RIGSARKIVET

Vejledning til bilag 9 i bekendtgørelse om arkiveringsversioner

Uddrag fra den fulde vejledning til bekendtgørelsen

Bilag 9 - Afleveringspakke for visse typer af forskningsdata

Læsevejledning

Bemærk at denne vejledning til bilag 9 primært er et opslagsværk. Her kan finde svar på hvordan kravene skal forstås og tolkes. Dette kan være anvendeligt, hvis du under produktion og test af en afleveringspakke finder specifikke fejl, der skal rettes.

Der henvises til Rigsarkivets øvrige vejledninger, fx *"Quickguide - til produktion og test af en afleveringspakke med ASTA"* og *"Brugervejledning til ASTA"*. Quickguiden giver dig en lettere introduktion til kravene i bilag 9 og ASTA brugervejledningen anviser hvordan du skaber og tester en afleveringspakke af udtræk fra statistikfiler i formaterne SAS, Stata og SPSS med Rigsarkivets værktøj ASTA. Rigsarkivets har også udarbejdet en vejledning om, hvordan du kan skabe en afleveringspakke med data fra regneark og csv-filer.

Se også eksempelafleveringspakken AVID.SA.18005.1 med forskningsdata fra alle tre typer statistikfiler på Rigsarkivets hjemmeside www.sa.dk.

Formålet med bilag 9

Formålet med bilag 9 er at sikre bevaring og genanvendelse af forskningsdata skabt i statistikformat.

Statistikdata er den type forskningsdata som oftest afleveres til arkiv og Rigsarkivet ønsker med dette bilag 9, at styrke indsatsen for bevaring af forskningsdata skabt i statistikformat. Reglerne skal sikre, at de bevaringsværdige forskningsdata afleveres til arkivet således, at der i forhold til format, struktur og dokumentation sker en sikker bevaring, samtidig med, at mulighederne for genanvendelse understøttes.

Reglerne i bilag 9 sikrer, at statistikdata afleveres i en systemuafhængig standard, som indebærer, at data altid afleveres i en fast mappestruktur, i særlige formater for data og metadata og med specifikke datatyper. Med aflevering efter standarden i bilag 9 kan arkivet konvertere de afleverede data til et systemuafhængigt bevaringsformat, som betyder, at Rigsarkivet altid vil kunne konvertere de arkiverede data til nye formater, som kan læses, også når der ikke længere findes statistikprogrammer som fx SAS og STATA. Det betyder også, at forskere, som senere ønsker at få statistikdata udleveret til brug i nye videnskabelige undersøgelser, vil kunne for data udleveret i de nye statistikformater, der findes på markedet.

Reglerne i bilag 9 sikrer også, at der stilles specifikke krav til dokumentationen af de arkiverede data, således at forskere og andre arkivbrugere i fremtiden vil kunne forstå og genbruge data. Det gøres ved krav til metadata og kontekstdokumentation, der udgøres af følgende elementer:

- ***Kontekstdokumenter***, der på et overordnet plan beskriver den kontekst data er skabt i. Det kan bl.a. være en projektbeskrivelse, en metoderapport, et eksempel på et spørgeskema etc.
- ***Kontekstdokumentationsfil***, der giver oplysninger om de enkelte kontekstdokumenter fx dokumenttitel, dato, dokumentkategori og forfatter.
- ***Arkivbeskrivelsesfil***, der fx beskriver hvilke data det er der er afleveret, hvilke adgangskriterier der er for udlevering af data, hvilket formål data er indsamlet til og hvilken periode dataindsamlingen strækker sig over. Oplysningerne i arkivbeskrivelsesfilen vil blive anvendt af forskere og andre arkivbrugere til at fremsøge data i Rigsarkivets søgekatalog. Da arkivbeskrivelsesfilen ikke er omfattet af arkivlovens tilgængelighedsfrister kan den straks offentliggøres som søgemiddel til Rigsarkivets samling af forskningsdata.
- ***Metadatafil*** med informationer om statistikfilens indhold, herunder fx variabelnavne, variabelbeskrivelser, value labels, koder for manglende værdier og dataformater.

Bilag 9 anviser hvordan disse krav sikres ved at statistikdata afleveres i form af en særlig "afleveringspakke".

Hvad er en afleveringspakke?

Forskningsdata skabt eller behandlet i statistikprogrammer eller tilsvarende, skal afleveres som en afleveringspakke, der overholder kravene i bilag 9.

En afleveringspakke er et afleveringsformat, som indeholder de informationer, data og metadata, der er nødvendige for at kunne konvertere til bevaringsformat.

En afleveringspakke skal først og fremmest indeholde data- og metadataudtræk fra selve statistikfilen. Derudover skal den også indeholde en arkivbeskrivelsesfil og en kontekstdokumentationsfil samt de tilhørende kontekstdokumenter. Indhold og struktur af en afleveringspakke fremgår af figur 9.1.

Andre typer af forskningsdata skal afleveres til Rigsarkivet i de formater og strukturer og med den dokumentation, som fremgår af bekendtgørelsens bilag 1-8. Dette gælder fx forskningsdata i databaser eller kvalitative undersøgelser, fx lydfiler eller transskriptioner af interviews eller andre dokumentsamlinger. Forskningsdata i form af mindre dokumentsamlinger kan efter aftale med det modtagende arkiv også indgå som en del af afleveringen af institutionens ESDH-system (system med dokumenter), jf. bilag 3.

Programmer der kan bruges til at skabe afleveringspakken:
Rigsarkivet stiller en række programmer til rådighed, som kan anvendes i processen med at skabe afleveringspakken:

- Programmet ***Skab archiveIndex.exe***, der anvendes til at lave arkivbeskrivelsesfilen archiveIndex.xml, som skal indgå i afleveringspakken.
- Programmet ***Skab contextDocumentationIndex.exe***, der anvendes til at lave kontekstdokumentationsfilen contextDocumentationIndex.xml, som skal indgå i afleveringspakken.
- Programmet ***Asta (Aflevering af Statistikfiler Til Arkiv)***, der kan anvendes til først at skabe afleveringspakken med udtræk fra statistikformaterne SAS, SPSS, STATA og efterfølgende teste om afleveringspakken overholder krav specificeret i bilag 9.

