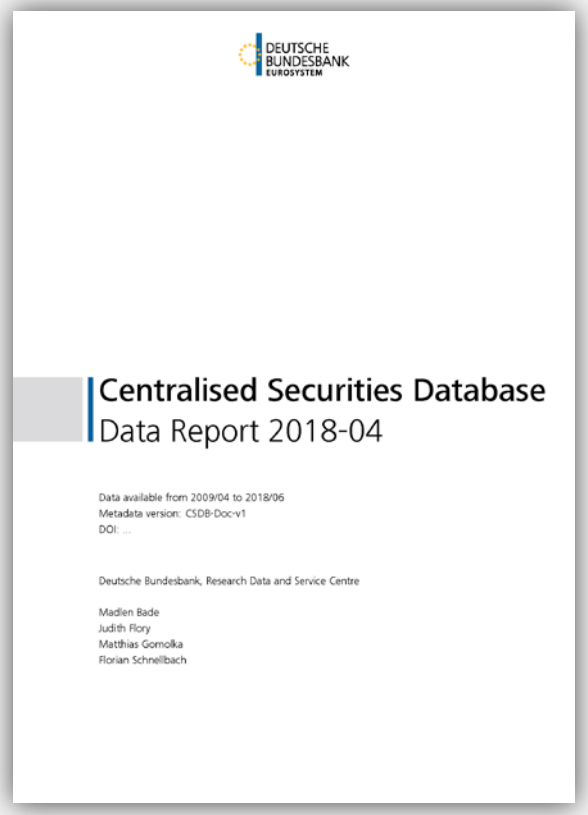
DataReportR

An Internal Package for (Semi-)Automated Metadata Documentation

Matthias Gomolka
Research Data and Service Centre
@El_Grund

What is a Data Report?

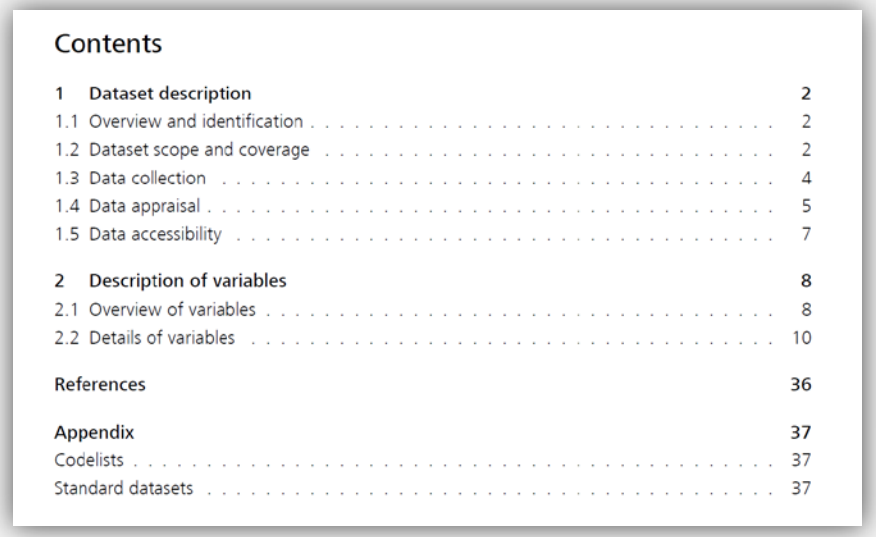
- The RDSC produces research datasets
- Each dataset should be accompanied by a Data Report
- A Data Report is a collection of metadata



What is a Data Report technically?

