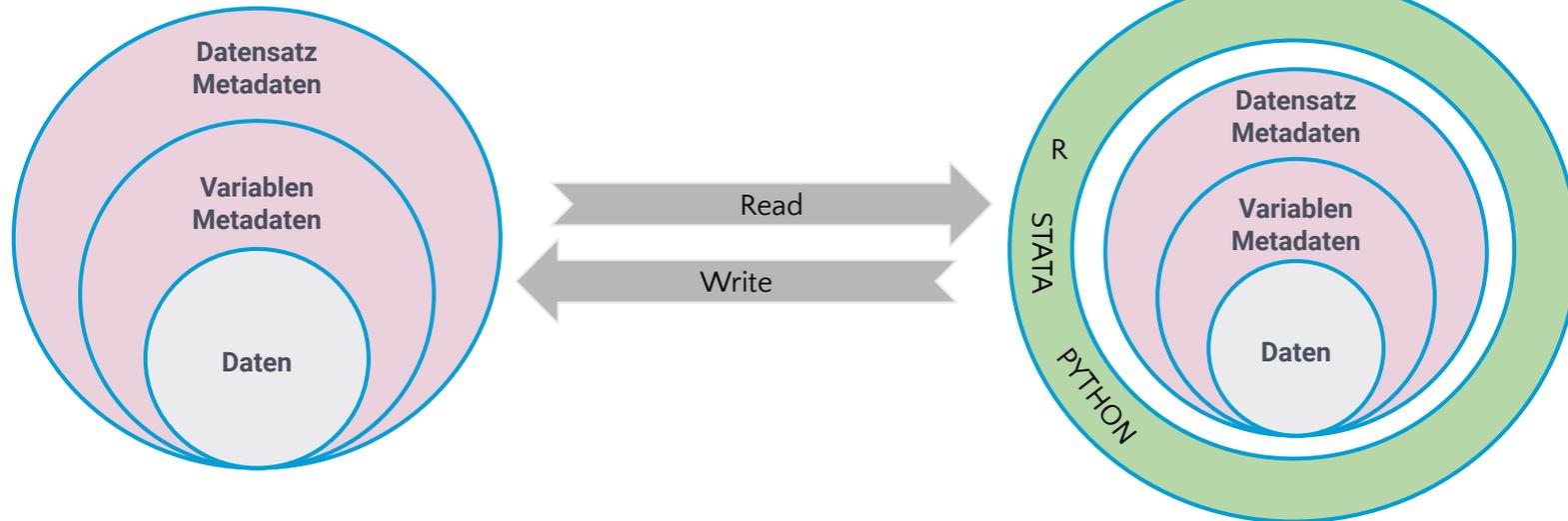
Variablen Metadaten

- Name
 - Label (mehrsprachig)
 - Beschreibung (mehrsprachig)
 - Url
 - Typ (numerisch, character)
 - Kategorien
 - Werte
 - Werte Labels
- } inkl. benutzerdefinierter Missings