Alle programmer kan tilgås fra Rigsarkivets hjemmeside www.sa.dk.

Hvis du har et regneark kan afleveringspakken eventuelt laves i hånden. Metadatafilen skal da udfyldes manuelt i en teksteditor. CSV-datafilen vil kunne udtrækkes fra regnearket. Se Rigsarkivets vejledning om udtræk fra regneark til en afleveringspakke på Rigsarkivets hjemmeside www.sa.dk.

Hvis myndigheden ikke selv kan producere en afleveringspakke kan en leverandør købes til at producere den.

Hvem har ansvaret for at lave afleveringspakken?

Det er myndighedens opgave at sørge for, at der skabes og afleveres en afleveringspakke til det modtagende arkiv af de bevaringsværdige data.

Arbejdet med at skabe afleveringspakken kan eventuelt udføres af en datamanager i tæt samarbejde med forskeren eller en anden person, som har indgående kendskab til data. En datamanager vil oftest have de relevante tekniske kundskaber og desuden have rettigheder til at behandle personfølsomme data.

Hvad sker der med data efter afleveringspakken er afleveret?

En afleveringspakke med statistikdata er ikke umiddelbart klar til den endelige arkivering. Afleveringspakken er kun et mellemstadie på vejen mod at konvertere de originale statistikdata til Rigsarkivets bevaringsformat som kaldes en arkiveringsversion og er defineret i bilag 1-8. Bevaringsformatet sikrer, at data kan langtidsbevares. Når det modtagende arkiv (fx Rigsarkivet) har modtaget afleveringspakken er det efterfølgende arkivets ansvar at udføre den sidste konvertering til det endelige bevaringsformat.

Udlevering af forskningsdata i statistikformat
Når forskere og andre arkivbrugere ønsker data udleveret igen, vil det modtagende arkiv sikre, at data fra statistikprogrammer afleveret efter bilag 9, kan udleveres til brugerne i statistikformat igen.[1]

Hvis data med personfølsomme oplysninger afleveres til det modtagende arkiv, vil de være beskyttet af arkivlovens bestemmelser. Det betyder, at data først er umiddelbart tilgængelige efter 75 år. En bruger kan dog søge om adgang til dem inden da. De første 20 år efter data er skabt, vil myndigheden blive spurgt om der kan gives adgang – og hvis ja, så skal Datatilsynet også høres inden brugeren får adgang til oplysninger om rent private forhold. Efter 20 år er det arkivet som kan give adgang til data – og igen skal Datatilsynet høres, inden data stilles til rådighed. Der kan som hovedregel fastsættes vilkår for brug af data, som der gives adgang til jf. Arkivlovens § 23, stk. 1.

Vejledning til punkterne i bekendtgørelsens bilag 9

9.A. Aflevering af forskningsdata

9.A.1 Reglerne i dette bilag gælder kun for data, som er skabt i forbindelse med forskning med anvendelse af videnskabelig metode, og som er skabt eller bearbejdet i statistikprogrammer eller tilsvarende.

9.A.2 Afleveringspakken konverteres til en arkiveringsversion, jf. reglerne i bilag 1-8, af det modtagende arkiv.

Reglerne i bilag 9 gælder kun data, som er indsamlet i forbindelse med en forskningsproces og kun hvis disse forskningsdata er skabt eller bearbejdet i statistikprogrammer og lignende værktøjer, der kan fortage statistiske analyser.

Forskningsdata, som er skabt eller bearbejdet i statistikprogrammer, kan fx være spørgeskemaundersøgelser eller målinger/registreringer, bl.a. biologiske, fysiologiske og neurologiske målinger, i formaterne SAS, STATA, SPSS og R.

Data anvendt til statistisk analyse kan også være skabt i andre tabulær formater end statistikformat, fx regneark eller CSV-filer. Disse skal også afleveres efter bilag 9.

Findes forskningsdata anvendt til statistisk analyse derimod i en relationel database, skal disse data afleveres i henhold til reglerne i bilag 1-8.

Det modtagende arkiv afgør om dine forskningsdata skal afleveres efter reglerne i bilag 9. Det fremgår af den afleveringsbestemmelse du modtager fra arkivet, hvis dine forskningsdata skal afleveres efter bilag 9.

9.B. Afleveringspakkens mappestruktur

9.B.1 I roden af filsystemet på afleveringsmediet, jf. bilag 7, skal der være placeret en mappe navngivet med afleveringspakkens navn. Afleveringspakkens navn består af præfikset »FD.« samt et unikt løbenummer for afleveringspakken.

9.B.2 Løbenummeret for afleveringspakken udleveres af Rigsarkivet.

En afleveringspakke har en fast mappestruktur. I roden af mappestrukturen placeres en mappe, som navngives med afleveringspakkens navn.

[1] For generelle regler vedr. tilgængelighedsfrister og adgang til ikke umiddelbart tilgængelige arkivalier se Arkivloven kapitel 6 og kapitel 7.

Afleveringspakkens navn består af præfikset »FD.« og et unikt løbenummer på minimum 5 cifre:
fx: FD.18999

Løbenummeret udleveres af det modtagende arkiv og fremgår af afleveringsbestemmelsen, som du modtager fra arkivet.

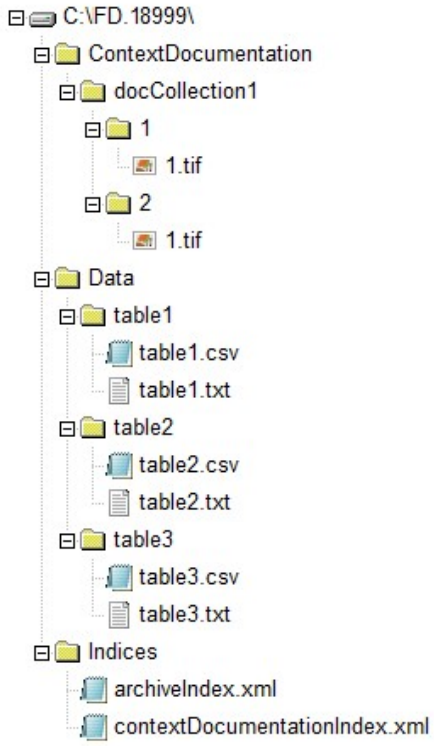
Løbenummeret i afleveringspakkens navn svarer til løbenummeret i det arkiveringsversions-ID, afleveringen får tildelt når afleveringspakken med statistikdata konverteres til en arkiveringsversion af det modtagende arkiv. Dvs. FD.18005 bliver til AVID.SA.18005.1.