A LaTeX document

- which is standardized
- and contains many tables



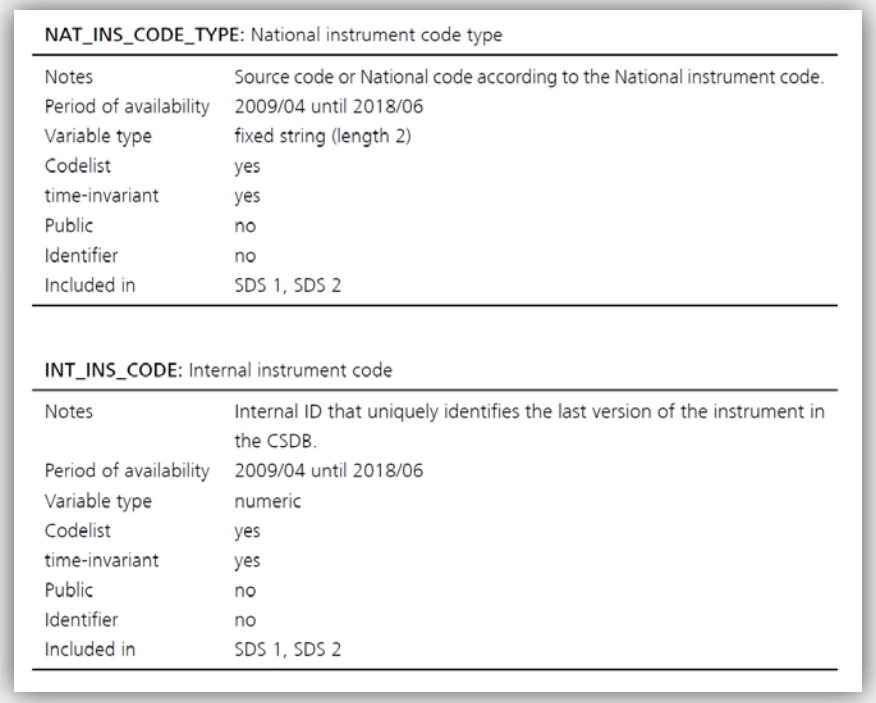
Contents

1 Dataset description 2
1.1 Overview and identification . . . 2
1.2 Dataset scope and coverage . . . 2
1.3 Data collection . . . 4
1.4 Data appraisal . . . 5
1.5 Data accessibility . . . 7

2 Description of variables 8
2.1 Overview of variables . . . 8
2.2 Details of variables . . . 10

References 36

Appendix 37
Codelists . . . 37
Standard datasets . . . 37



NAT_INS_CODE_TYPE: National instrument code type

Notes	Source code or National code according to the National instrument code.
Period of availability	2009/04 until 2018/06
Variable type	fixed string (length 2)
Codelist	yes
time-invariant	yes
Public	no
Identifier	no
Included in	SDS 1, SDS 2

INT_INS_CODE: Internal instrument code

Notes	Internal ID that uniquely identifies the last version of the instrument in the CSDB.
Period of availability	2009/04 until 2018/06
Variable type	numeric
Codelist	yes
time-invariant	yes
Public	no
Identifier	no
Included in	SDS 1, SDS 2

Why create an R package?

Two simple questions:

1. Who likes copy / paste?
2. Who enjoys writing LaTeX tables by hand?

I'm lazy, so I don't.

```
\subsection{Overview of variables}\label{sec:variableoverview}

\begin{tabularx}{\textwidth}{lX}

  \textbf{Name} & \textbf{Label}\\
  \midrule
  \endfirsthead
  \multicolumn{2}{r@{}}{\textcolor{gray}{\itshape \footnotesize Overview of variables continued from previous page}}\\
  \textbf{Name} & \textbf{Label}\\
  \midrule
  \endhead
  \multicolumn{2}{r@{}}{\textcolor{gray}{\itshape \footnotesize Overview of variables continues on next page}}
  \endfoot
  \bottomrule
  \endlastfoot

  \\
    \multicolumn{2}{l}{\bfseries General Attributes} \\
    {\hyperref[var:EXTRACTION_DT]{EXTRACTION\_DT}} & Extraction date \\
    {\hyperref[var:DT_LAST_MODIFIED]{DT\_LAST\_MODIFIED}} & Date last modified \\
    \\
    \multicolumn{2}{l}{\bfseries Instrument identification} \\
    {\hyperref[var:ISIN]{ISIN}} & International Securities Identification Number (ISIN) \\
    {\hyperref[var:NAT_INS_CODE]{NAT\_INS\_CODE}} & National instrument code \\
    {\hyperref[var:NAT_INS_CODE_TYPE]{NAT\_INS\_CODE\_TYPE}} & National instrument code type \\
    {\hyperref[var:INT_INS_CODE]{INT\_INS\_CODE}} & Internal instrument code \\
    \\
    \multicolumn{2}{l}{\bfseries Instrument attributes} \\
    {\hyperref[var:SEC_STATUS]{SEC\_STATUS}} & Security status \\
    \\
    \multicolumn{2}{l}{\bfseries Instrument attributes} \\
    {\hyperref[var:SEC_STATUS_DT]{SEC\_STATUS\_DT}} & Security status date \\
    {\hyperref[var:SHORT_NAME]{SHORT\_NAME}} & Short name \\
    {\hyperref[var:QUOTATION_BASIS]{QUOTATION\_BASIS}} & Quotation basis \\
    {\hyperref[var:NOMINAL_CURRENCY]{NOMINAL\_CURRENCY}} & Nominal currency \\
    {\hyperref[var:AMOUNT_ISSUED]{AMOUNT\_ISSUED}} & Amount issued \\
    {\hyperref[var:AMOUNT_OUTSTANDING]{AMOUNT\_OUTSTANDING}} & Amount outstanding \\
    {\hyperref[var:AMOUNT_OUTST_EUR]{AMOUNT\_OUTST\_EUR}} & Amount outstanding in EUR \\
    {\hyperref[var:NUMBER_OUTST]{NUMBER\_OUTST}} & Number outstanding \\
    {\hyperref[var:NOMINAL_VALUE]{NOMINAL\_VALUE}} & Nominal value \\
    {\hyperref[var:MARKET_CAPITAL]{MARKET\_CAPITAL}} & Market capitalisation \\
    {\hyperref[var:MARKET_CAP_EUR]{MARKET\_CAP\_EUR}} & Market Capitalisation in EUR \\
    {\hyperref[var:POOL_FACTOR]{POOL\_FACTOR}} & Pool factor \\
```

What can DataReportR do for me?

- `create_skeleton()`

- `create_var_overview()`
- `create_var_details()`
- `create_codelists()`

Advantages of having DataReportR

- reduced risk of wrong or outdated information
- no more manual LaTeX table writing
- Data Reports stick closer to Corporate Design

better result in less time

How does it work?

`create_skeleton(path, ...)`:

```
> dir(path, recursive = TRUE)
 [1] "appendices/codelists.tex"          "appendices/int_orgs.tex"
 [3] "data_report.tex"                   "graphics/bbklogo.pdf"
 [5] "sections/0_1_titlepage.tex"        "sections/0_2_abstract.tex"
 [7] "sections/1_1_dataoverview.tex"     "sections/1_2_datascope.tex"
 [9] "sections/1_3_datacollection.tex"   "sections/1_4_listofaggregates.tex"
[11] "sections/1_5_dataappraisal.tex"    "sections/1_6_dataaccessibility.tex"
[13] "sections/2_1_variableoverview.tex" "sections/2_2_vardetails.tex"
[15] "sections/3_definitions.tex"
```

- copies files from **`inst/extdata`** to **`path`**
- inserts title and other information into **`tex`** files using regular expressions and the **`stringr`** 📦

How does it work?

`create_[tables]():`

- insert information into Excel file
- some data wrangling with the **`tidyverse`**
- create LaTeX files using **`xtable`** and **`purrr`**

How does it work?