Ziel

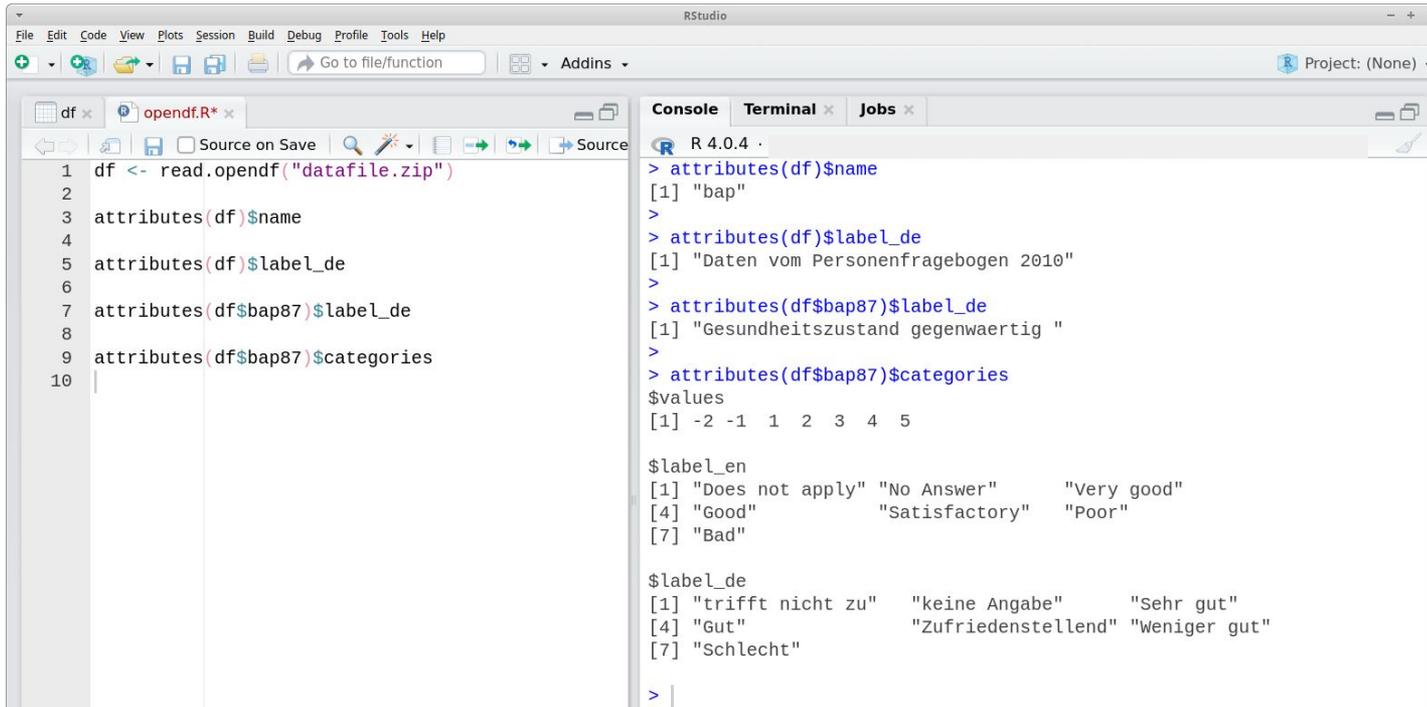
Metadata enriched
Open Data Format

Statistical
Software



Features

Zugang zu Metadaten



```
df <- read.opendf("datafile.zip")
attributes(df)$name
attributes(df)$label_de
attributes(df$bap87)$label_de
attributes(df$bap87)$categories
```

```
R 4.0.4 .
> attributes(df)$name
[1] "bap"
>
> attributes(df)$label_de
[1] "Daten vom Personenfragebogen 2010"
>
> attributes(df$bap87)$label_de
[1] "Gesundheitszustand gegenwaertig "
>
> attributes(df$bap87)$categories
$values
[1] -2 -1  1  2  3  4  5

$label_en
[1] "Does not apply" "No Answer"      "Very good"
[4] "Good"           "Satisfactory"   "Poor"
[7] "Bad"

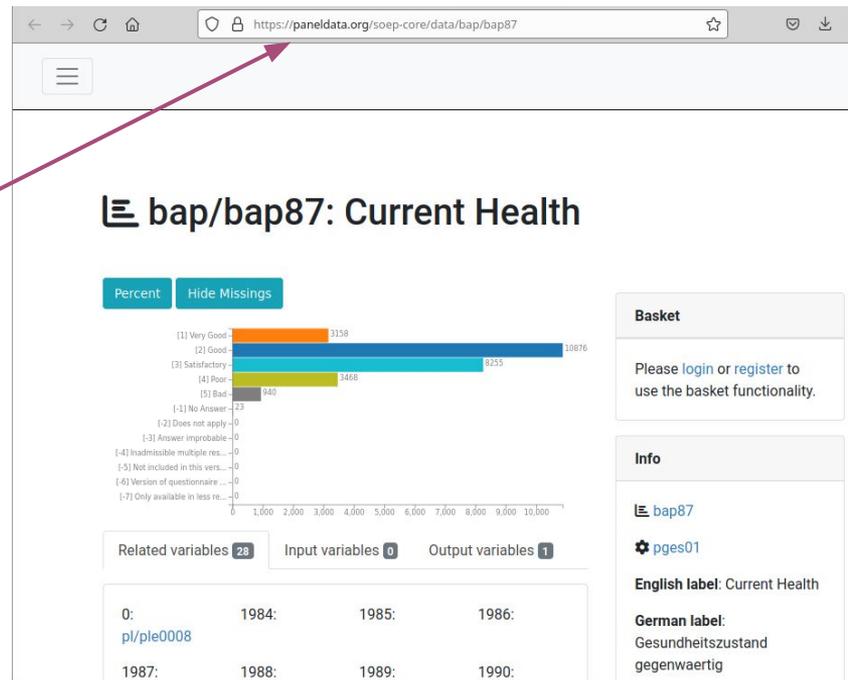
$label_de
[1] "trifft nicht zu" "keine Angabe"   "Sehr gut"
[4] "Gut"             "Zufriedenstellend" "Weniger gut"
[7] "Schlecht"

> |
```