9.B.3 Afleveringspakkens indhold fordeles i mapper, som angivet i figur 9.2.

9.B.4 Mapperne skal navngives som angivet i figur 9.2.

Der skal være tre undermapper under rodmappen, som navngives henholdsvis *ContextDocumentation*, *Data* og *Indices*.

Figur 9.1 Grafisk oversigt over elementer og struktur i en afleveringspakke



Mappen **ContextDocumentation** indeholder dokumenter konverteret til bevaringsformat fx tif, der beskriver de data, afleveringspakken indeholder. Fx hvordan data er indsamlet, metoderapport eller et spørgeskema.

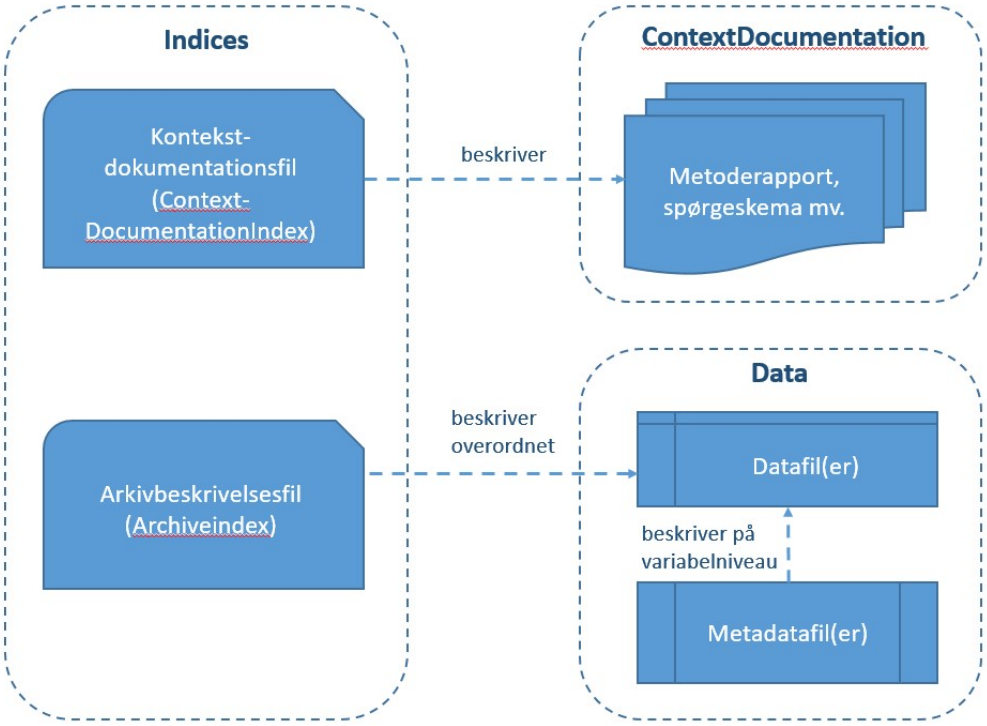
Mappen **Data** indeholder både en datafil og en metadatafil, der begge overholder kravene i bilag 9. data udtrukket fra de originale statistikfiler skal afleveres i en semikolonsepareret csv-fil (**table1.csv**). Metadata udtrukket fra statistikfilen, fx variabelbeskrivelser (variable labels), svarkategorier (value labels) og koder for manglende værdier (missing values), skal afleveres som en metadatafil i txt-format (**table1.txt**)

Mappen **Indices** indeholder to indeksfiler med metadata på et mere overordnet niveau. **archiveIndex.xml** filen indeholder fx oplysninger som navn på datasættet, der afleveres, navn på forsker der har indsamlet data, perioden data dækker, adgangsbegrænsninger til data osv. Filen **contextDocumentationIndex.xml** indeholder oplysninger om kontekstdokumenterne placeret i mappen ContextDocumentation, fx dokumentet titel, forfatter samt emnekategorisering af dokumentet.

Figur 9.2 Afleveringspakkens mapper

Navn på mappe	Indhold
ContextDocumentation	Kontekstdokumentation, jf. 9.D
Data	Datafil og metadatafil, jf. 9.E
Indices	Indeksfiler, jf. 9.C

Afleveringspakkens tre mapper og sammenhængen mellem dem



9.C. Mappen *Indices*

9.C.1 Mappen *Indices* skal indeholde følgende indeksfiler med oplysninger om afleveringspakken og dens indhold:

- **archiveIndex.xml**
- **contextDocumentationIndex.xml**

Disse to indeksfiler skabes af myndigheden og godkendes af det modtagende arkiv før afleveringspakken produceres. Rigsarkivet stiller indtastningsprogrammer til indeksfilerne til rådighed på www.sa.dk.

9.C.2 Indexfilerne skal overholde deres tilhørende skema, jf. bilag 8.

Når indeksfilerne udfyldes ved brug af de indtastningsprogrammer, som Rigsarkivet stiller til rådighed på www.sa.dk, valideres indtastningerne automatisk op imod de gældende skemaer for indeksfilerne, både når de udfyldes og gemmes.

9.C.3 Arkivbeskrivelsesfilen archiveIndex.xml skal overholde reglerne i bilag 6 punkt 6.A.

Arkivbeskrivelsesfilen *archiveIndex.xml* er en metadatafil, der indeholder generelle oplysninger om de data, der afleveres. Fx datasættets navn, start og slutdato for data, navn på afleverende myndighed, oplysninger om tilgængelighedsfrister osv.

Reglerne vedrørende arkivbeskrivelsesfilen er fremsat i bilag 6, punkt 6.A.1-3 og er beskrevet herunder:

1. Enhver afleveringspakke skal indeholde en arkivbeskrivelsesfil med angivelse af oplysninger iht. Figur 6.1, som vises herunder. Den viste figur er tilpasset aflevering af forskningsdata efter bilag 9.

Af figuren fremgår også hvilke oplysninger der er obligatoriske at udfylde samt øvrig vejledning til udfyldelse af oplysningerne.