	A	B	C	D	E	F	H
1	position	variable	short_description	Notes	available_from	available_until	Variable type
2	1	REFERENCE_MONTH	Data is reported for last day of this month	Date in the format "YYYYMM"	2005/12	2018/06	numeric
3	2	BAID	Random ID for the reporting agent	Random ID for the respective reporting agent. By using this ID, information about the respective reporting agent can be linked to the SHS-Base plus (e.g. information from ZENTK, BISTA or NUTS 3-Codes etc.).	2005/12	2018/06	numeric
4	3	ISIN	International Securities Identification Number	The ISIN code is a 12-character alpha-numerical code which uniquely identifies a security. The structure of the ISIN is defined in ISO 6166. By using the ISIN, information about the respective securities can be linked to the SHS-Base plus (e.g. information on prices, dividends, stock splits etc.).	2005/12	2018/06	fixed string
5	4	HOLDER_SECTOR_ESA1995	Sector of the holder according to the ESA 1995	For the economic sector of the holder four-digit internal codes are used. The classification of the sector of the holder is in accordance with the European System of Accounts (ESA). For the purpose of the Securities Holdings Statistics a more detailed breakdown in some sectors is necessary. The additional breakdowns are sectors 1212 to 1299 in the list below.\newline \emph{CAUTION:} Until December 2012 the classification of the sector of the holder followed the ESA 95 scheme. Since January 2013 this classification follows the ESA 2010 scheme.	2005/12	2012/12	numeric
	5	HOLDER_SECTOR_ESA2010	Sector of the holder according to the ESA 2010	For the economic sector of the holder four-digit internal codes are used. The classification of the sector of the holder is in accordance with the European System of Accounts (ESA). For	2013/13	2018/06	numeric

How does it work?

```
# A tibble: 11 x 6
   variable   short_description  Notes       Source `Variable type` `Period of avai~
   <chr>      <chr>              <chr>       <lgl>  <chr>           <chr>
 1 REFERENCE~ Data is reported ~ Date in th~ NA     numeric         2005/12 until 2~
 2 BAID       Random ID for the~ Random ID ~ NA     numeric         2005/12 until 2~
 3 ISIN       International Sec~ The ISIN c~ NA     fixed string    2005/12 until 2~
 4 HOLDER_SE~ Sector of the hol~ "For the e~ NA     numeric         2005/12 until 2~
 5 HOLDER_SE~ Sector of the hol~ "For the e~ NA     numeric         2013/13 until 2~
 6 HOLDER_CO~ Country of the ho~ "Country c~ NA     fixed string    2005/12 until 2~
 7 STOCK_TYPE Stock type         Indicates ~ NA     fixed string    2005/12 until 2~
 8 NOMINAL_C~ Currency / Unit    Indicates ~ NA     fixed string    2005/12 until 2~
 9 STOCK_RAW  Stock raw          For securi~ NA     numeric         2005/12 until 2~
10 STOCK_NOM~ Stock nominal val~ "Nominal v~ NA     numeric         2005/12 until 2~
11 STOCK_MAR~ Stock market value "Stock val~ NA     numeric         2005/12 until 2~
```

How does it work?

```
# A tibble: 1 x 6
  variable   short_description     Notes    Source `Variable type` `Period of avail~
  <chr>      <chr>                 <chr>    <lgl>  <chr>           <chr>
1 REFERENCE~ Data is reported for~ Date in~ NA     numeric         2005/12 until 20~
```

How does it work?

```
# A tibble: 3 x 2
  attribute              value
  <chr>                  <chr>
1 Notes                  Date in the format “YYYYMM”
2 Period of availability 2005/12 until 2018/06
3 Variable type          numeric
```

How does it work?

```
\begin{tabularx}{\textwidth}{lX}
  \multicolumn{2}{p{\textwidth-2\tabcolsep}}{\textbf{REFERENCE\_MONTH:} Data is
  reported for last day of this month\label{var:REFERENCE_MONTH}}\\
  \midrule
  \endfirsthead
  \multicolumn{2}{r@{}}{\textcolor{gray}{\itshape \footnotesize Variable
  REFERENCE\_MONTH continued from previous page}}\\
  \endhead
  \multicolumn{2}{r@{}}{\textcolor{gray}{\itshape \footnotesize Variable
  REFERENCE\_MONTH continues on next page}}
  \endfoot
  \bottomrule
  \endlastfoot
Notes & Date in the format “YYYYMM” \\
  Period of availability & 2005/12 until 2018/06 \\
  Variable type & numeric \\
  \end{tabularx}
```

How does it work?

xtable

From xtable v1.8-3
by David Scott

99.5th Percentile

Create Export Tables

Convert an R object to an `xtable` object, which can then be printed as a LaTeX or HTML table.

