Re-encoding `people` in the EDH dataset

Antonio Rivero Ostoic

September 2022

```r
# load and check versions
library(sdam)
packageVersion("sdam")
```

```
[1] '1.0.0'
```

EDH people

- `EDH` is a dataset in `"sdam"` that contains the texts of Latin and Latin-Greek inscriptions of the Roman Empire, which have been retrieved from the Epigraphic Database Heidelberg API repository through routines `get.edh()` and `get.edhw()`.

Since the year 2022 and still today, the API repository does not support people variables, and the `EDH` dataset serves as an alternative for the analysis of people-related inscriptions.

One challenge with people variables in `EDH` is that some records contain characters in Greek and Latin extended that need re-encoding for a proper rendering and display.

Re-encoding `people` in `EDH`

Ancient inscriptions in some Roman provinces have Greek characters written and, due to encoding and decoding steps in the process of extraction, loading, and transformation of the data (perhaps Treating UTF-8 Bytes as Windows-1252?), Greek and other Latin characters are not displayed properly with the actual version of the `EDH` dataset. Most of the encoding issues are in variables related to people, and some examples with inscriptions in Roman provinces are next.

Achaia

The Roman province of **Achaia** in the `EDH` dataset has inscriptions related to people.

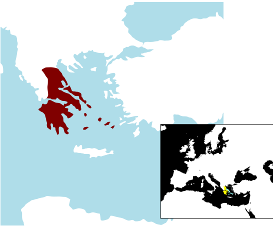

Figure 2: Roman province of Achaia (ca 117 AD).

Function `edhw()` is to obtain the available inscriptions per province in the EDH dataset, which is a list that is the input for the same function to extract `people` variables *cognomen* and *nomen*. In this case, the `'province'` argument is `Ach` that stands for `Achaia`.

```r
# select two people variables from Achaia
Ach <- edhw(province="Ach") |>
  edhw(vars="people", select=c("cognomen","nomen"))
```

There are 1539 records with people in `Ach` that corresponds to the number of rows in this data frame.

```r
# number of people entries in Achaia
nrow(Ach)
```

```
[1] 1539
```

However, some records have either missing data or are inscriptions where *cognomen* and *nomen* are not available.

```r
# also remove NAs
Ach <- edhw(province="Ach") |>
  edhw(vars="people", select=c("cognomen","nomen"), na.rm=TRUE)

nrow(Ach)
```

```
[1] 1465
```

Clean function for re-encoding

Treating with `people` attribute variables requires many times re-encoding that is one option in function `cln()`. For instance, values in *cognomen* in the first entries of `Ach` are likely in Greek.

```r
# some people entries in Achaia
head(Ach)
```

```
        id                                             cognomen                   nomen
1 HD001917                                                Rufus Ponponius (= Pomponius)
2 HD001917                                                  Eia    Ponponia (= Pomponia)
3 HD001917                           Î<U+0094>á½¹Î¾Î± Î\235á½·ÎºÎ·                       <NA>
4 HD002097 Î<U+0092>Î±Î»Î»ÎµÎ½Ï<U+0084>Î¹Î½Î¹Î±Î½á½¹Ï<U+0082>+                    <NA>
5 HD002097                            Î<U+0092>á½±Î»Î·Ï<U+0082>                    <NA>
6 HD002097                                            Arcadius+                    <NA>
```

Function `cln()` serves to re-encode Greek and Latin characters to render Greek, Greek extended, and Latin extended glyphs.

```r
# re-encode in Ach cognomen
Ach$cognomen |>
  head() |>
  cln()
```

```
cognomen

Rufus
Eia
ΔόξαΝίκη
Βαλλεντινιανός+
Βάλης
Arcadius+
```

