KonsortSWD

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

Open Data Format

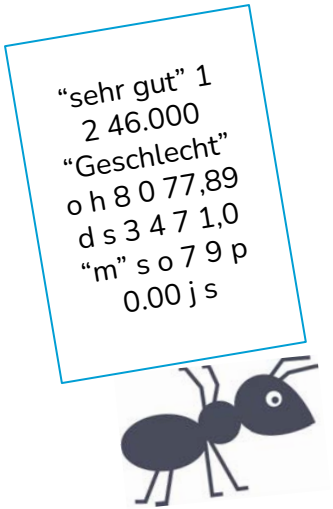
TA.3-M.5

KonsortSWD Community Conference, 07.04.2022

Xiaoyao Han,
Claudia Saalbach &
Knut Wenzig
DIW, Berlin

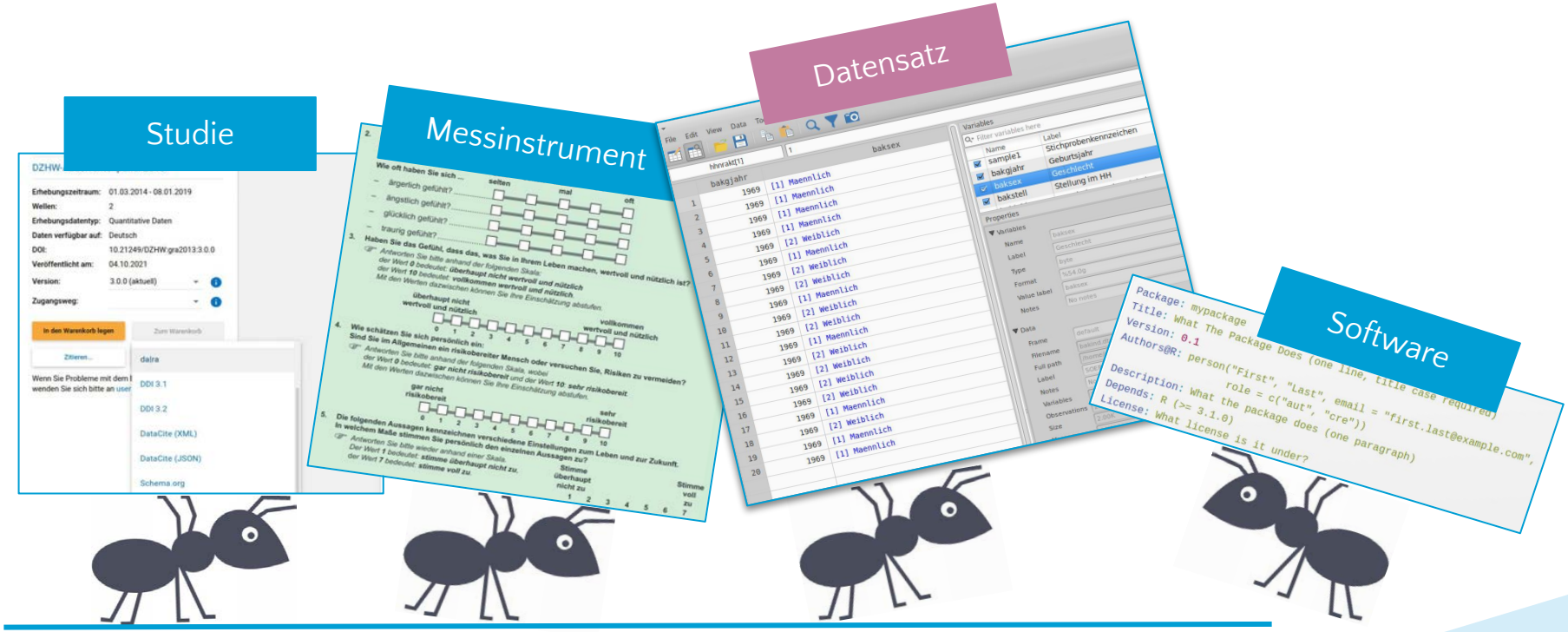
DIW SOEP

Inhalt



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

Metadaten

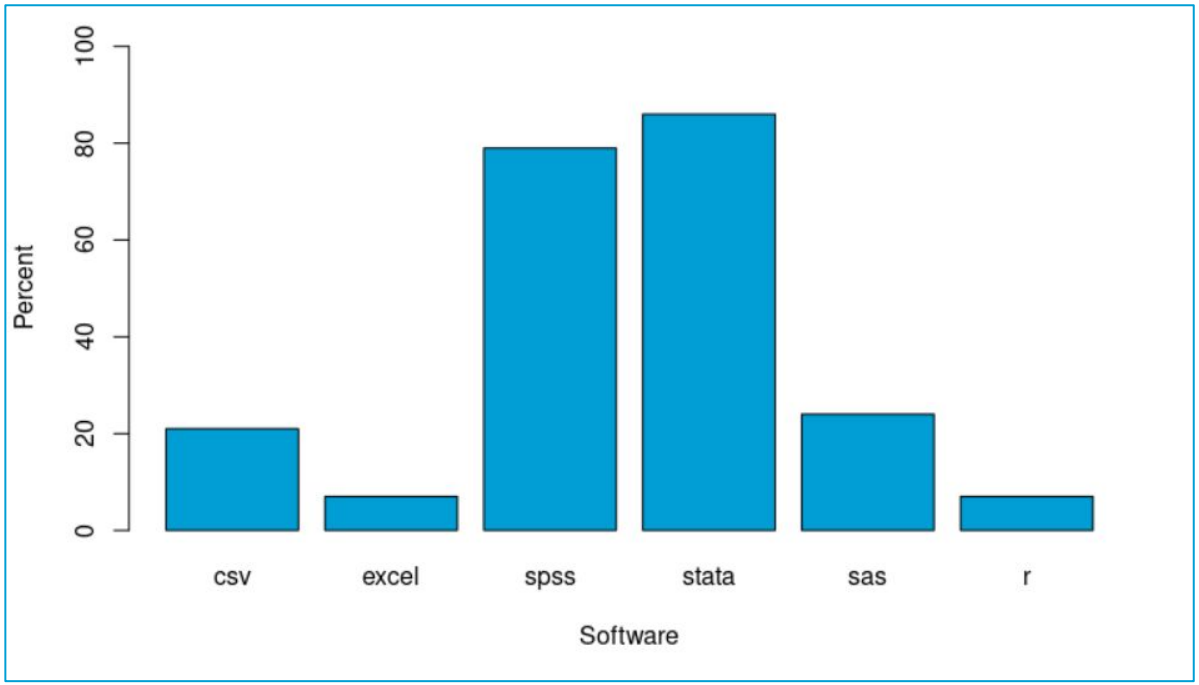

Studie
Messinstrument
Datensatz
Software
Erhebungszeitraum: 01.03.2014 - 08.01.2019
Wellen: 2
Erhebungsdatentyp: Quantitative Daten
Daten verfügbar auf: Deutsch
DOI: 10.21249/DZHW:gra2013:3.0.0
Veröffentlicht am: 04.10.2021
Version: 3.0.0 (aktuell)
Zugangsweg:
In den Warenkorb legen
Zitieren...
dalra
DDI 3.1
DDI 3.2
DataCite (XML)
DataCite (JSON)
Schema.org
Wie oft haben Sie sich ...
ärgerlich gefühlt?
ängstlich gefühlt?
glücklich gefühlt?
traurig gefühlt?
selten
mal
oft
Haben Sie das Gefühl, dass das, was Sie in Ihrem Leben machen, wertvoll und nützlich ist?
überhaupt nicht wertvoll und nützlich
vollkommen wertvoll und nützlich
Wie schätzen Sie sich persönlich ein:
gar nicht risikobereit
sehr risikobereit
Die folgenden Aussagen kennzeichnen verschiedene Einstellungen zum Leben und zur Zukunft. In welchem Maße stimmen Sie persönlich den einzelnen Aussagen zu?
Stimme überhaupt nicht zu
Stimme voll zu
Variables
Properties
Package: mypackage
Title: What The Package Does (one line, title case required)
Version: 0.1
Authors@R: person("First", "Last", email = "first.last@example.com", role = c("aut", "cre"))
Description: What the package does (one paragraph)
Depends: R (>= 3.1.0)
License: What license is it under?

