

# UTF-8 vejledning

---

*Vejledning til indkodning af statistikfiler og tekstfiler som UTF-8. Supplement til brugervejledning til ASTA (Aflevering af Statistikfiler Til Arkiv).*

*Rigsarkivet december 2019*

## Indhold

<b>0. Læsevejledning til UTF-8 vejledningen .....</b>	<b>2</b>
A. Vejledningens målgruppe og anvendelse.....	2
B. Henvi sning til øvrig vejledning.....	2
C. Lovgivning og retsforskrifter.....	2
D. Definitioner.....	2
<b>1. Hvad er UTF-8 indkodning? .....</b>	<b>3</b>
<b>2. Hvad er konsekvensen, hvis data ikke er indkodet som UTF-8 tegnsæt? .....</b>	<b>3</b>
<b>3. Hvordan aflæses og ændres statistikfilers indkodning i statistikprogrammer? .....</b>	<b>3</b>
A. SAS – syntakser og procedurer til tegnsæt.....	4
B. SPSS – syntakser og procedurer til tegnsæt.....	5
C. Stata – syntakser og procedurer til tegnsæt .....	7
<b>4. Hvordan kontrolleres tegnsæt i en tekstfil? .....</b>	<b>9</b>
<b>5. Hvordan aflæses UTF-8 hex-værdier i en tekstfil?.....</b>	<b>9</b>
<b>6. UTF-8 support i Rigsarkivet.....</b>	<b>10</b>

## 0. Læsevejledning til UTF-8 vejledningen

Offentlige myndigheder, herunder forskningsinstitutioner, afleverer data til Rigsarkivet i form af arkiveringsversioner og afleveringspakker. Krav til disse afleveringer er beskrevet i Rigsarkivets bekendtgørelse om arkiveringsversioner. Et af kravene er, at de data, der afleveres, skal indkodes som UTF-8.

**UTF-8 vejledningen er en teknisk vejledning, der forklarer, hvad UTF-8 indkodning er, hvordan du tjekker, om tegnsættet er indkodet som UTF-8 samt, hvordan du indkoder tegnsæt som UTF-8 i statistikprogrammerne SAS, Stata og SPSS eller i en teksteditor.**

### A. Vejledningens målgruppe og anvendelse

UTF-8 vejledningen henvender sig til dem, som producerer afleveringspakker med dataudtræk fra statistikfiler til arkivet.

### B. Henvisning til øvrig vejledning

Foruden UTF-8 vejledningen har Rigsarkivet udarbejdet andre vejledninger, der har betydning for produktion og aflevering af afleveringspakker:

- Quickguide – til produktion og test af en afleveringspakke med ASTA
- Vejledning til bilag 9 om afleveringspakker i bekendtgørelse om arkiveringsversioner
- Brugervejledning til ASTA
- Vejledning til produktion af afleveringspakke med data fra regneark eller csv-filer
- Vejledning til Skab archiveIndex
- Vejledning til Skab contextDocumentationIndex
- Vejledning om konvertering af dokumenter til TIFF
- Eksempelafleveringspakke med statistikdata AVID.SA.18005.1

Alt vejledningsmateriale findes i ASTA og kan tilgås fra Rigsarkivets hjemmeside [www.sa.dk](http://www.sa.dk).

### C. Lovgivning og retsfor skrifter

Information om lovgivning m.v. findes på Rigsarkivets hjemmeside [www.sa.dk](http://www.sa.dk).

### D. Definitioner

**Afleveringspakke med data fra statistikfiler** består overordnet set af kontekstdokumenter, der skal afleveres i Rigsarkivets arkivformater, udtræk af data og metadata fra de statistikfiler, som skal afleveres, samt to indeksfiler i xml-format, der indeholder overordnet metadata om de afleverede data og kontekstdokumenterne.

## 1. Hvad er UTF-8 indkodning?

Ord og sætninger i tekst er lavet af tegn, som er de bogstaver, som man kan se. Tegn gemmes i computeren som en eller flere bytes, som er repræsenteret af tal. For at lave en oversættelse mellem tegn og byte anvender man specielle koder. Indkodning er et sæt af koder imellem bytes i computeren og karakter og tegn, og som hjælper med oversættelse imellem de to. Et tegn kan repræsenteres af mere end én byte. For eksempel er tegnet "a" oversat til bytes "97", mens "@" er oversat til bytes "90".