2. Arkivbeskrivelsesfilen navngives archiveIndex.xml og skal overholde det tilhørende skema, jf. bilag 8.

3. Indholdet af arkivbeskrivelsesfilen fastlægges efter drøftelse med den afleverende myndighed og den modtagende arkiv.

Figur 6.1 Oplysninger i archiveIndex.xml (arkivbeskrivelsesfilen)

Vejledning i denne figur er tilpasset forskningsdata der afleveres efter bilag 9

Elementnavn	Betegnelse	Beskrivelse	Udfaldsrum	Fore-komst	Obliga-torisk	Vejledning
archiveInforma-tionPackageID	Afleveringspak-keløbenummer	**Entydigt ID som tildeles afleveringspakken af Rigsarkivet, som består af arkivkode og løbenummer.**	**Afleveringspa kkeløbenumm er, som defineret i denne bekendtgørelse**	**1**	**Ja**	Arkivkoden på det arkiv, som skal modtage afleveringspakken. Hvis det modtagende arkiv fx er Rigsarkivet er arkivkoden SA (= Statens Arkiv).
archiveInforma-tionPackageID-Previous	Tidligere afleveringer (fx kohorte)	**Entydigt Arkiveringsversion sID på tidligere aflevering fra samme forskningsprojekt.**	**Arkiveringsve rsionsID, som defineret i denne eller tidligere bekendtgørelse .**	**0-m**	**Nej**	ArkiveringsversionsID (fx AVID.SA.12345) på tidligere afleveringer af forskningsprojektet med kontinuerlig dataindsamling (fx kohorter og dataindsamling ifm overvågning af fænomener). ArkiveringsversionsID oplyses af Rigsarkivet.
archivePeriod-Start	Dataindsamling startdato	**Startdato for indsamlingen af de afleverede data.**	**År, år-måned, eller år-måned-dag.**	**1**	**Ja**	
archivePeriodEnd	Dataindsamling slutdato	**Slutdato for indsamling af de afleverede data.**	**År, år-måned, eller år-**	**1**	**Ja**	

archiveInforma-tionPacketType	Slutaflevering	**Angivelse af, om afleveringspakken er sidste aflevering af data, herunder også sidste dataaflevering fra et forskningsprojekt med kontinuerlig dataindsamling over en længere periode (evt. kohorte).**	**Boolsk værdi**	**1**	**Ja**	Har værdien true, hvis datasættet på tidspunktet for produktion af afleveringspakken er lukket for ny dataindsamling/opdateringer. Dette gælder fx den afsluttende dataindsamling fra en kohorteundersøgelse. Har ligeledes værdien true, hvis Rigsarkivet på tidspunktet for produktion af afleveringspakken har truffet bestemmelse om, at der ikke fremover skal afleveres data fra forskningsprojektet. Har værdien false i øvrige tilfælde.
creatorName	Forskningsin-stitutionen og Primærforsker	**Både forskningsinstituti onen og primærforsker, som har skabt datasættet.**	**Fri tekst**	**1-m**	**Ja**	Det aftales med det modtagende arkiv, hvilke dataskabere der skal anføres.
creationPeriod-Start	Dataindsamlings startdato for hver dataskaber	**Det angives hvornår data indsamlingen er påbegyndt for hver af de dataskabere/insti-tutioner, der har bidraget med indsamling af data til projektet.**	**År, år-måned, eller år-måned-dag**	**1-m**	**Ja**	Datering skal være så præcis som mulig, f.eks. 2008-06-01.
creationPeriodEnd	Dataindsamlings slutdato for hver dataskaber	**Det angives hvornår data indsamlingen er**	**År, år-måned, eller år-måned-dag**	**1-m**	**Ja**	Datering skal være så præcis som mulig, f.eks. 2010-12-31.

		afsluttet for hver af de dataskabere/institutioner, der har bidraget med indsamling af data til projektet.				
archiveType	Arkivtype	**Angivelse af om der er tale om en afsluttet aflevering eller en kontinuerlig aflevering.**	**Boolsk værdi**	1	Ja	Oplyses af det modtagende arkiv. Har værdien true, hvis der er tale om en afsluttet aflevering. Det kan f.eks. være: en tværsnitsundersøgelse af befolkningens holdninger til diverse emner eller en afsluttet naturvidenskabelig observationsundersøgelse. Har værdien false, hvis der er tale om en kontinuerlig aflevering, hvor dataindsamling er kontinuerlig over en længere periode og hvor der skal ske kontinuerlig aflevering af data i form af et øjebliksbillede, hvor de afleverede data også vil indgå i næste afleveringspakke. Det kan fx være en kohorteundersøgelse eller forløbsundersøgelser eller dataindsamling ifm overvågning af fænomener, hvor der løbende akkumuleres data.
systemName	Forskningsprojektets titel	**Den officielle danske titel på forskningsprojektet, hvor alle forkortelser er opløst.**	**Fritekst**	**1**	**Ja**	
alternativeName	Forskningsprojektets engelske titel	**Den officielle engelske titel på forskningsprojektet, hvor alle**	**Fritekst**	**0-m**	**Nej**	

		forkortelser er opløst				
systemPurpose	Forskningspro-jektets formål	**Beskrivelse af forskningsprojekte ts formål (evt. = forskningsprojekte ts abstract)**	**Fritekst**	**1**	**Ja**	Teksten skal redegøre for projektets centrale problemstillinger. Dokumenter med detaljerede oplysninger indgår i kontekstdokumentation.
systemContent	Beskrivelse af central population og centrale variable	**Beskrivelse af den centrale population og centrale variable i data**	**Fritekst**	**1**	**Ja**	Teksten skal besvare spørgsmålet om, hvem eller hvad (= den centrale population) der primært registreres hvilke data om (= de centrale variable). Fx et forskningsprojekt befolkningen i Danmark (= population) undersøges for holdning til Brexit (der vil være en vifte af centrale variable) eller et forskningsprojekt om størrelsen på ål i Lillebælt, hvor ål = population og centrale variable = længde og vægt.
regionNum	Regionsnumre	**Angivelse af, om der i systematisk form er registreret regionsnumre i datasættet**	**Boolsk værdi**	**1**	**Ja**	For hver af de angivne identifikatorer skal det oplyses, om den er registreret i datasættet. Har værdien true, hvis den pågældende identifikator er registreret systematisk. Har værdien false, hvis den pågældende identifikator ikke er registreret systematisk. Hvis der i datasættet findes andre fagspecifikke identifikatorer, der anvendes i flere datasæteller registre, f.eks. et geografisk referencesystem, Dansk branchekode, eller beskæftigelseskoder for baggrundsvariable, diagnosekoder eller medicinkoder kan det angives i forbindelse med oplysninger om central population og centrale variable (Beskrivelse af central population og centrale variable /systemContent).
komNum	Kommune-numre	**Angivelse af, om der i systematisk form er registreret kommunenumre i**	**Boolsk værdi**	**1**	**Ja**	