Erwartungen

Datenquellen- KonsortSWD Forschungsdatenzentren

Erwartungen

Bereitgestellte Datenformate von FDZs des KonsortSWD

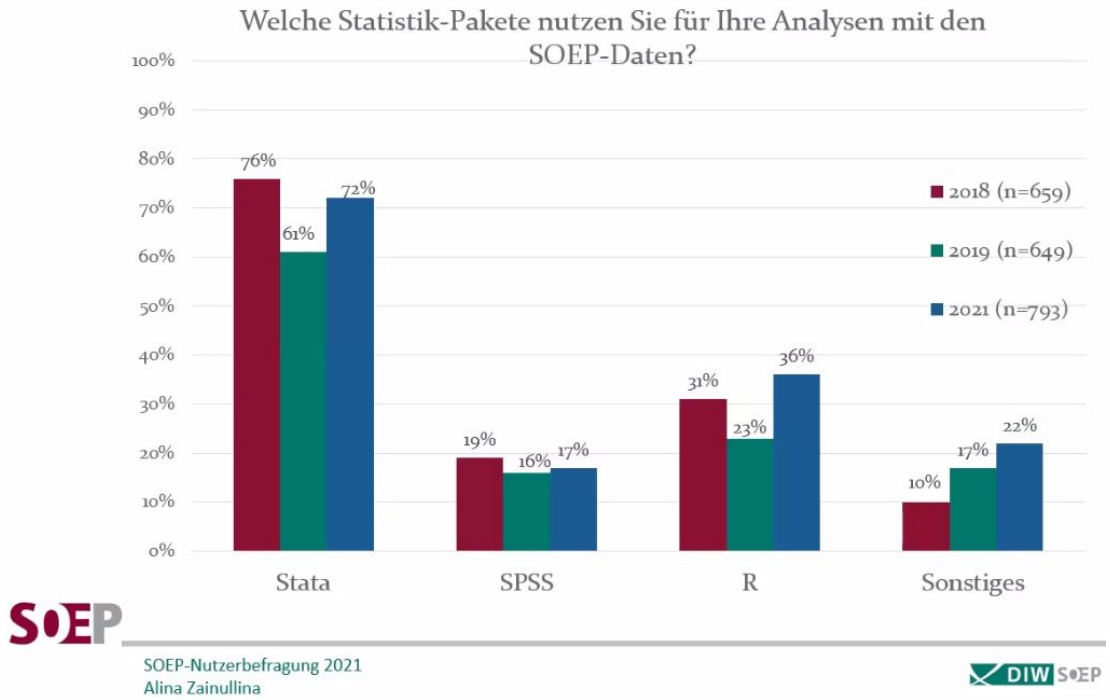


Quelle

Eigene Analyse der FDZ Webseiten, Januar 2022

Erwartungen

Genutzte Datenformate – SOEP User Survey



Problem Statement

Datenformate im Kontext der FAIR Prinzipien

Findability

Accessibility

Interoperability

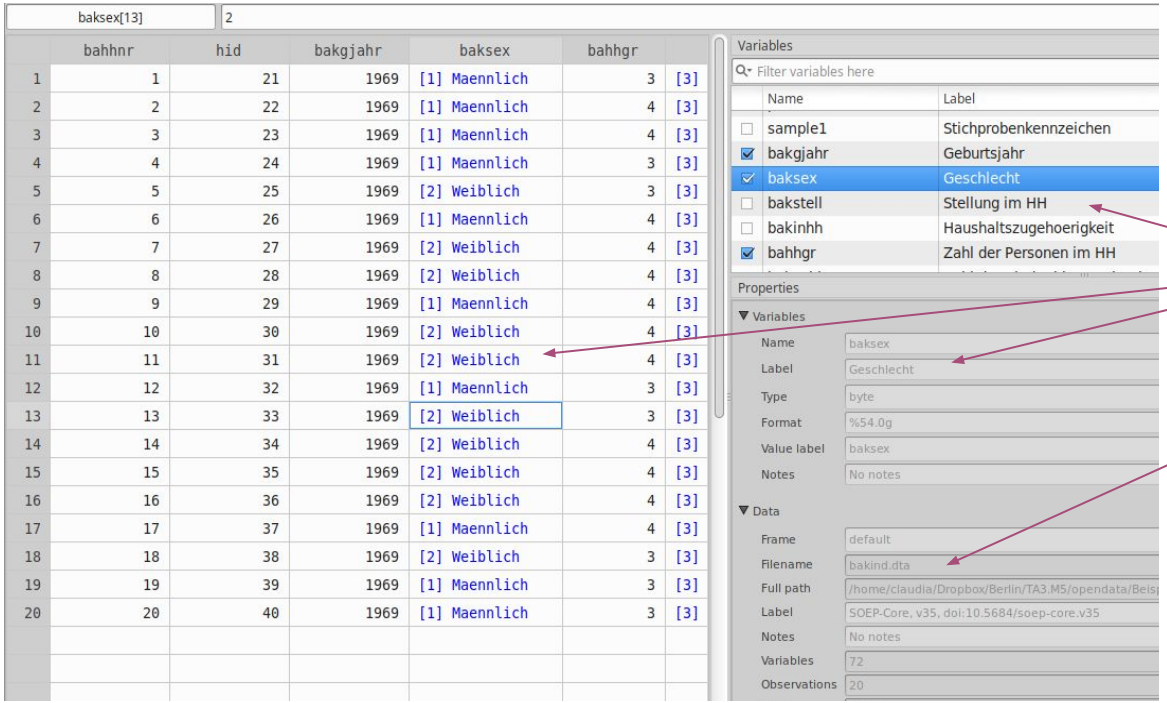
- (Proprietäre) Datenformate sind nicht interoperabel zwischen verschiedener Software oder Software Versionen
  - FDZs produzieren verschiedene Datenformate > kostenintensiv
