Constructing a data dictionary and appending it to your data in R

Mairead Bermingham

22 June, 2022

Overview

A data dictionary is an important tool in data management. A data dictionary is a supplementary document that details the information about variables, data collection, and other important features of a data set, i.e. metadata, data that describes other data. This metadata is crucial and helps others find, access, understand, and reuse your data. Without proper documentation, the data you store in online repositories may be rendered unfindable and unusable by others and indexing search engines. The *dataMeta* R package is designed to create a data dictionary and append it to the original dataset's attributes list along with other information generally provided as metadata. In this lesson, we will use the *dataMeta* R package to construct a data dictionary for a subset of the NHS England accident and emergency (A&E) attendances and admissions (`ae_attendances`) data from the *NHSRdatasets* package; we will then map the data and then save it as an R dataset (.rds) in your 'RawData' folder.

Load packages and read in the data

Let's load the packages and data needed for this document. The *dataMeta* packages is not installed on Noteable. Therefore, you will need to `install.packages('dataMeta')` every time you want to knit this document in RStudio in Notable. We will use the `read_csv()` function from the *readr* package from *tidyverse* to read in the data. *tidyverse* is a collection of essential R packages for data science. The *readr* package provides a fast and friendly way to read rectangular data from delimited files, such as comma-separated values (CSV) and tab-separated values (TSV). The * readr * package is loaded by the *tidyverse* package as one of its core components. We will use the *here* package to build a path relative to the top-level directory to read the raw ae_attendances data from our 'RawData' folder. The *here* package enables easy file referencing in project-oriented workflows. In contrast to using `setwd()` function, which is fragile and dependent on the way you organise your files, here uses the top-level directory of a project to easily build paths to files. The *lubridate* provides tools that make it easier to manipulate dates in R.

```
library(dataMeta)
library (tidyverse)
```

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.6
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(here)
```

```
## here() starts at /home/jovyan/B209978/B209978_assessment
```

Data

The data you will be managing on the course are from the NHSRdatasets package. This package has been created to support skills development in the NHS-R community and contains several free datasets. The dataset set I have chosen to manage from the NHSRdatasets package is the NHS England accident and emergency (A&E) attendances and admissions (`ae_attendances`) data. The `ae_attendances` data includes reported attendances, four-hour breaches and admissions for all A&E departments in England for 2016/17 through 2018/19 (Apr-Mar). We previously selected a subset of the variables needed for my data capture tool, including period, attendances and breaches, and subsetted the data into test and training data. However, for this lesson, we will use the data collected from the full `ae_attendances` dataset to demonstrate how to use the *dataMeta* to construct a data dictionary and append it to your collected data. The Jupyter Notebook "./ipynbScripts/CollectingDataUsingInteractiveJupyterWidgets.ipynb" was used to to collect the data.

Note, you only need to construct and append of data dictionary for the subset of the variables required for your data capture tool. We are using the full data set here, as you will be using the *dataMeta* R package construct and append a data dictionary for different variables from your `ae_attendances` data subset collected by your data capture tool.

Let us use the `read_csv()` function from the *readr* package to read your collected data from the Raw data folder.

```
CollectedData=read_csv(here("RawData", "CollectedDataFinal.csv"))
```

```
## Rows: 11 Columns: 9
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): org_code
## dbl  (6): index, attendances, breaches, admissions, breach_performance, admi...
## lgl  (1): consent
## date (1): period
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Let's view the CollectedData ae_attendances data

The `glimpse()` function is from *tibble* package and is used to view the columns/variables in a data frame. It also shows data type and some of the data in the data frame in each row. The *tibble* package provides utilities for handling tibbles. The *tibble* package is loaded by the *tidyverse* package, as one of its core components.

```
glimpse(CollectedData)
```

```
## Rows: 11
## Columns: 9
## $ index              <dbl> 2881, 2896, 4258, 4281, 5043, 6471, 7137, 7509, 957~
## $ period             <date> 2016-07-01, 2016-07-01, 2018-03-01, 2018-03-01, 20~
## $ org_code           <chr> "RXK", "RNA", "RXK", "RRK", "RLQ", "RWP", "RJC", "R~
## $ attendances        <dbl> 1488, 8947, 13805, 9936, 4532, 9817, 5811, 10313, 1~
## $ breaches           <dbl> 2128, 596, 3556, 2154, 1263, 2716, 297, 2824, 4432,~
## $ admissions         <dbl> 3141, 2599, 3429, 3896, 1437, 2921, 1617, 3174, 374~
## $ breach_performance <dbl> -0.4301075, 0.9333855, 0.7424122, 0.7832126, 0.3170~
## $ admission_rate     <dbl> 2.1108871, 0.2904884, 0.2483883, 0.3921095, 0.31707~
## $ consent            <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRU~
```

Here is the output of the 'glimpse()' function. It starts off with the number of rows and columns and each column in separate rows.