		datasættet			
cprNum	CPR-numre	**Angivelse af, om der i systematisk form er registreret CPR-numre i datasættet**	**Boolsk værdi**	**1**	**Ja**
cvrNum	CVR-numre	**Angivelse af, om der i systematisk form er registreret CVR-numre i datasættet**	**Boolsk værdi**	**1**	**Ja**
matrikNum	Matrikelnumre	**Angivelse af, om der i systematisk form er registreret matrikelnumre i datasættet**	**Boolsk værdi**	**1**	**Ja**
bbrNum	BBR-numre	**Angivelse af, om der i systematisk form er registreret BBR-numre i datasættet**	**Boolsk værdi**	**1**	**Ja**
whoSygKod	WHOs sygdomskoder	**Angivelse af, om der i systematisk form er registreret WHO-sygdomskoder i datasættet**	**Boolsk værdi**	**1**	**Ja**

sourceName	Evt. registre/andre datasæt der er udtrukket data fra	**Eventuelle registre der er udtrukket data fra/ andre datasæt der er udtrukket data fra (opslag, overførsel, samkøring osv.)**	**Fritekst**	**0-m**	**Nej**	Teksten skal besvare spørgsmålet om der er anvendt data udtrukket fra registre eller andre datasæt og fra hvilke. I kontekstdokumentationen kan indgå dokumenter med detaljerede oplysninger, f.eks. om hvilke data (kolonner) der er overført fra andre registre eller datasæt og hvordan overførsel er sket.
userName	Evt. andre projekter, der anvender dette projekts data	**Eventuelle andre forskningsprojekter, der anvender dette projekts data**	**Fritekst**	**0-m**	**Nej**	Teksten skal besvare spørgsmålet om, hvorvidt data anvendes i andre forskningsprojekter. I kontekstdokumentationen kan indgå dokumenter med detaljerede oplysninger, f.eks. om hvilke data (kolonner) der er anvendt i andre systemer, og hvordan data er anvendt.
predecessorName	Opfølgning på eksisterende forskningspro-jekt	**Eksisterende forskningsprojekt som dette forskningsprojekt følger op på.**	**Fritekst**	**0-m**	**Nej**	Hvis der er tale om en opfølgende forskningsundersøgelse til et tidligere udført forskningsprojekt, skal titlen på det tidligere udførte forskningsprojekt angives.
archiveApproval	Arkiv der godkender afleveringspak-ken (fx SA)	**Angivelse af, hvilket offentligt arkiv, der godkender afleveringspakken**	**Identifikation af det pågældende arkiv (2-4 tegn)**	**1**	**Ja**	Oplyses af det modtagende arkiv. Hvis det modtagende arkiv er Rigsarkivet skrives forkortelsen SA (= Statens Arkiv).
personalData-RestrictedInfo	Personhenfør-bare oplysninger	**Angivelse af, om der i data findes personfølsomme oplysninger i henholdt il databeskyttelsesfor ordningen**	**Boolsk værdi**	**1**	**Ja**	Har værdien true, hvis der i data findes følsomme personoplysninger (oplysninger om racemæssig eller etnisk baggrund, politisk, religiøs eller filosofisk overbevisning, fagforeningsmæssige tilhørsforhold, oplysninger om helbredsmæssige og seksuelle forhold samt oplysninger om strafbare forhold og væsentlige sociale problemer).

otherAccessTypeRestrictions	Særlige tilgængeligheds-bestemmelser	**Angivelse af, om der i data og den tilhørende dokumentation findes oplysninger, der kan betinge længere tilgængelighedsfrist i øvrigt**	**Boolsk værdi**	**1**	**Ja**	Har værdien true, hvis tilgængelighedsfristen er forlænget. Hvis den afleverende myndighed ønsker forlængelse af tilgængelighedsfristen, skal det i forbindelse med aflevering drøftes med det modtagende arkiv, jf. lovbekendtgørelse nr. 1035 af 21. august 2007 (Arkivloven), § 27.
archiveRestrictions	Beskrivelse af betingelser for adgang til data	**Angivelse af nærmere bestemmelser for adgang til materialet. Feltet kan anvendes efter det modtagende arkivs nærmere retningslinjer.**	**Fritekst**	**0-1**	**Nej**	Det modtagende arkiv træffer aftale med den afleverende myndighed om, hvorvidt og med hvilke oplysninger dette felt skal udfyldes.

9.C.4 Kontekstdokumentationsfilen contextDocumentationIndex.xml skal overholde reglerne i bilag 4 punkt 4.C.4.a samt 4.C.4.b.

Kontekstdokumentationsfilen *contextDocumentationIndex.xml* er en metadatafil, der indeholder oplysninger om kontekstdokumenterne, der vedlægges afleveringspakken jf. 9.D., fx dokumenttitel, dokumentdato, forfatter og dokumentkategori. Titel og dokumentkategori er obligatorisk at udfylde.
Reglerne vedrørende kontekstdokumentationsfilen er fremsat i bilag 4, punkt 4.C.4.a, 4.C.4.b samt 6.B. og er beskrevet herunder:

1. Enhver afleveringspakke skal indeholde kontekstdokumenter, som på et overordnet plan beskriver den kontekst data er skabt i. Det kan bl.a. bestå af en projektbeskrivelse, en metoderapport, et eksempel på et spørgeskema etc., jf. 6.B.1.