For *cognomen* in the last people entries in `Achaia`.

```r
# last entries
tail(Ach)
```

```
           id                                                                        cognomen
1534 HD068263                                              Î<U+009A>á½±Î»Î»Ï<U+0085>Ï<U+0082>
1535 HD068315 Î¦Ï\201Î¿Î½Ï<U+0084>Îµá¿<U+0096>Î½Î¿Ï<U+0082> Î\235ÎµÎ¹Îºá½µÏ\201Î±Ï<U+0084>Î¿Ï<
1536 HD068319 Î¦Ï\201Î¿Î½Ï<U+0084>Îµá¿<U+0096>Î½Î¿Ï<U+0082> Î\235ÎµÎ¹Îºá½µÏ\201Î±Ï<U+0084>Î¿Ï<
1537 HD072342                                          Î<U+0091>á¼°Î¼Î¹Î»Î¹Î±Î½á½¹Ï<U+0082>+
1538 HD072342                                              Î<U+009A>Î±Î¹Î»Î¹Î±Î½á½¹Ï<U+0082>+
1539 HD078079                                                                           Eburo
                               nomen
1534                            <NA>
1535 Î<U+009A>Î»Î±á½»Î´Î¹Î¿Ï<U+0082>
1536 Î<U+009A>Î»Î±á½»Î´Î¹Î¿Ï<U+0082>
1537 Î<U+009F>á½\220á½±Ï\201Î¹Î¿Ï<U+0082>+
1538                            <NA>
1539                            <NA>
```

After re-encoding the last records in `Ach` with `cln()`, it is easier to see, for example, that some have identical *cognomen* where entries having `<NA>` in the input become `NA`.

```r
# clean last entries of cognomen
Ach$cognomen |>
  tail() |>
  cln()
```

```
cognomen

Κάλλυς
ΦροντεῖνοςΝεικήρατος
ΦροντεῖνοςΝεικήρατος
Αἰμιλιανός+
Καιλιανός+
Eburo
```

```r
# clean last entries of nomen
Ach$nomen |>
  tail() |>
  cln()
```

```
nomen

NA
Κλαύδιος
Κλαύδιος
Οὐάριος+
NA
NA
```

Re-encode Greek and Latin within data frames

Aegyptus

In the case of the province of **Aegyptus**, three people variables have a mixing og Greek and Latin characters scripted that need *re-codification* as well.

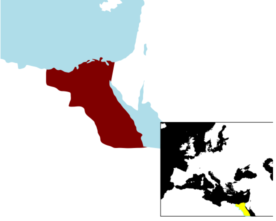

Figure 3: Roman province of Aegyptus (ca 117 AD).

```
# Aegyptus people
Aeg <- edhw(province="Aeg") |>
  edhw(vars="people")
```

```
# three variables of the last eight records
Aeg[ , c(3,5:6)] |>
  tail(8)
```

```
                                                                         cognomen
81                             Augustus+ / Î£ÎµÎ²Î±Ï<U+0083>Ï<U+0084>á½¹Ï<U+0082>
82                                                   Aquila / á¼<U+0088>Îºá½»Î»Î±
83 Traianus Hadrianus / Î¤Ï\201Î±Î¹Î±Î½á½¸Ï<U+0082> á¼<U+0089>Î´Ï\201Î¹Î±Î½á½¹Ï<U+0082>
84                                               Serenus / Î£ÎµÏ\201Î·Î½á½¹Ï<U+0082>
85                     Domitianus+ / Î<U+0094>Î¿Î¼Î¹Ï<U+0084>Î¹Î±Î½á½¹Ï<U+0082>++
86                               Vegetus / Î<U+009F>á½\220ÎÎ³ÎµÏ<U+0084>Î¿Ï<U+0082>
87                                Î<U+009B>Ï<U+0085>Ï<U+0083>á¾¶Ï<U+0082> / Lysas
88                                            Î Î»á½¹ÎºÎ±Î¼Î¿Ï<U+0082> / Plocamus

81 Imp. Caesar divi f. August. / Î<U+0091>á½\220Ï<U+0084>Î¿ÎºÏ\201á½±Ï<U+0084>Ï<U+0089>Ï\201 Î
82                                                                                         C.
83
84                      Sulpic. Serenus / Î£Î¿Ï<U+0085>Î»Ï<U+0080>á½·ÎºÎ¹Î¿Ï<U+0082> Ï<U+0085>
85
86                                                        G. Septimio Vegeto / Î<U+0093>Î±á¿<U
87                                           Î<U+009B>Ï<U+0085>Ï<U+0083>á¾¶Ï<U+0082> Î Î¿Ï<U+0
88                                                                                         Î Î
                                                   nomen
81             Caesar / Î<U+009A>Î±á¿<U+0096>Ï<U+0083>Î±Ï\201
82                        Iulius / á¼¸Î¿á½»Î»Î¹Î¿Ï<U+0082>
83                                                    <NA>
84 Sulpicius* / Î£Î¿Ï<U+0085>Î»Ï<U+0080>á½·ÎºÎ¹Î¿Ï<U+0082>
85                                                    <NA>
86    Septimius / Î£ÎµÏ<U+0080>Ï<U+0084>á½·Î¼Î¹Î¿Ï<U+0082>
87                                                    <NA>
88                    á¼<U+008C>Î½Î½Î¹Î¿Ï<U+0082> / Annius
```