UTF-8 står for "Unicode Transformation Format". UTF-8 er et af de tre standard indkodninger (tegn sæt) som bruges til at repræsentere Unicode som computertekst (de andre er UTF-16 og UTF-32). UTF-8 bruger en algoritme til at dekode data imellem en binær form, der bruges af computere, til et sæt af tal, der kan oversættes til tegn. '8' i UTF-8 betyder, at indkodningen bruger 8-bit blokke til at repræsentere et tegn.

UTF-8 er normalt den mest effektive måde at lagre Unicode-tekst. Desuden understøtter UTF-8 mange forskellige sprog. Dette har medført, at det er den mest anvendte Unicode-indkodning i dag.

Der er mange forskellige indkodninger. Hvis indkodningsinformation ikke er korrekt angivet for filen, bliver visualisering af teksten ødelagt, dvs. man ser ikke de korrekte tegn fx å vises som □ eller lignende.

## 2. Hvad er konsekvensen, hvis data ikke er indkodet som UTF-8 tegnsæt?

I forbindelse med aflevering af statistikfiler i form af en afleveringspakke til arkivet er det den afleverende myndigheds ansvar at udtrække data og metadata fra statistikfilerne til afleveringspakken. Det er den, der producerer afleveringspakken, der skal sikre, at alle tegn er korrekt indkodet som UTF-8 før udtræk.

Hvis et datasæt oprindeligt stammer fra en af de nyere versioner af SAS, SPSS eller STATA, er tegnsættet i statistikfilen højst sandsynligt indkodet som UTF-8, fordi nyere versioner af statistikprogrammerne anvender denne som default indkodning. Hvis statistikprogrammets opsætning ikke er Unicode, kan du ændre denne opsætning i 'Preference'/'Options' i statistiskprogrammet og derefter gemme filen som Unicode før udtræk af data til afleveringspakken via programmet ASTA og dermed sikre dig, at alle tegn vises korrekt.

Hvis dit datasæt stammer fra et andet program eller oprindeligt har en anden indkodning fx ANSI og importeres til et statistikprogram, som anvender Unicode som default, kan dette forårsage at nogle tegn vises forkert fx et ord som 'stå' kan vises som 'st□', da transformationen kan påvirke tegnene. Er dette sket, er det vigtigt at rette tegnene, som er forkerte, så andre i fremtiden kan bruge, læse og forstå datasættet.

Rigsarkivets værktøj, der kan anvendes til test af afleveringspakken (ASTA) før aflevering til arkiv, tester ikke automatisk for tegnsættet i data- og metadatafilen. Rigsarkivet tester visuelt alle afleverede data- og metadatafiler efter aflevering. Hvis der er ugyldige tegn i afleveringen, som ikke er UTF-8 tegn, får arkiverskaber besked herom, og der kræves nogle justeringer i datafilen eventuelt nye dataudtræk og en genaflevering vil muligvis være nødvendig.

Derfor er det vigtigt, at du visuelt tjekker din datafil udtrukket til afleveringspakken for at sikre dig, at alle tegn vises korrekt og kan tydes.

## 3. Hvordan aflæses og ændres statistikfilers indkodning i statistikprogrammer?

Sørg for at kontrollere, at statistikfilens indkodning er UTF-8, før du laver udtræk af data til afleveringspakken.

Hvert statistikprogram har sin egen syntaks til at undersøge tegnsættet i et datasæt og ændre dette til et andet tegnsæt. Nedenfor finder du procedurer og syntakser for SAS, Stata og SPSS. Ved brug af disse kan du sikre dig, at data er indkodet som UTF-8 tegnsæt.

## A. SAS – syntakser og procedurer til tegnsæt

### Undersøg SAS-filens indkodning/tegnset

For at identificere datasættets indkodning/tegnset i en SAS-datafil, skal du følge disse trin, som er anbefalet af SAS<sup>1</sup>:

- Kør følgende SAS-syntaks for at bestemme indkodningen/tegnsettet for et datasæt i SAS. Du skal kun erstatte libref.data\_set\_name med dit biblioteks navn og filnavn (for eksempel "mylib.mydata").