- Software und Services zur Konvertierung von Datenformaten existieren
  - konzeptionell inkompatible Spezifikationen
  - schwer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



Variablen Metadaten

Datensatz Metadaten

Spezifikation

Metadaten Elemente und Attribute

Datensatz Metadaten

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

Variablen Metadaten

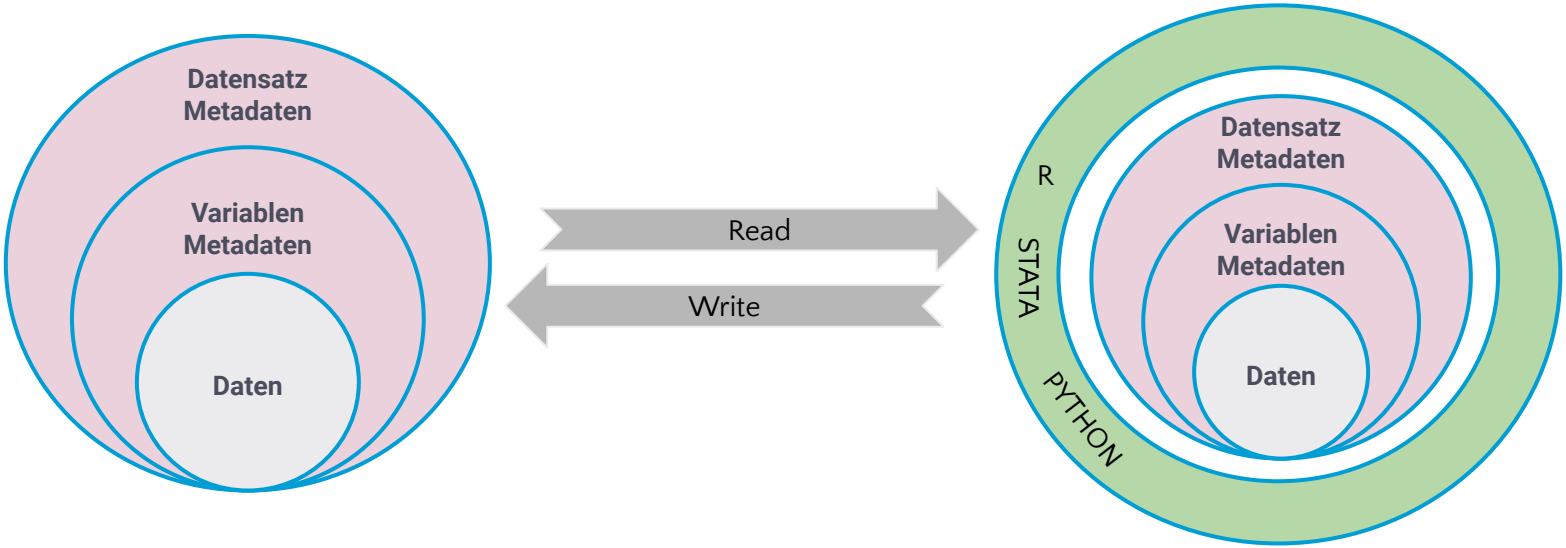
- Name
- Label (mehrsprachig)
- Beschreibung (mehrsprachig)
- Url
- Typ (numerisch, character)
- Kategorien
  - Werte
  - Werte Labels