The CollectedData dataset contains:

- **index:** the index column that allows us to link the data collected to the original ae_attendances data in the 'RawData' folder.
- **period:** the month that this activity relates to, stored as a date (1st of each month).
- **org_code:** the Organisation data service (ODS) code for the organisation. The ODS code is a unique code created by the Organisation data service within NHS Digital and used to identify organisations across health and social care. ODS codes are required in order to gain access to national systems like NHSmail and the Data Security and Protection Toolkit. If you want to know the organisation associated with a particular ODS code, you can look it up from the following address: https://odsportal.digital.nhs.uk/Organisation/Search. For example, the organisation associated with the ODS code 'AF003' is Parkway health centre.
- **type:** the Department Type for this activity, either
  - **1:** Emergency departments are a consultant-led 24-hour service with full resuscitation facilities and designated accommodation for the reception of accident and emergency patients,

  - **2:** Consultant-led mono speciality accident and emergency service (e.g. ophthalmology, dental) with designated accommodation for the reception of patients, or

  - **other:** Other types of A&E/minor injury activity with designated accommodation for the reception of accident and emergency patients. The department may be doctor-led or nurse-led and treats at least minor injuries and illnesses and can be routinely accessed without an appointment. A service mainly or entirely appointment-based (for example, a GP Practice or Outpatient clinic) is excluded even though it may treat a number of patients with minor illness or injury. Excludes NHS walk-in centres.(National Health Service, 2020)
- **attendances:** the number of attendances for this department type at this organisation for this month.
- **breaches:** the number of attendances that breached the four-hour target.
- **admissions:** the number of attendances that resulted in an admission to the hospital.(Chris Mainey, 2021)
- **performance:** the performance (`[1 - breaches]/attendances`) calculated for the whole of England.
- **consent:** the consent from the end-user to process and share the data collected with the data capture tool.

Build a data dictionary for the data collected by the data capture tool

Build a linker data frame

We first need to build a linker data frame. To do this, we need to create two string vectors representing the different variable descriptions and the different variable types.

Variable descriptions

We need to create a string vector representing the different variable descriptions.

```
variable_description <- c("The index column that allows us to link the data collected to the original a
"The month that this activity relates to, stored as a date (1st of each month).",
"The Organisation data service (ODS) code for the organisation. If you want to know the organisation as
"The number of attendances for this department type at this organisation for this month.",
"The number of attendances that breached the four-hour target.",
"The number of attendances that resulted in an admission to the hospital.",
"The performance ([1 - breaches]/attendances)",
"The rate of admission (admissions/attendances)",
"The consent from the end-user to process and share the data collected with the data capture tool.")
print(variable_description)
```

```
## [1] "The index column that allows us to link the data collected to the original ae_attendances data
## [2] "The month that this activity relates to, stored as a date (1st of each month)."
## [3] "The Organisation data service (ODS) code for the organisation. If you want to know the organisa
## [4] "The number of attendances for this department type at this organisation for this month."
## [5] "The number of attendances that breached the four-hour target."
## [6] "The number of attendances that resulted in an admission to the hospital."
## [7] "The performance ([1 - breaches]/attendances)"
## [8] "The rate of admission (admissions/attendances)"
## [9] "The consent from the end-user to process and share the data collected with the data capture too
```

Variable types

We need to create a string vector representing the different variable types. It is a vector of integers with values 0 or 1. We need to use 0 for a variable with quantitative values (measured values) variables and 1 for fixed values (allowable values or codes) variables. Let us use The `glimpse()` function from *tibble* package to view the variable types in the CollectedData data frame.

```
glimpse(CollectedData)
```

```
## Rows: 11
## Columns: 9
## $ index              <dbl> 2881, 2896, 4258, 4281, 5043, 6471, 7137, 7509, 957~
## $ period             <date> 2016-07-01, 2016-07-01, 2018-03-01, 2018-03-01, 20~
## $ org_code           <chr> "RXK", "RNA", "RXK", "RRK", "RLQ", "RWP", "RJC", "R~
## $ attendances        <dbl> 1488, 8947, 13805, 9936, 4532, 9817, 5811, 10313, 1~
## $ breaches           <dbl> 2128, 596, 3556, 2154, 1263, 2716, 297, 2824, 4432,~
## $ admissions         <dbl> 3141, 2599, 3429, 3896, 1437, 2921, 1617, 3174, 374~
## $ breach_performance <dbl> -0.4301075, 0.9333855, 0.7424122, 0.7832126, 0.3170~
## $ admission_rate     <dbl> 2.1108871, 0.2904884, 0.2483883, 0.3921095, 0.31707~
## $ consent            <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRU~
```