2. Det modtagende arkiv fastlægger efter drøftelse med den afleverende myndighed, hvilke kontekstdokumenter der skal afleveres, herunder hvilke punkter, specificeret i bekendtgørelsens figur 6.2, som ikke er relevante at dokumentere i den konkrete aflevering, jf. 6.B.2.

   Hvis der ikke er en tilstrækkelig dokumentation, skal den udarbejdes i forbindelse med produktion af afleveringspakken. I henhold til arkivlovgivningen er kontekstdokumentationen som udgangspunkt belagt med en tilgængelighedsfrist på 20 år. Det betyder, at såfremt en arkivbruger ønsker at gøre brug af materialet, før der er gået 20 år, skal den afleverende myndighed give sit samtykke, før materialet kan udleveres, jf. § 33 i arkivloven.

3. Kontekstdokumentationsfilen *contextDocumentationIndex.xml* skal indeholde et indeks over de kontekstdokumenter, som findes i afleveringspakkens kontekstdokumentation, jf. 4.C.4.a.

4. For hvert kontekstdokument skal der i kontekstdokumentationsfilen angives specifikke oplysninger[2], jf. 4.C.4.b og 6.B.3.b og vælges én eller flere dokumentkategorier[3], jf. 6.B.3.a og figur 6.2.

   Oplysninger i kontekstdokumentationsfilen:

   Aktuelt dokument
   - 'Dokumenttitel' (Obligatorisk): fx *Projektbeskrivelse*
   - 'Dokumentbeskrivelse': fx *Projektbeskrivelse af forskningsprojektet*
   - 'Dato': fx *2019-01-04*

   Forfattere
   - 'Forfatternavn': fx *Professor Erik Eriksen* (udlades ved Rigsarkivets Afleveringsbestemmelse)
   - 'Forfatterinstitution': fx *Københavns Universitet* (Ved dokumentet 'Rigsarkivet Afleveringsbestemmelse' skrives her 'Rigsarkivet').

   Dokumentkategorier
   - Der skal være valgt mindst én kategori for hvert dokument.

   Det er primært dokumentkategorierne 7.a-7.e i figur 6.2, der er relevante for forskningsdata. Dog kan andre dokumentkategorier være relevante, fx skal Rigsarkivets Afleveringsbestemmelse' tilknyttes dokumentkategorien *3.a 'Arkivets bestemmelse, herunder afleveringsbestemmelse'*.

[2] Oplysningerne fremgår også af figur 4.3 i bilag 4.
[3] Dokumentkategorierne fremgår også af figur 6.2 i bilag 6.

Figur 6.2 Dokumentkategorier i contextDocumentationIndex.xml (kontekstdokumentationsfilen)

Primært dokumentkategorierne 7.a-7.e er relevante for forskningsdata

1. Dokumentation vedrørende administrativ brug af it-systemet		
	Elementnavn	**Beskrivelse**
1.a	systemPurpose	It-systemets formål
1.b	systemRegulations	It-systemets lov- og regelgrundlag
1.c	systemContent	It-systemets indhold, population og særlige begreber
1.d	systemAdministrativeFunctions	It-systemets administrative funktioner
1.e	systemPresentationStructure	It-systemets præsentationsstruktur
1.f	systemDataProvision	Tilvejebringelse af data
1.g	systemDataTransfer	Videregivelse af data
1.h	systemPreviousSubsequentFunctions	Data og funktioner fælles med forgænger- og efterfølgersystemer
1.i	systemAgencyQualityControl	Myndighedens egen kvalitetskontrol
1.j	systemPublication	Publikation af og om data
1.k	systemInformationOther	Andet
1.l	systemTaxonomy	Registreringssystematik
1.m	systemInstruction	Instruks for anvendelse af systemet
2. Dokumentation vedrørende it-systemets tekniske udformning, drift og udvikling		
	Elementnavn	**Beskrivelse**
2.a	operationalSystemInformation	Driftsversionens opbygning
2.b	operationalSystemConvertedInformation	Konvertering hos myndigheden
2.c	operationalSystemSOA	Dokumentation af sammensætning af data og eventuelle dokumenter fra flere forskellige it-systemer i en serviceorienteret arkitektur
2.d	operationalSystemInformationOther	Andet
3. Dokumentation vedrørende arkivskabers aflevering af data		
	Elementnavn	**Beskrivelse**
3.a	archivalProvisions	Arkivets bestemmelser herunder afleveringsbestemmelse
3.b	archivalTransformationInformation	Dokumentation af konvertering fra driftsversion til arkiveringsversion
3.c	archivalInformationOther	Andet
4. Dokumentation vedrørende arkivets modtagelse af data (udfyldes af modtagende arkiv)		
	Elementnavn	**Beskrivelse**
4.a	archivistNotes	Arkivarnoter
4.b	archivalTestNotes	Testnoter
4.c	archivalInformationOther	Andet
5. Dokumentation vedrørende arkivets bevaring af arkiveringsversionen (udfyldes af det modtagende arkiv)		
	Elementnavn	**Beskrivelse**
5.a	archivalMigrationInformation	Konvertering hos arkivet

5.b	archivalInformationOther	Andet
6. Anden dokumentation		
	Elementnavn	**Beskrivelse**
6.a	informationOther	Andet
7. Dokumentation af forskningsdata		
7.a	researchProjectDescription	Projektbeskrivelse gældende for de afleverede data
7.b	researchQuestionnaire	Spørgeskema, interviewguide og/eller registreringsskema anvendt til at indsamle og analysere de afleverede data
7.c	researchProtocol	Protokoller og metoderapporter
7.d	researchPublication	Publikationer, som er udgivet på basis af de afleverede data
7.e	researchInformationOther	Andet

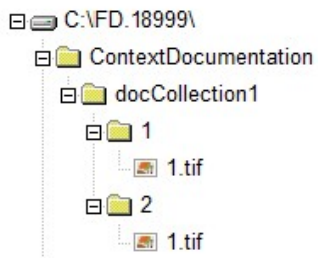
9.D. Mappen *ContextDocumentation*

9.D.1 Mappen *ContextDocumentation* skal indeholde kontekstdokumentation, jf. 4.E.

Kontekstdokumentationsfilen er beskrevet under 9.C.4. Her beskrives udelukkende regler for placering og navngivning af kontekstdokumenterne, jf. 4.E.