} inkl. benutzerdefinierter Missings

Ziel

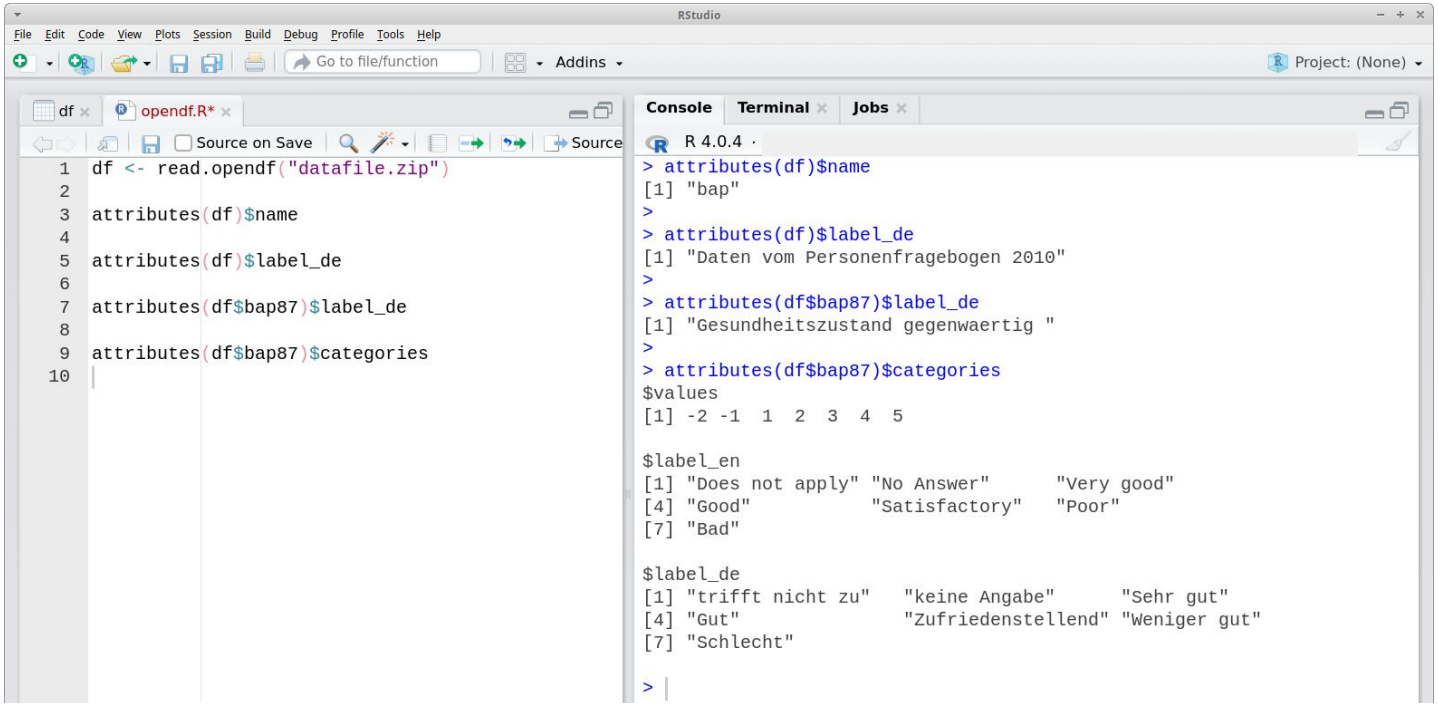
Metadata enriched Open Data Format

Statistical Software



Features

Zugang zu Metadaten



Features

Zugang zu Metadaten

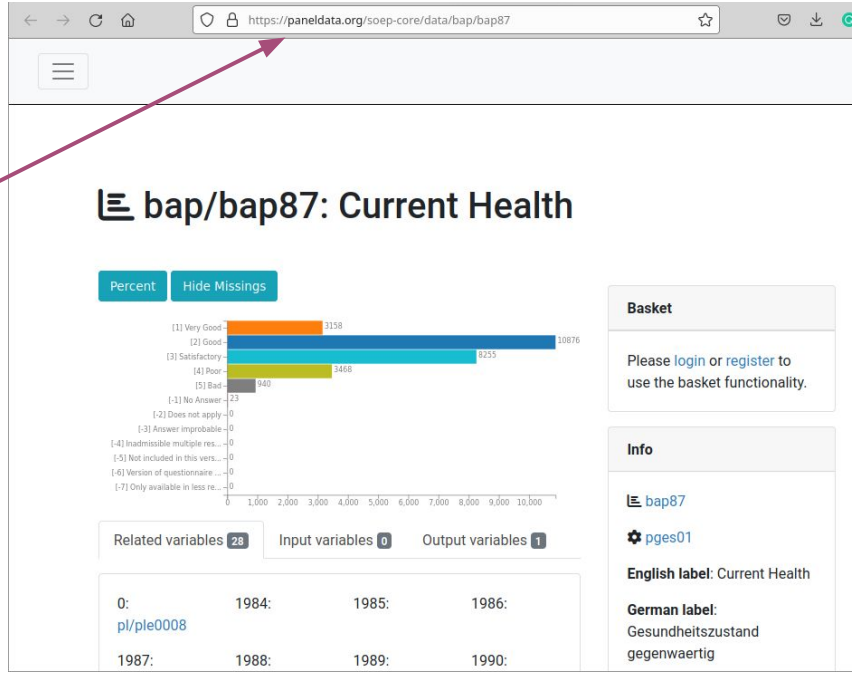