#### **SAS-Syntaks**

```
%let dsn=libref.data_set_name;  
%let dsid=%sysfunc(open(&dsn,i));  
%put &dsn ENCODING is: %sysfunc(attrc(&dsid,encoding));
```

#### **Eksempel**

```
%let dsn=dgi.customerdaga;  
%let dsid=%sysfunc(open(&dsn,i));  
%put &dsn ENCODING is: %sysfunc(attrc(&dsid,encoding));
```

En anden måde at finde SAS-filens indkodning/tegnset er at køre en "proc contents", som i output viser filens indkodning/tegnset. Sådan:

#### **SAS-Syntaks**

```
PROC CONTENTS <option-1 <...option-n>>;  
run;
```

#### **Eksempel**

```
PROC CONTENTS data=dgi.customerdata;
```

### Skift SAS-filens indkodning/tegnset til UTF-8:

For at ændre indkodning/tegnset af en SAS-fil, der ikke tidligere har været defineret som UTF-8, kan nedenstående syntaks anvendes<sup>2</sup>. Bemærk at du skal erstatte følgende:

- 1) libref samt dets placering
- 2) Angiv den placering, du ønsker at gemme den nye UTF-8-fil i (Syntaksens anden linje)
- 3) Angiv det ønskede datasætnavn til den nye UTF-8-fil (Syntaksens fjerde linje)

<sup>1</sup> <http://support.sas.com/kb/14/290.html>

<sup>2</sup> <http://support.sas.com/kb/15/597.html>

**SAS-Syntaks**

```
libname inlib libref 'c:\xxxx';  
libname outlib 'c:\yyy' outencoding='UTF-8';  
proc copy noclone in=inlib out=outlib;  
select dataset_name;  
run;
```

**Eksempel**

```
libname inlib dgi 'c:\temp';  
libname outlib 'c:\temp\out' outencoding='UTF-8';  
proc copy noclone in=inlib out=outlib;  
select customerdata;  
run;
```

**B. SPSS – syntakser og procedurer til tegnsæt****Indkodning/tegnset i forskellige SPSS-versioner:**

Som beskrevet af IBM SPSS<sup>3</sup>:

- Frem til version 15 er alt indkodning/tegnset i SPSS baseret på code pages.
- Fra version 16 til 20 er Unicode (som UTF-8) også understøttet. UTF-8 er kaldt **”Unicode mode”** i SPSS 16. Bemærk at UTF-8 indkodning er både understøttet i datasæt og i syntaksfiler.
- Fra SPSS-version 21 og derefter spørger programmet, om **”Unicode mode”** skal anvendes, når det startes.

**Undersøg SPSS-filens indkodning/tegnset**

Følgende syntax kan køres i SPSS for at identificere om SPSS opsætning er i Unicode.

```
SPSS-Syntaks  
SHOW UNICODE
```