Features

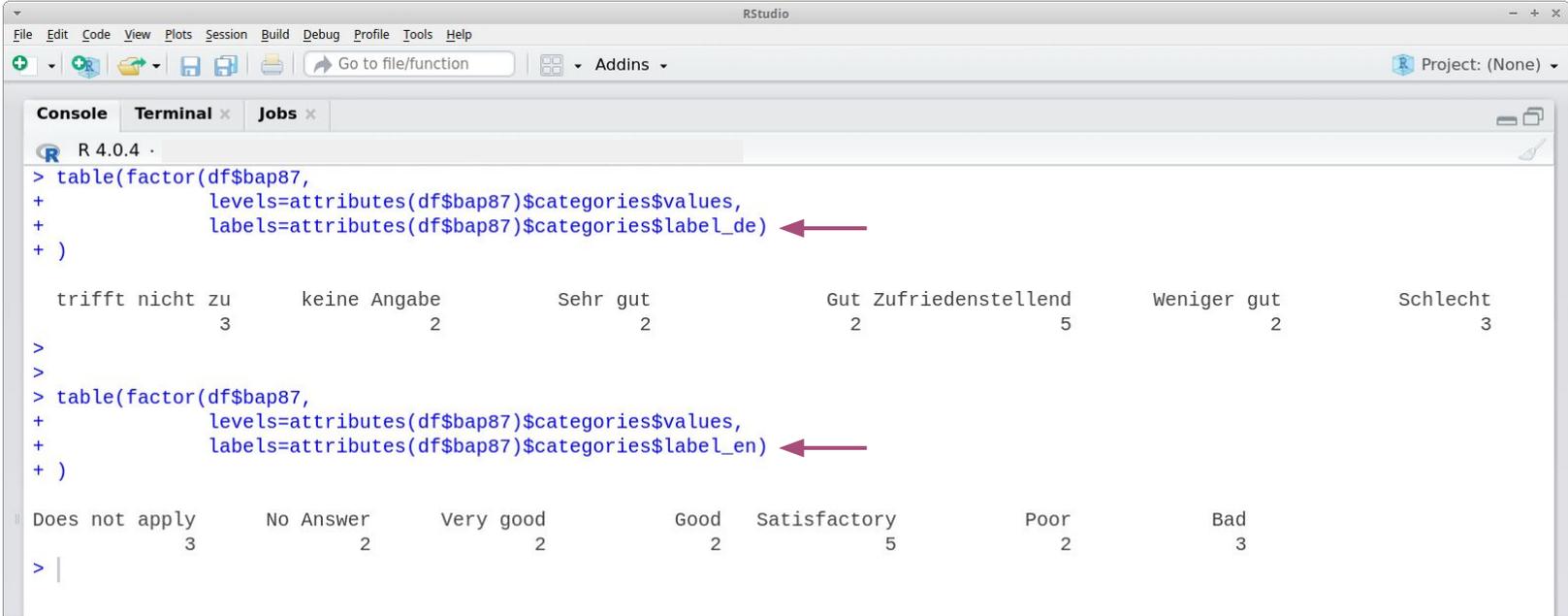
Zugang zu Metadaten

```
RStudio  
Project:  
Console Terminal Jobs  
R 4.0.4  
> attributes(df$bap87)$url  
[1] "https://paneldata.org/soep-core/data/bap/bap87"  
> browseURL(attributes(df$bap87)$url)  
> |
```