RStudio

Project:

Console | Terminal | Jobs

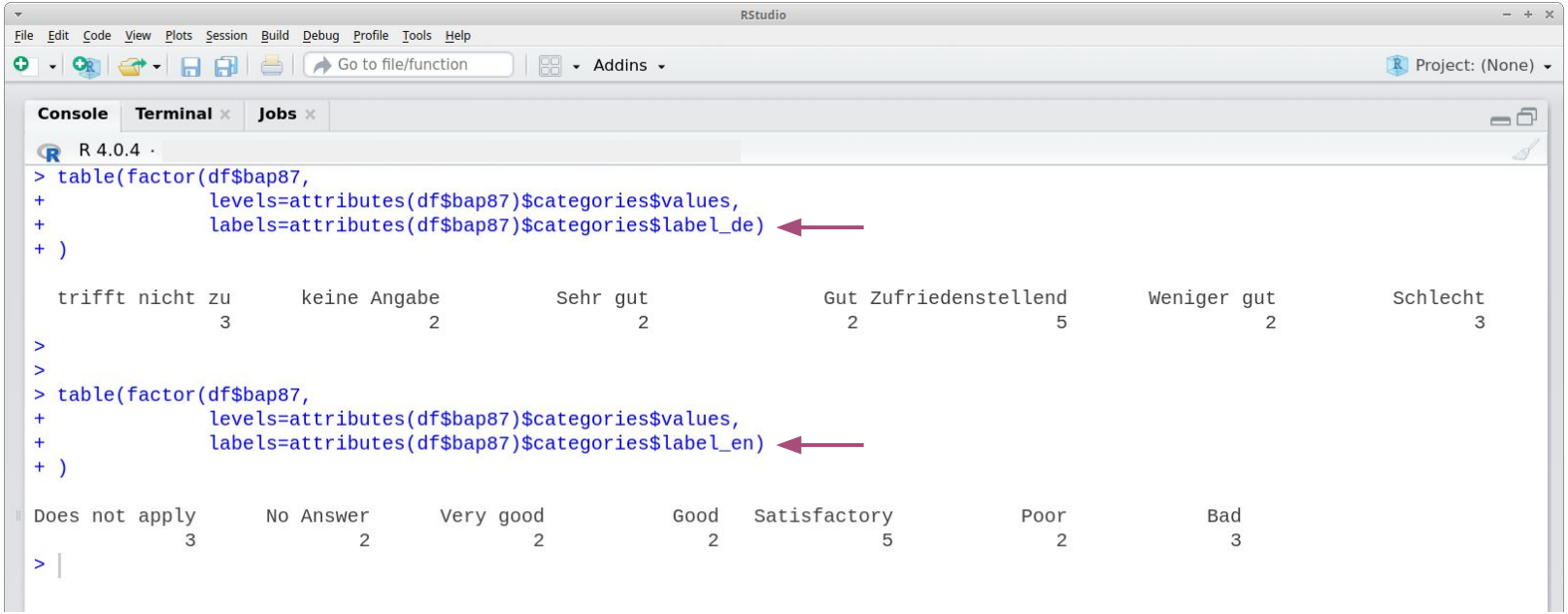
R 4.0.4 ·

```
> attributes(df$bap87)$url
[1] "https://paneldata.org/soep-core/data/bap/bap87"
>
> browseURL(attributes(df$bap87)$url)
>
```