For people in `Aegyptus`, columns three, and five to six correspond to *cognomen*, *name*, and *nomen*, where the output from `cln()` in the console is a dataframe.

```r
# re-encode three variables from last entries
Aeg[ ,c(3,5:6)] |>
  tail() |>
  cln()
```

`cognomen`

Augustus+ / Σεβαστός
Aquila / Ἀκύλα
Traianus Hadrianus / ΤραιανὸςἉδριανός
Serenus / Σερηνός
Domitianus+ / Δομιτιανός++
Vegetus / Οὐέγετος
Λυσᾶς / Lysas
Πλόκαμος / Plocamus

`name`

Imp. Caesar divi f. August. / ΑὐτοκράτωρΚαῖσαρθεοῦυἱὸςΣεβαστὸς
C. Iulio Aquila / ΓαΐουἸουλίουἈκύλα
Traiani Hadriani / ΤραιανοῦἉδριανοῦ
Sulpic. Serenus / ΣουλπίκιοςυἱὸςΓναίουΚουιρίναΣερηνὸς
[Domitiani] / [[Δομιτια
G. Septimio Vegeto / ΓαΐουΣεπτιμίουΟὐεγέτου
ΛυσᾶςΠοπλίουἈννίουΠλοκάμου / Lysas P. Anni Plocami
ΠοπλίουἈννίουΠλοκάμου / P. Anni Plocami

`nomen`

Caesar / Καῖσαρ
Iulius / Ἰούλιος
NA
Sulpicius* / Σουλπίκιος
NA
Septimius / Σεπτίμιος
NA
Ἄννιος / Annius

Some entries in `Aeg` have Greek extended characters, and one entry in Latin has a special character at the end (`Sulpicius*`), which can be omitted for further computations by raising the cleaning level to 2.

nomen in Aegyptus

Benefits from re-encoding and cleaning text from the `EDH` dataset are evident like when counting occurrences in the different attribute variables as with `nomen` in `Aeg`.

```r
# default cleaning level 1
Aeg$nomen |>
  cln() |>
  table() |>
  sort(decreasing=TRUE)
```

Sempronius+

```
[1] 4
```

Κούρτιος

```
[1] 2
```

Μέμμιος

```
[1] 2
```

Ἰούλιος

```
[1] 2
```

etc.

```
...
```

By raising the cleaning level to 2, all special characters are removed from the end, and it is possible to see that, in the Roman province of Aegyptus, `Sempronius`, `Sentius`, `Valerius` are the three most common *nomen* in inscriptions with four occurrences each.

```
# raise cleaning level and remove NAs
Aeg$nomen |>
  cln(level=2, na.rm=TRUE) |>
  table() |>
  sort(decreasing=TRUE)
```

Sempronius

```
[1] 4
```

Sentius

```
[1] 4
```

Valerius

```
[1] 4
```

Κούρτιος

```
[1] 2
```

etc.

```
...
```

Caveats

See `Warnings` section in manual.