Features

Mehrsprachige Labels

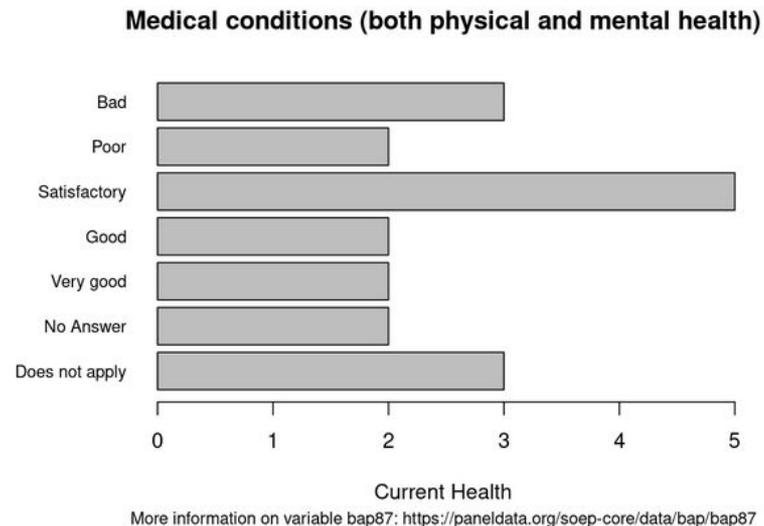


```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins Project: (None)
Console Terminal Jobs
R 4.0.4
> table(factor(df$bap87,
+           levels=attributes(df$bap87)$categories$values,
+           labels=attributes(df$bap87)$categories$label_de)
+ )
  trifft nicht zu   keine Angabe   Sehr gut   Gut Zufriedenstellend   Weniger gut   Schlecht
                3             2             2             2             5             2             3
>
>
> table(factor(df$bap87,
+           levels=attributes(df$bap87)$categories$values,
+           labels=attributes(df$bap87)$categories$label_en)
+ )
Does not apply   No Answer   Very good   Good   Satisfactory   Poor   Bad
                3             2             2             2             5             2             3
> |
```

Features

Flexible Nutzung der Metadaten

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
df x opendf.R* x
Source on Save Run Source
1 df <- read.opendf("datafile.zip")
2
3 bp <- barplot(
4   table(
5     factor(
6       df$bap87,
7       levels=attributes(df$bap87)$categories$value, # values
8       labels=attributes(df$bap87)$categories$label_en # value labels
9     )
10  ),
11  horiz=TRUE, las=1, cex.names = .8,
12  xlab=attributes(df$bap87)$label_en, # variable label
13  main=attributes(df$bap87)$description_en, # variable description
14  sub=paste0(
15    "More information on the variable ",
16    attributes(df$bap87)$name, # variable name
17    ": ",
18    attributes(df$bap87)$url, # variable url
19    cex.sub=.8
20  )
21 )
22
23
```



Spezifikation

Format

Metadata

xml