**Undersøg og skift SPSS-filens indkodning/tegnset**

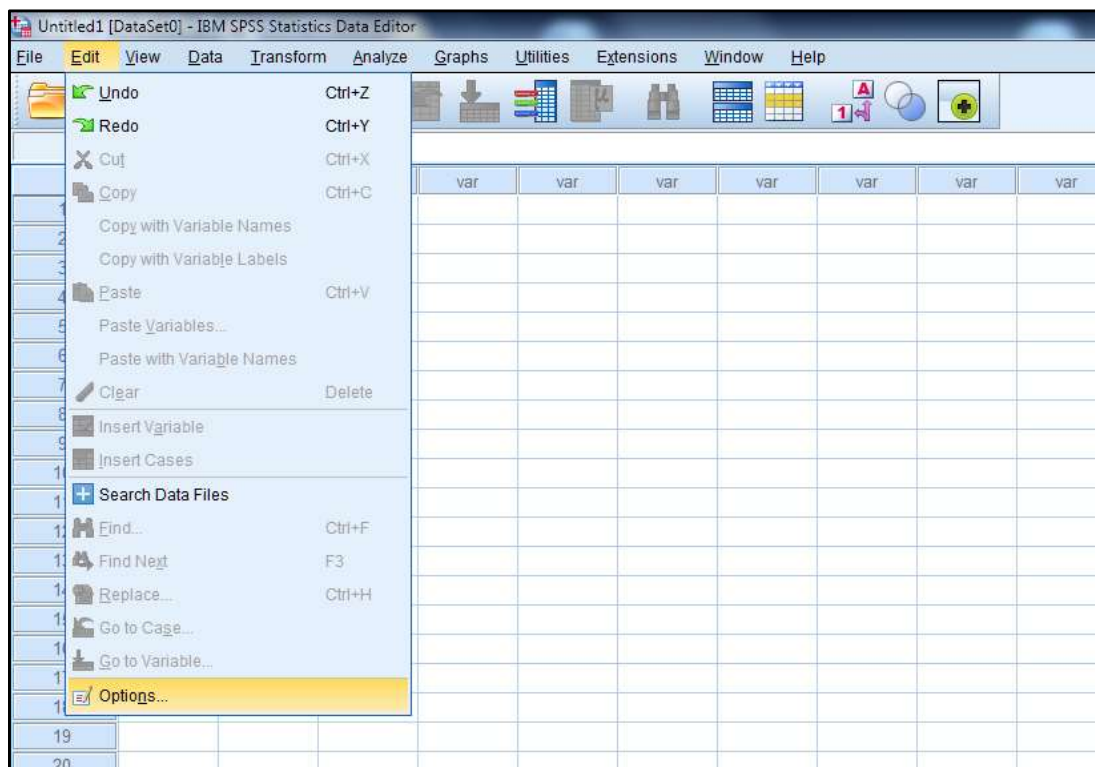
For at identificere og skifte indkodning i SPSS, skal du åbne SPSS, og før du åbner datasættet, som du vil undersøge, skal du klikke på **”Edit”** og vælge **’Options’** i menuen (se figur 1).

Et vindue åbnes nu med alle **”Options”**. Vælg fanebladet **”Language”** (se figur 2). Marker **”Unicode (universal character set)”** for at vælge UTF-8 som SPSS default indkodning/tegnset for data og syntakser. Klik på **’OK’**.

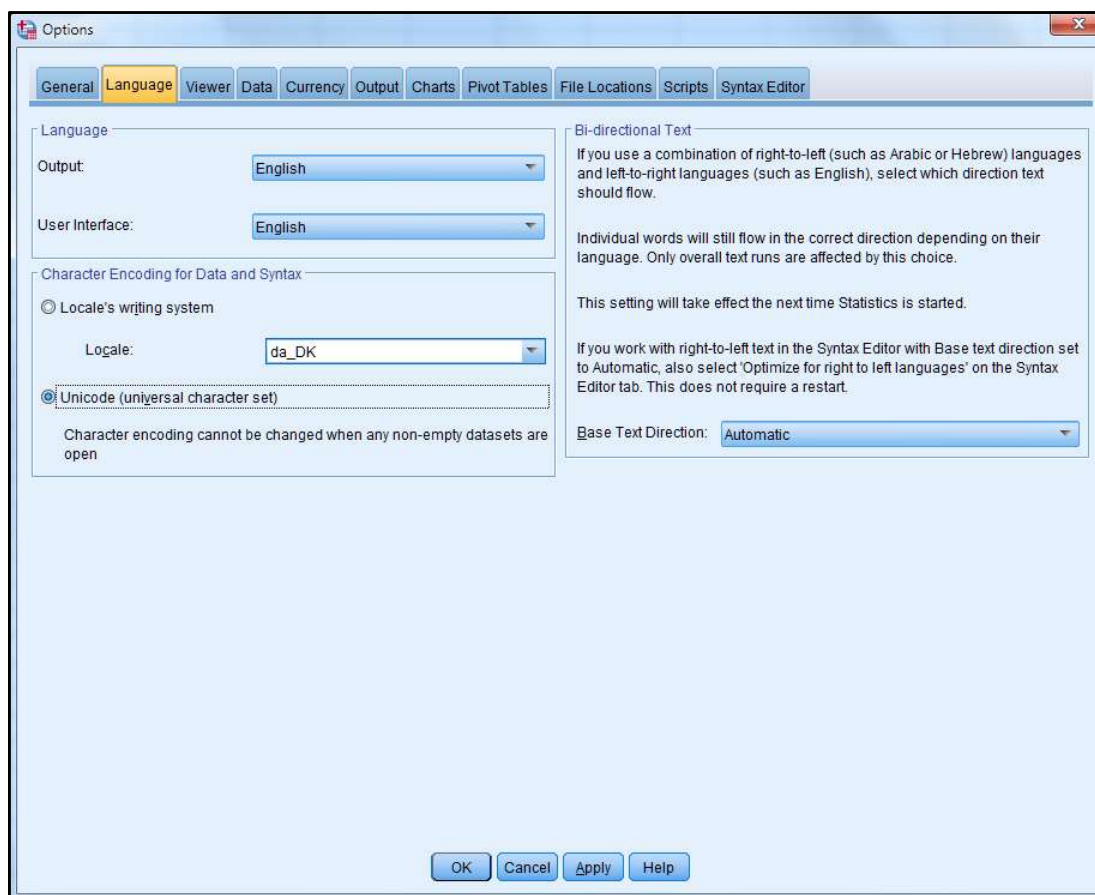
Du kan kontrollere, om ændringen er sket ved at kigge i bunden til højre af SPSS program-vinduet, som skal vise **’Unicode: ON’** (se figur 3).