We have six quantitative values (measured values) variables and three fixed values (allowable values or codes) variables.

```
variable_type <- c(0, 1, 1, 0, 0, 0, 0, 0, 1)
print(variable_type)
```

```
## [1] 0 1 1 0 0 0 0 0 1
```

Now let us use the `build_linker()` function from the *dataMeta* package to constructs an intermediary (linker) data frame between the CollectedData and the data dictionary. For this function to run, it requires the CollectedData data frame and variable_description and variable_type string vectors as inputs.

```r
linker<-build_linker(CollectedData, variable_description, variable_type)
print(linker)
```

```
##              var_name
## 1               index
## 2              period
## 3            org_code
## 4         attendances
## 5            breaches
## 6          admissions
## 7 breach_performance
## 8      admission_rate
## 9             consent
##
## 1
## 2
## 3 The Organisation data service (ODS) code for the organisation. If you want to know the organisatio
## 4
## 5
## 6
## 7
## 8
## 9
##   var_type
## 1        0
## 2        1
## 3        1
## 4        0
## 5        0
## 6        0
## 7        0
## 8        0
## 9        1
```

Data dictionary

We are now going to use the `build_dict()` function from the *dataMeta* to constructs a data dictionary for a CollectedData data frame with the aid of the linker data frame between. For this function to run, it requires the CollectedData and linker data frames and variable_description as inputs.

```r
dictionary <- build_dict(my.data = CollectedData, linker = linker)
```

```
## Enter description for variable 'admission_rate' and option '0.237557194882809 to 2.11088709677419':
## Enter description for variable 'admissions' and option '1437 to 4407':
## Enter description for variable 'attendances' and option '1488 to 13805':
## Enter description for variable 'breach_performance' and option '-0.43010752688172 to 0.93338549234438
## Enter description for variable 'breaches' and option '297 to 4432':
## Enter description for variable 'consent' and option 'TRUE':
## Enter description for variable 'index' and option '2881 to 12530':
## Enter description for variable 'org_code' and option 'RXK':
## Enter description for variable 'org_code' and option 'RNA':
## Enter description for variable 'org_code' and option 'RRK':
## Enter description for variable 'org_code' and option 'RLQ':
## Enter description for variable 'org_code' and option 'RWP':
## Enter description for variable 'org_code' and option 'RJC':
```

```
## Enter description for variable 'org_code' and option 'RKB':
## Enter description for variable 'org_code' and option 'RL4':
## Enter description for variable 'period' and option '16983':
## Enter description for variable 'period' and option '17591':
## Enter description for variable 'period' and option '17532':
## Enter description for variable 'period' and option '17410':
## Enter description for variable 'period' and option '17348':
## Enter description for variable 'period' and option '17318':
## Enter description for variable 'period' and option '17866':
## Enter description for variable 'period' and option '17805':
## Enter description for variable 'period' and option '17622':
```

```
glimpse(dictionary)
```

```
## Rows: 24
## Columns: 4
## $ `variable name`        <chr> "admission_rate", "admissions", "attendances", ~
## $ `variable description` <chr> "The rate of admission (admissions/attendances)~
## $ `variable options`     <chr> "0.237557194882809 to 2.11088709677419", "1437 ~
## $ notes                  <chr> "", "", "", "", "", "", "", "", "", "", "", "",~
```

```
# dictionary[7,4]<-"RDZ: NHS Trust - The Royal Bournemouth and Christchurch Hospitals NHS Foundation Tr
# #You should do that for all Organisation data service (ODS) in the data dictionary
# dictionary[27,4] <-"other: Other types of A&E/minor injury activity with designated accommodation for
# dictionary[28,4]<- "1: Emergency departments are a consultant-led 24-hour service with full resuscita
# dictionary[29,4] <- "2: Consultant-led mono speciality accident and emergency service (e.g. ophthalmo
```

Let’s save the data dictionary for CollectedData to the ‘RawData’ folder

Of note, when naming folders and files, you must do so in a consistent, logical and predictable way means that information may be located, identified and retrieved by your and your colleagues as quickly and efficiently as possible. With this in mind, let’s name this file “CollectedData_DataDictionary” and write it to the raw data folder.

```
glimpse(dictionary)
```

```
## Rows: 24
## Columns: 4
## $ `variable name`        <chr> "admission_rate", "admissions", "attendances", ~
## $ `variable description` <chr> "The rate of admission (admissions/attendances)~
## $ `variable options`     <chr> "0.237557194882809 to 2.11088709677419", "1437 ~
## $ notes                  <chr> "", "", "", "", "", "", "", "", "", "", "", "",~
```

```
write_csv(dictionary, here("RawData", "CollectedData_DataDictionary.csv"))
```