Selve kontekstdokumenterne skal placeres i undermappen *ContextDocumentation* i afleveringsmappen.

Illustration af mappen *ContextDocumentation* og dens indhold i afleveringspakken



Reglerne vedrørende placering af kontekstdokumenterne i afleveringspakken er følgende[4]:

Regler til Dokumentsamlingsmappen:
Mappen *ContextDocumentation* skal indeholde en undermappe kaldet en dokumentsamlingsmappe. Dokumentsamlingsmappen navngives "docCollection[fortløbende nummer]", begyndende med 1.

En dokumentsamlingsmappe med kontekstdokumentation må indeholde op til 10.000 dokumentmapper, jf. 4.E.2. Der vil derfor sjældent være behov for flere dokumentsamlingsmapper, men opstår behovet skal næste dokumentsamlingsmappe navngives med det næste fortløbende nummer (*docCollection2*), således at navnet på hver dokumentsamlingsmappe er unikt inden for *ContextDocumentation* mappen, jf. 4.E.3.

Regler til dokumentmapperne:
I dokumentsamlingsmappen *docCollection1* placeres en eller flere undermapper, kaldet dokumentmapper, hvor der placeres én mappe til hvert kontekstdokument, 4.E.2.

Hvert kontekstdokument er i kontekstdokumentationsfilen tildelt et dokumentID. Når kontekstdokumenterne

[4] Reglerne for placeringen af kontekstdokumenterne i afleveringspakken er desuden angivet i bilag 4, punkt 4.E.

placeres i afleveringspakkens *docCollection1* mappe, skal de placeres i overensstemmelse med rækkefølgen angivet i kontekstdokumentationsfilen, fx hvis kontekstdokumentet *"Afleveringsbestemmelse"* har dokumentID *1* skal det placeres i dokumentmappen navngivet "1", jf. 4.E.4 og 4.E.5.

Regler for kontekstdokumenterne:
En dokumentmappe skal indeholde ét dokument, som består af én eller flere filer af samme format, jf. 4.E.5.

Et dokuments fil (eller filer) navngives fortløbende med et nummer, begyndende med 1 samt formatets ekstension, jf. 4.G.6., fx 1.tif. Hvis dokumentet består af flere filer navgives de *1.tif*, *2.tif*, *3.tif* osv.

Dokumenterne skal afleveres i ét af de i arkiveringsversionen tilladte dokumentformater, jf. 6.B.4 samt 5.E.-5.F. Følgende dokumentformater er tilladt:

- .tif (TIFF) og .jp2 (JPEG-2000) til dokumenter og billeder
- .mp3 (MP3) og .wav (WAVE) til lydfiler
- .mpg (MPEG-2 og MPEG-4) til video

Det er vigtigt, at dokumenterne konverteres efter reglerne efter bekendtgørelsens bilag 5.E. og 5.F., som er beskrevet herunder. Se også Rigsarkivets vejledning i konvertering af dokumenter til TIFF på Rigsarkivets hjemmeside www.sa.dk.

5.E. Digitale dokumenter

5.E.1 Et digitalt dokument, jf. dog 5.F. og 5.G., skal lagres i et af følgende formater:

- **det grafiske bitmapformat TIFF, version 6.0 baseline.**
- **JPEG-2000 efter standarden ISO/IEC 15444-1:2004. Information technology - JPEG 2000 image coding system - Part 1: Core coding system.**

5.E.1.a Det er tilladt at benytte begge formater inden for samme arkiveringsversion.

5.E.2 Dokumenter i TIFF skal komprimeres efter følgende kompressionsregler:

5.E.2.a Sort/hvide dokumenter skal komprimeres med CCITT/TSS Grp3, Grp4, PackBit eller LZW.

5.E.2.b Dokumenter med gråtoner eller farver skal komprimeres med PackBit eller LZW.

5.E.3 Dokumenter i TIFF RGB må udelukkende benytte følgende bitdybder: 1, 2, 4, 8, 24 og 32.

5.E.3.a Dokumenter i TIFF RGB må maksimalt benytte 3 farvekanaler med en maksimal bitdybde på 24 bit (8x8x8 bit) evt. suppleret med maksimalt en alfakanal (8 bit), således at den samlede bitdybde for en billedfil ikke kan overstige 32 bit.

5.E.4 Dokumenter i TIFF CMYK må udelukkende benytte følgende bitdybder: 1, 2, 4, 8, 32 og 40.

5.E.4.a Dokumenter i TIFF CMYK må maksimalt benytte 4 farvekanaler med en maksimal bitdybde på 32 bit (8x8x8x8 bit) evt. suppleret med maksimalt én alfakanal (8 bit), således at den samlede bitdybde for en billedfil ikke kan overstige 40 bit.

5.E.5 TIFF dokumenters anvendelse af XResolution og YResolution (TIFF Tag 282 og 283), skal ske på en sådan måde, at forholdet mellem bredde og højde ved anvendelse af disse værdier (skalering) svarer til sidernes dimensioner i det oprindelige dokument.

Der kan anvendes to forskellige formater til arkivering af dokumenter, TIFF (version 6.0 baseline) og JPEG2000:

- TIFF 1 bit med kompression Grp4 samt LZW er velegnet til sort/hvide dokumenter
- TIFF 4-24 bit med kompression LZW er velegnet til almindelige dokumenter, som indeholder tegninger og fotos i varierede størrelse og antal
- TIFF 24 bit er velegnet til tegninger, kort, fotos etc., hvor den eksakte bevaring af dokumentet er af betydning. Som eksempel kan nævnes satellitfoto af afgrøder, hvor tolkningen af farven på de enkelte pixels kan være af betydning
- JPEG-2000 er overordentlig velegnet til store tegninger, kort, fotos etc., såfremt farven på de enkelte pixels er af mindre betydning. Som eksempel kan nævnes indscannede matrikelkort eller affotograferede skibstegninger

Kvaliteten af den grafiske kopi skal svare til den kvalitet, som myndigheden selv anvendte, hvilket for digitalt fødte dokumenter svarer til en passende udskriftkvalitet (f.eks. 300 DPI). For indscannet materiale skal kvaliteten modsvare kvaliteten på det materiale, som danner grundlag for kopien.