Hvis der i feltet markeret med rød cirkel i figur 3 vises **’Unicode:OFF’**, betyder det, at der i **”language settings”** (se figur 2) er markeret **’Locale’s writing systems’** og ikke **’Unicode (universal character set)’**.

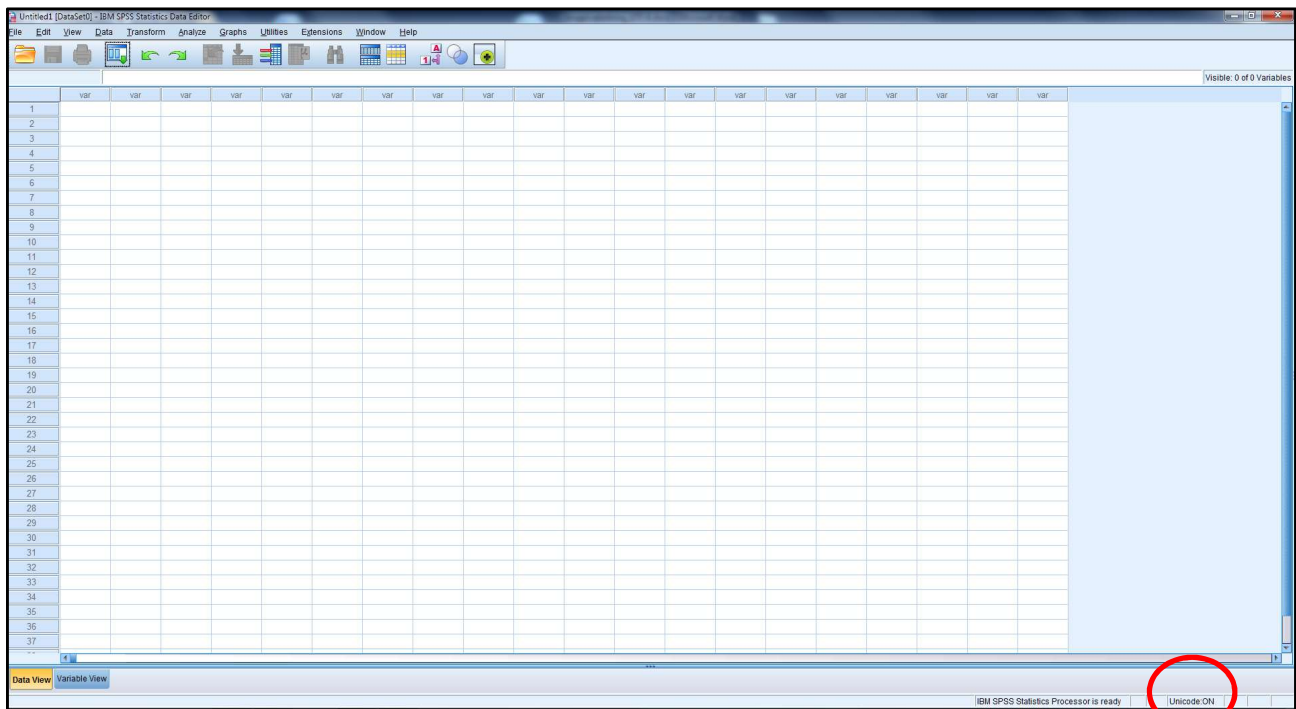
<sup>3</sup> <https://www.spss-tutorials.com/spss-unicode-mode/>