Features

Mehrsprachige Labels

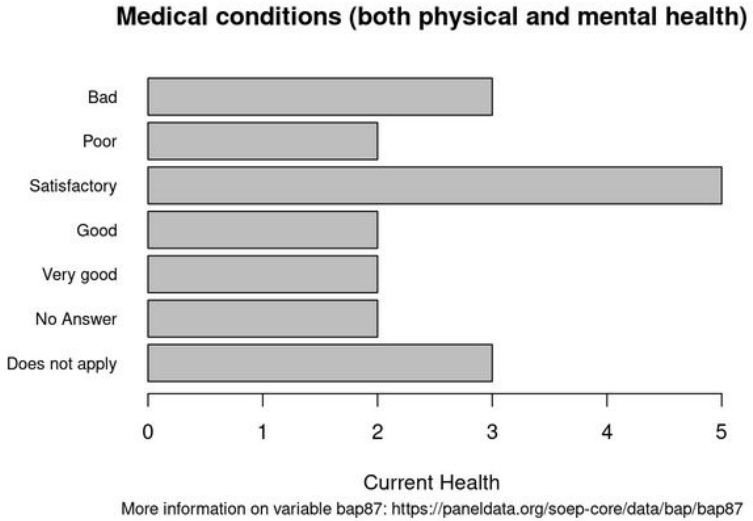


Features

Flexible Nutzung der Metadaten

```
df <- read.opendf("datafile.zip")

bp <- barplot(
    table(
        factor(
            df$bap87,
            levels=attributes(df$bap87)$categories$value,# values
            labels=attributes(df$bap87)$categories$label_en # value labels
        )
    ),
    horiz=TRUE, las=1, cex.names = .8,
    xlab=attributes(df$bap87)$label_en, # variable label
    main=attributes(df$bap87)$description_en, # variable description
    sub=paste0(
        "More information on the variable ",
        attributes(df$bap87)$name, # variable name
        ": ",
        attributes(df$bap87)$url, # variable url
        cex.sub=.8
    )
)
```



Spezifikation

Format

Metadata
xml

```
<dataDscr>
        <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 healt
            <txt xml:lang="de">Gesundheitszustand (sowohl körperliche als auch ge
            <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>
```

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
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

```
<pr:Used xpath="/codeBook/dataDscr/var/labl" isRequired="false">
        <r:Description>
                <r:Content>Required: Mandatory if 'var' element is present.</r:Co
                <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>

<pr:Used xpath="/codeBook/dataDscr/var/labl/@xml:lang" isRequired="false">
        <r:Description>
                <r:Content>Required: Mandatory if 'labl' element is present.</r:C
                <r:Content>ElementType: Attribute element</r:Content>
                <r:Content>Usage: Attribute to specify the language of the variab
        </r:Description>
        <pr:Instructions>
                <r:Content>
                        <![CDATA[<Constraints><MandatoryNodeIfParentPresentConstr
                </r:Content>
        </pr:Instructions>
</pr:Used>
```

Spezifikation

Veröffentlichung

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

Herzlichen Dank!