Keywords file

Usage

```
xtable(x, caption = NULL, label = NULL, align = NULL, digits = NULL,
       display = NULL, auto = FALSE, ...)
```

How does it work?

```
var_details_df %>%
    xtable::xtable(align = c("l", "l", "X")) %>%
    print(tabular.environment    = "tabularx",
          booktabs               = TRUE,
          width                  = "\\textwidth",
          hline.after            = c(),
          include.colnames       = FALSE,
          include.rownames       = FALSE,
          floating               = FALSE,
          comment                = FALSE,
          add.to.row             = tableheader,
          sanitize.text.function = function(text){text}) %>%
    readr::write_file(tex_file, append = TRUE)
```

How does it work?

```
  tableheader <- list()
  tableheader$pos <- list(0)
  tableheader$command <- paste0(
    "\n  \\textbf{Name} & \\textbf{Label}\\\\\n",
    "  \\midrule\n",
    "  \\endfirsthead\n",
    "  \\multicolumn{2}{r@{}}{\\textcolor{gray}{\\itshape \\footnotesize
Overview of variables continued from previous page}}\\\\\n",
    "  \\textbf{Name} & \\textbf{Label}\\\\\n",
    "  \\midrule\n",
    "  \\endhead\n",
    "  \\multicolumn{2}{r@{}}{\\textcolor{gray}{\\itshape \\footnotesize
Overview of variables continues on next page}}\n",
    "  \\endfoot\n",
    "  \\bottomrule\n",
    "  \\endlastfoot\n\n"
  )
```

What's the future?

- make use of existing information for other sections
- make DataReport generation as reproducible as possible

```
library(DataReportR)

create_skeleton(
  path               = "../LaTeX",
  replace_sections   = FALSE,
  dataset            = "Securities Holdings Statistics",
  datasetsecondline  = "Base plus",
  dataset_abbr       = "SHS-Base plus",
  heightmarkenbalken = 4,
  datareport         = "2018-??",
  version            = "2-0",
  authors            = "Madlen Bade; Judith Flory; Matthias Gomolka; Tobias
Schönberg",
  authors_abbr       = "Bade, M., Flory, J., Gomolka, M., and T. Schönberg",
  timespanavailable  = "2005/12 to 2018/06",
  doi                = "...",
  keywords           = "Security holdings ESCB master data, security-by
-security database, debt securities, equity, investment funds, price data,
issuer information"
)

include_appendix("../LaTeX/data_report.tex", "codelists")
include_appendix("../LaTeX/data_report.tex", "int_orgs")

create_var_overview(
  var_details_xlsx = "var_details_shs_base_plus.xlsx",
  tex_file         = "../LaTeX/sections/2_1_variableoverview.tex"
)

create_var_details(
  var_details_xlsx = "var_details_shs_base_plus.xlsx",
  tex_file         = "../LaTeX/sections/2_2_vardetails.tex",
  allow_latex      = TRUE
)
```

Difficulties

- People are unfamiliar with R
- new custom LaTeX class was introduced just before DataReportR

How to overcome the difficulties?

- create great documentation (and make it accessible for non-useRs)
- help colleages gettings started with DataReportR
- be open for improvements

Why not R Markdown?

- Data Reports were created using LaTeX before I joined the RDSC
- Most of my colleagues are unfamiliar with R

DataReportR

See me afterwards …

… for more questions

… if you want to see / get the whole code

Matthias Gomolka
Research Data and Service Centre
@El_Grund