Figur 1: Valg af Edit > Option i SPSS



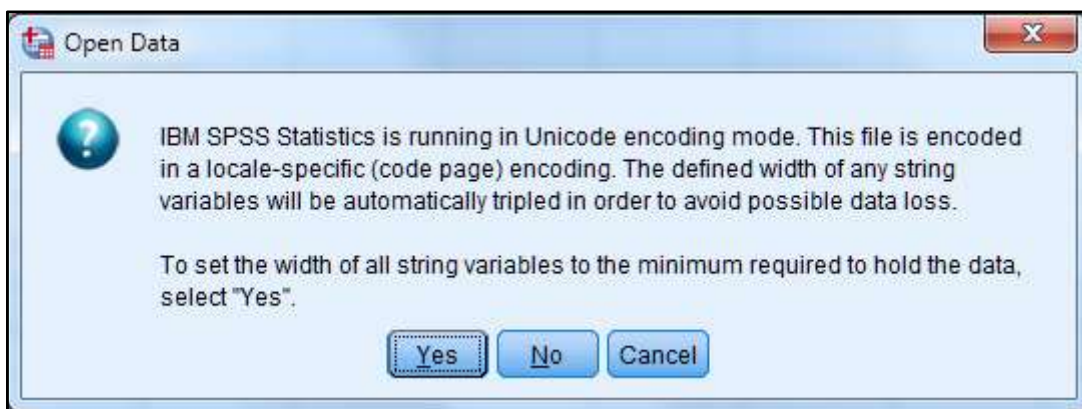
Figur 2: Fanebladet "Language" under Options i SPSS



Figur 3: Kontrol af 'Unicode: ON' i SPSS

### Skift SPSS-filers indkodning til UTF-8:

Hvis du i SPSS åbner en fil, som ikke er indkodet som Unicode, efter du har aktiveret Unicode i SPSS, fremkommer et pop-up vindue med følgende besked:



Figur 4: Pop-up vindue i SPSS

Du skal vælge 'Yes' for at optimere antallet af bytes. Filen er nu gemt som UTF-8 format.

## C. Stata – syntakser og procedurer til tegnsæt

### Indkodning/tegnset i forskellige Stata-versioner:

Som beskrevet af Stata<sup>4</sup>:

- Stata 13 og tidligere versioner bruger ASCII som standard til indkodning/tegnset.
- I Stata 14 og efterfølgende versioner er UTF-8 default indkodning/tegnset til datasæt, do-filer, ado-filer og 'help' filer.

<sup>4</sup> <https://www.stata.com/manuals/dunicodeencoding.pdf>



### Undersøg Stata-filens indkodning/tegnset

Til at analysere Stata-filens indkodning/tegnset i Stata, skal du køre følgende syntaks:

**Stata-Syntaks**

```
unicode analyze datasetname.dta
```

**Eksempel**

```
unicode analyze customerdata.dta
```

### Skift Stata-filens indkodning/tegnset til UTF-8

Stata kan også oversætte filer fra 'extended ASCII' indkodning til Unicode (UTF-8). Først skal du definere, hvilken indkodning/tegnset du ønsker at oversætte filen til. Dette kan gøres ved at køre følgende syntaks:

**Stata-Syntaks**

```
unicode encoding set encodingnavn
```

**Eksempel**

```
unicode encoding set unicode
```

Dernæst kan du bruge følgende syntaks til at transformere Stata-filen til Unicode:

**Stata-Syntaks**

```
unicode translate myfile.dta
```

**Eksempel**

```
Unicode translate customerdata.dta
```

Hvis du kender source-filens (srcencoding) indkodning/tegnset og den indkodning, du ønsker at transformere den til (dstencoding), kan du anvende følgende syntaks:

**Stata-Syntaks**

```
unicode convertfile srcfilename destfilename , options
```

**Eksempel**

```
unicode convertfile "C:\Temp\customerdata.txt" "  
C:\Temp\customerdata2.txt", srcencoding(ANSI1251)  
dstencoding(UNICODE)
```

#### 4. Hvordan kontrolleres tegnsæt i en tekstfil?

Når du har udtrukket data og metadata fra statistikfilen til .csv- og .txt-filer, der overholder afleveringspakkens format for datafiler og metadatafiler, skal du også kontrollere, at alle tegn kan læses korrekt og er UTF-8 tegn.

Den udtrukne datafil (fx table1.csv) kan indeholde forkerte tegn, hvis den statistikfil, udtrækket er lavet fra, ikke var i UTF-8 (Unicode) før udtræk, eller hvis den oprindeligt indeholdt ikke gyldige UTF-8 tegn.

Du kan inspicere .csv- og .txt-filer i afleveringspakken for forkerte tegn på følgende måde:

- Find placeringen af din afleveringspakke. Mappen kaldes fx FD.12345
- Find datafilen, du vil kontrollere for ikke gyldige UTF-8 tegn, fx table1.csv ved at klikke ned i mappestrukturen: FD.12345 > Data > table1 > table1.csv
- Højreklik på 12345.csv og vælg 'åben med' fra popup-listen. Vælg at åbne filen med en teksteditor, fx *Notesblok* eller *Notepad++*.

**OBS:** Du skal ikke dobbeltklikke på filen for at åbne den, da dette automatisk kan åbne den op i Excel. Excel gætter ofte på tegnsæt og formaterne af dine data og kan derfor indlæse dine data forkert).

- Kontrollér indholdet af datafilen ved at kigge efter tegn, der ser mærkelige ud. Søg efter tegn som æ, ø og å, da disse ofte vises forkert, hvis tegnsættet ikke er UTF-8
- Hvis du finder forkerte tegn, skal disse rettes i din originale datafil. Efter korrektionen skal et nyt udtræk laves (fx med programmet ASTA), og den udtrukne datafil skal igen visuelt testes for læsbarhed og ikke gyldige UTF-8 tegn.

#### 5. Hvordan aflæses UTF-8 hex-værdier i en tekstfil?

Hvis du ønsker at vide præcis, om et tegn er et gyldigt UTF-8 tegn, kan du undersøge det binære indhold af et tegn i tekstfilen. Til det formål skal du anvende en binær fieditor fx HxD-fil. Denne Hex-editor viser den numeriske standardrepræsentation af et tegn i et binært format i form af en hex-værdi.

Du kan således se tekstfilens normale indkodning ved siden af det binære format (se figur 5). Når du markerer et tegn i teksten til højre (fil i tekstformat), markeres den binære repræsentation af tegnet i venstre side.

Søg efter specialtegn fx æ, ø og å, som har følgende binære repræsentationer (UTF-8 hex-værdier):

Æ = E6

Ø = F8

Å = E5

File Edit Search View Analysis Extras Window ?

