4-Datenbereinigung

4a-combined-data.R

Dieses Programm liest alle annotierten Stücke[1] ein (im .tsv Format) und kombiniert sie zu einer grossen .tsv Tabelle. Folgende Sonderheiten wurden beachtet:

- Interpretation der Einträge als Text und nicht Werte (im Annotationsstandard wird '%' für eine halbverminderten Akkord benutzt, dasselbe Zeichen wird in R zur Kommentierung benutzt; weiter wird in der Annotation 'F' für die Note verwendet, R versteht das jedoch als logischen Wert 'FALSE') Dazu wurde beim Einlesen der Tabellen folgendes festgelegt:

```
21  data <- read.table(file, header = TRUE, sep = "\t", fill = TRUE, comment.char = "", stringsAsFactors =
            FALSE, colClasses = "character")
```

- Unterschiedliche Spaltenüberschriften (und dadurch fehlende Spalten), was durch das Hinzufügen von Spalten folgendermassen gelöst wird:

```
40  combine_data <- function(data_list, all_column_names) {
41      combined_data <- data.frame()
42      for (data in data_list) {
43          missing_cols <- setdiff(all_column_names, names(data))
44          for (col in missing_cols) {
45              data[[col]] <- ""
46          }
47          combined_data <- rbind(combined_data, data)
48      }
49      return(combined_data)
50  }
```

[1] Die Tabellen müssen, damit sie von den Skripten erkannt und eingelesen werden können, wie folgt benannt werden:

[KOM]-[A][xx]-M[x].tsv

wobei KOM für das Komponistenkürzel steht (LVB für L. v. Beethoven, WAM für W. A. Mozart), A für die Stückart (S für Sonate, Q für Streicherquartett) mit der üblichen Nummerierung (xx aus 00-99) und M für den Satz mit Nummer (x aus 0-9)

Die Datei zum dritten Satz der ersten Beethovensonate heisst also: LVB-S01-M3.tsv