```
<dataDescr>
  <var name="bap87">
    <labl xml:lang="en">Current Health</labl>
    <labl xml:lang="de">Gegenwärtiger Gesundheitszustand</labl>
    <txt xml:lang="en">Medical conditions (both physical and mental health)
    <txt xml:lang="de">Gesundheitszustand (sowohl körperliche als auch geistige)
    <notes>
      <ExtLink URI="https://paneldata.org/soep-core/data/bap/bap87"/>
    </notes>
    <varFormat type="numeric"/>
    <catgry>
      <catValu>-2</catValu>
      <labl xml:lang="en">Does not apply</labl>
      <labl xml:lang="de">trifft nicht zu</labl>
    </catgry>
    <catgry>
      <catValu>-1</catValu>
      <labl xml:lang="en">No Answer</labl>
      <labl xml:lang="de">keine Angabe</labl>
    </catgry>
    <catgry>
      <catValu>1</catValu>
      <labl xml:lang="en">Very good</labl>
      <labl xml:lang="de">Sehr gut</labl>
    </catgry>
    <catgry>
      <catValu>2</catValu>
      <labl xml:lang="en">Good</labl>
      <labl xml:lang="de">Gut</labl>
    </catgry>
  </var>
</dataDescr>
```

Data

csv

```
bap87, bap9201, bap9001, bap9002, bap9003, bap96, name
4, -2, 1, -1, 2, -2, Jakob
3, 5, -2, 1, 4, 1.57, Luca
, -1, -1, 2, -1, 1.92, Emilia
1, 9, -2, 2, 4, 1.85, Charlotte
-1, 4, 2, 3, 1, 1.91, Johanna
3, 4, -1, 4, -2, 1.8, Paul
1, 9, 2, -1, -1, 1.8,
5, 6, 1, -1, 1, 1.96, Mia
5, 5, 5, 3, 1, 1.64, Ben
-2, 4, 4, -1, -2, 1.93, Jakob
-1, 4, 2, 1, 5, 1.93, Anton
-2, 5, 3, -2, 4, , Charlotte
3, -1, 2, 1, 2, 1.74, Luca
2, -2, -2, 4, -1, 1.65, Maria
5, -1, -2, -1, -1, 1.8, Johanna
4, 5, 1, 3, -1, 1.58, Emma
3, 7, 1, 2, -2, 1.95, Felix
3, , 5, 3, -2, 1.98, David
-2, 8, 1, 4, 5, 1.61, Emma
- - - - -
```

Spezifikation

Metadaten Profil, Validierung & Interoperabilität



DDI Codebook 2.5 OpenDF Profil

variable label	A short description of the variable. In the variable label, the length of this phrase may depend on the statistical analysis system used (e.g., some versions of SAS permit 40-character labels, while some versions of SPSS permit 120 characters), although the DDI itself imposes no restrictions on the number of characters allowed.	//codeBook/dataDscr/var/labl	mandatory if ,var' element is present	<pre> <pr:Used xpath="/codeBook/dataDscr/var/labl" isRequired="false"> <r:Description> <r:Content>Required: Mandatory if 'var' element is present.</r:Content> <r:Content>ElementType: Content element</r:Content> <r:Content>Usage: A short description of the parent element. In t phrase may depend on the statistical analysis system used (e.g., labels, while some versions of SPSS permit 120 characters), altho on the number of characters allowed.</r:Content> </r:Description> <pr:Instructions> <r:Content> <![CDATA[<Constraints><MandatoryNodeIfParentPresentConstr </r:Content>] </pr:Instructions> </pr:Used> </pre>
language tag	Attribute to specify the language of the <variable label>. <Use ISO-639-1-Code for language subtags, e.g. en for English.>	//codeBook/dataDscr/var/labl[@xml:lang]	mandatory if ,labl' element is present	<pre> <pr:Used xpath="/codeBook/dataDscr/var/labl/@xml:lang" isRequired="false"> <r:Description> <r:Content>Required: Mandatory if 'labl' element is present.</r:Content> <r:Content>ElementType: Attribute element</r:Content> <r:Content>Usage: Attribute to specify the language of the variat </r:Description> <pr:Instructions> <r:Content> <![CDATA[<Constraints><MandatoryNodeIfParentPresentConstr </r:Content>] </pr:Instructions> </pr:Used> </pre>

Spezifikation

Veröffentlichung

- Repository
 - Fiktives Datenbeispiel im Offenen Datenformat
 - Profil und Beschreibung
- Arbeitsbericht

Herzlichen Dank!