16 ANSI hex

File i binært format

Userinstruction\_UTF-8\_final.txt brugervejledning\_UTF-8.txt

File i tekst format

Offset (h)	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00000000	42	72	75	67	65	72	76	65	6A	6C	65	64	6E	69	6E	67
00000010	20	0D	0A	49	6E	64	6B	6F	64	6E	69	6E	67	20	6F	67
00000020	20	55	54	46	2D	38	0D	0A	48	76	61	64	20	65	72	20
00000030	55	54	46	2D	38	20	46	6F	72	6D	61	74	3F	0D	0A	0D
00000040	0A	4F	72	64	20	6F	67	20	73	E6	74	6E	69	6E	67	65
00000050	72	20	69	20	74	65	6B	73	74	20	65	72	20	6C	61	76
00000060	65	74	20	61	66	20	74	65	67	6E	20	73	6F	6D	20	65
00000070	72	20	64	65	20	62	6F	67	73	74	61	76	65	72	20	73
00000080	6F	6D	20	6D	61	6E	20	6B	61	6E	20	73	65	2E	20	54
00000090	65	67	6E	20	67	65	6D	6D	65	73	20	69	20	63	6F	6D
000000A0	70	75	74	65	72	65	6E	20	73	6F	6D	20	65	6E	20	65
000000B0	6C	6C	65	72	20	66	6C	65	72	65	20	62	79	74	65	73
000000C0	2C	20	73	6F	6D	20	65	72	20	72	65	70	72	E6	73	65
000000D0	6E	74	65	72	65	74	20	61	66	20	74	61	6C	2E	20	46
000000E0	6F	72	20	61	74	20	6C	61	76	65	20	65	6E	20	6F	76
000000F0	65	72	73	E6	74	74	65	6C	73	65	20	6D	65	6C	6C	65
00000100	6D	20	74	65	67	6E	20	6F	67	20	62	79	74	65	20	61
00000110	6E	76	65	6E	64	65	72	20	6D	61	6E	20	73	70	65	63
00000120	69	65	6C	6C	65	20	6B	6F	64	65	72	2E	20	49	6E	64
00000130	6B	6F	64	6E	69	6E	67	20	65	72	20	65	74	20	73	E6
00000140	74	20	61	66	20	6B	6F	64	65	72	20	69	6D	65	6C	6C
00000150	65	6D	20	62	79	74	65	73	20	69	20	63	6F	6D	70	75
00000160	74	65	72	65	6E	20	6F	67	20	6B	61	72	61	6B	74	65
00000170	72	20	6F	67	20	74	65	67	6E	2C	20	6F	67	20	73	6F
00000180	6D	20	68	6A	E6	6C	70	65	72	20	6D	65	64	20	6F	76
00000190	65	72	73	E6	74	74	65	6C	73	65	20	69	6D	65	6C	6C
000001A0	65	6D	20	64	65	20	74	6F	2E	20	45	74	20	74	65	67
000001B0	6E	20	6B	61	6E	20	72	65	70	72	E6	73	65	6E	74	65
000001C0	72	65	73	20	61	66	20	6D	65	72	65	20	65	6E	64	20
000001D0	E9	6E	20	62	79	74	65	2E	20	46	6F	72	20	65	6B	73
000001E0	65	6D	70	65	6C	20	65	72	20	74	65	67	6E	65	74	20
000001F0	22	61	22	20	6F	76	65	72	73	61	74	20	74	69	6C	20
00000200	62	79	74	65	73	20	22	39	37	22	2C	20	6D	65	6E	73
00000210	20	22	40	22	20	65	72	20	6F	76	65	72	73	61	74	20

Brugervejledning  
..Indkodning og  
UTF-8..Hvad er  
UTF-8 Format?...  
.Ord og sætninge  
r i tekst er lav  
et af tegn som e  
r de bogstaver s  
om man kan se. T  
egn gemmes i com  
puteren som en e  
ller flere bytes  
, som er repræse  
nteret af tal. F  
or at lave en ov  
ersættelse melle  
m tegn og byte a  
nvender man spec  
ielle koder. Ind  
kodning er et sæ  
t af koder imell  
em bytes i compu  
teren og karakte  
r og tegn, og so  
m hjælper med ov  
ersættelse imell  
em de to. Et teg  
n kan repræsent  
eres af mere end  
én byte. For eks  
empel er tegnet  
"a" oversat til  
bytes "97", mens  
"@" er oversat

Figur 5: Visning af tegn og binære værdier af tegn i en hexeditor

## 6. UTF-8 support i Rigsarkivet

Hvis du oplever problemer med at identificere tegnsæt i filer og ændre tegnsæt til UTF-8, kan du kontakte datamanageren for forskningsdata i Rigsarkivet på følgende e-mail: [mailbox@sa.dk](mailto:mailbox@sa.dk).