Append data dictionary to the CollectedData

We will now incorporate attributes as metadata to the CollectedData as metadata using the ‘incorporate_attr()’ function from the *dataMeta* package. For this function to run, it requires the CollectedData and dictionary and main_string main_string as inputs. main_string is a character string describing the CollectedData data frame.

```

main_string <- "This data describes the NHS England accident and emergency (A&E) attendances and breache
main_string
```

Create main_string for attributes

```
## [1] "This data describes the NHS England accident and emergency (A&E) attendances and breaches of fo
```

Incorporate attributes as metada We are using the ‘incorporate_attr()’ function to return an R dataset containing metadata stored in its attributes. The attributes we are going to add include: * a data dictionary * number of columns * number of rows * the name of the author who created the dictionary and added it, * the time when it was last edited * a brief description of the original dataset.

```
complete_CollectedData <- incorporate_attr(my.data = CollectedData, data.dictionary = dictionary,
main_string = main_string)
#Change the author name
attributes(complete_CollectedData)$author[1]<-"Mairead Bermingham"
complete_CollectedData
```

```
## # A tibble: 11 x 9
##    index period     org_code attendances breaches admissions breach_performance
##  * <dbl> <date>     <chr>          <dbl>    <dbl>      <dbl>              <dbl>
##  1  2881 2016-07-01 RXK             1488     2128       3141             -0.430
##  2  2896 2016-07-01 RNA             8947      596       2599              0.933
##  3  4258 2018-03-01 RXK            13805     3556       3429              0.742
##  4  4281 2018-03-01 RRK             9936     2154       3896              0.783
##  5  5043 2018-01-01 RLQ             4532     1263       1437              0.317
##  6  6471 2017-09-01 RWP             9817     2716       2921              0.298
##  7  7137 2017-07-01 RJC             5811      297       1617              0.278
##  8  7509 2017-06-01 RWP            10313     2824       3174              0.308
##  9  9577 2018-12-01 RXK            13604     4432       3744              0.275
## 10 10327 2018-10-01 RKB            12519     1937       4407              0.352
## 11 12530 2018-04-01 RL4            10709     1704       2544              0.238
## # ... with 2 more variables: admission_rate <dbl>, consent <lgl>
```

```
attributes(complete_CollectedData)
```

```
## $row.names
##  [1]  1  2  3  4  5  6  7  8  9 10 11
##
## $names
## [1] "index"              "period"             "org_code"
## [4] "attendances"        "breaches"           "admissions"
## [7] "breach_performance" "admission_rate"     "consent"
##
## $spec
## cols(
##   index = col_double(),
##   period = col_date(format = ""),
##   org_code = col_character(),
##   attendances = col_double(),
##   breaches = col_double(),
##   admissions = col_double(),
##   breach_performance = col_double(),
##   admission_rate = col_double(),
##   consent = col_logical()
## )
##
## $problems
## <pointer: 0x562fbc624330>
##
```

```
## $class
## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
##
## $main
## [1] "This data describes the NHS England accident and emergency (A&E) attendances and breaches of fou
##
## $dictionary
##         variable name
## 1      admission_rate
## 2          admissions
## 3         attendances
## 4  breach_performance
## 5            breaches
## 6             consent
## 7               index
## 8            org_code
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16             period
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8  The Organisation data service (ODS) code for the organisation. If you want to know the organisati
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
##                           variable options notes
## 1   0.237557194882809 to 2.11088709677419
## 2                             1437 to 4407
## 3                            1488 to 13805
## 4  -0.43010752688172 to 0.933385492343802
## 5                              297 to 4432
## 6                                     TRUE
## 7                            2881 to 12530
## 8                                      RXK
## 9                                      RNA
## 10                                     RRK
## 11                                     RLQ
## 12                                     RWP
## 13                                     RJC
## 14                                     RKB
## 15                                     RL4
## 16                                   16983
## 17                                   17591
## 18                                   17532
## 19                                   17410
## 20                                   17348
## 21                                   17318
## 22                                   17866
## 23                                   17805
## 24                                   17622
##
## $last_edit_date
## [1] "2022-06-22 16:08:49 UTC"
##
## $author
## [1] "Mairead Bermingham"
```

Save the CollectedData with attributes We are using the 'save_it()' function to save the CollectedData with attributes stored as metadata as an R dataset (.rds) into the 'current working directory'RawData' folder. This is the final function used in this package. For the function to run, the complete_CollectedData, and the name of the file as a text string to name the file are required as inputs.

```
save_it(complete_CollectedData, here("RawData", "complete_CollectedData"))
```

If you would like to load this data later, here is the code to do so

```
complete_CollectedData<-readRDS(here("RawData", "complete_CollectedData.rds"))
```

Well done! You now have now saved your CollectedData_DataDictionary and enriched CollectedData_DataDictionary.rds file to your 'Rawdata' folder. Happy coding!