Valg af kompressionstype og bitdybde (antal mulige farver) må gerne variere fra side til side inden for samme dokument.

Myndigheden skal sikre sig, at dokumenterne også efter konvertering til arkivformatet er læsbare f.eks. ved at kontrollere at kommaer fremstår korrekt i dokumentet, og ikke ligner punktummer. Hvis det er tilfældet, er opløsningen for lav, og dokumentet skal lagres med en højere opløsning.

Rigsarkivet anbefaler anvendelsen af Grp4 og LZW-kompression eftersom dokumenter komprimeret med disse algoritmer fylder mindre end hvis der benyttes Grp3 og PackBit.

5.F. Lyd og video

5.F.1 Lydfiler skal lagres efter standarden MP3 DS/EN ISO/IEC 11172-3.

5.F.2 Det modtagende arkiv kan tillade, at lydfiler afleveres i formatet WAVE LPCM som specificeret i *Multimedia Programming Interface and Data Specifications 1.0. IBM Corporation and Microsoft Corporation, August 1991*. Dog begrænset til bitdybder, der er hele multipla af 8.

5.F.3 Videofiler skal lagres efter en af følgende standarder:

- **MPEG-2 DS/EN ISO/IEC 13818-2. Eventuel lyd indkodes som MP3, som specificeret i ISO/IEC 13818-3.**
- **MPEG-4 AVC DS/EN ISO/IEC 14496-10 (ITU-T H.264). Video indkodes som specificeret i ISO/IEC 14496-10. Eventuel lyd indkodes som AAC, som specificeret i ISO/IEC 14496-3. Video og lyd indpakkes i MPEG-4 formatet som defineret i ISO/IEC 14496-14.**

Kontekstdokumenter i form af lyd og video skal konverteres til de anviste formater på en sådan måde, at den oprindelige kvalitet i størst muligt omfang beholdes.

Der kan anvendes to forskellige lydformater ved aflevering til offentligt arkiv, MP3 og WAVE. Det vil normalt være MP3, som skal anvendes, men hvis myndigheden selv har anvendt WAVE kan det modtagende arkiv tillade, at WAVE anvendes som arkiveringsformat. Det samme gælder for lydoptagelser, hvor en meget præcis lydgengivelse er vigtig. Et eksempel på dette er optagelser, hvor baggrundslyde har betydning for forståelsen, og konteksten for den pågældende optagelse.

9.E. Mappen Data

9.E.1 En afleveringspakke skal indeholde et eller flere datasæt. Hvert datasæt skal bestå af en datafil og en tilhørende metadatafil.

Hvis aflevering af forskningsdata kun består af én statistikfil svarer det til ét datasæt. Hvis et datasæt består af flere statistikfiler, vil afleveringspakken bestå af flere datasæt med et datasæt for hver statistikfil.

Datafilen indeholder data fra statistikfilen, regnearket, csv-filen eller tilsvarende.

Metadatafilen indeholder metadata fra selve statistikfilen eller regnearket, fx variabelnavne, variabelbeskrivelser og value labels. Derudover indeholder den informationerne SYSTEMNAVN, DATAFILNAVN og DATAFILBESKRIVELSE samt eventuelt NØGLEVARIABEL og REFERENCE som indtastes af arkivskaberen, jf. figur 9.4.

Hvis der findes vigtige informationer i statistikfilen, som ikke kan udtrækkes til metadatafilen i afleveringspakken, fx noter om hvilken kilde en variabel stammer fra i notefelter til variable, kan disse informationer placeres i et kontekstdokument i afleveringspakken.

9.E.2 Datafil og metadatafil placeres i mappen Data i en undermappe, der navngives efter bilag 4 punkt 4.D.2.a samt 4.D.2.b.

9.E.2.a Datafil navngives med undermappens navn efterfulgt af ekstensionen ».csv«.

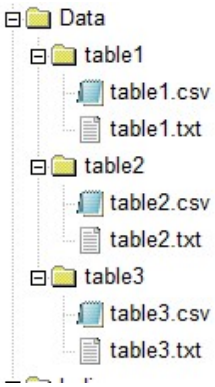
9.E.2.b Metadatafil navngives med undermappens navn efterfulgt af ekstensionen ».txt«.

Hver datafil og dennes tilhørende metadatafil placeres i en undermappe i mappen *Data*.

Undermappen navngives tabel [fortløbende nummer], fx table1. Den fortløbende nummerering begynder med 1. Foranstillede nuller må ikke anvendes, jf. 4.D.2.a samt 4.D.2.b.

Datafilen navngives med undermappens navn, efterfulgt af ekstensionen ».csv«, fx: tabel 1.csv.
Metadatafil navngives med undermappens navn, efterfulgt af ekstensionen ».txt«, fx: tabel 1.txt.

Illustration af mappen *Data* og dens indhold i afleveringspakken



9.F. Tekstformat

9.F.1 Tegnsættet for henholdsvis datafil og metadatafil skal være indkodet som UTF-8, som angivet i ISO/IEC 10646:2003 Annex D og som beskrevet i *The Unicode Standard 5.1*, kapitel 3, og skal i øvrigt overholde bestemmelserne i bilag 5 punkt 5.D.1.b – 5.D.1.d.

Tegnsættet for både data- og metadatafilen skal være indkodet som UTF-8. I Rigsarkivets særskilte vejledning om UTF-8 kan du læse mere om 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 som NotePad++.

Rigsarkivets udtræksværktøj ASTA, der kan anvendes til at skabe og teste afleveringspakken, tester ikke automatisk for om tegnsættet i data- og metadatafilen er indkodet som UTF-8. Hvis dette ikke er tilfældet skal det opdages i den manuelle visuelle test af afleveringspakken, fx ved at teste om nationale karakterer som æ, ø og å vises korrekt.

9.G. Datafil

9.G.1 Datafil afleveres som en semikolonsepareret tekstfil i henhold til RCF 4180 Common Format and MIME Type for CSV Files, der overholder syntaksen beskrevet som EBNF i figur 9.12.

Datafilen skal afleveres som en semikolonsepareret tekstfil, der overholder syntaksen beskrevet som EBNF i figur 9.12. EBNF'en overholder men begrænser også csv-standarden RCF4180.

Eksempel på datafil (table1.csv) der overholder EBNF i figur 9.12

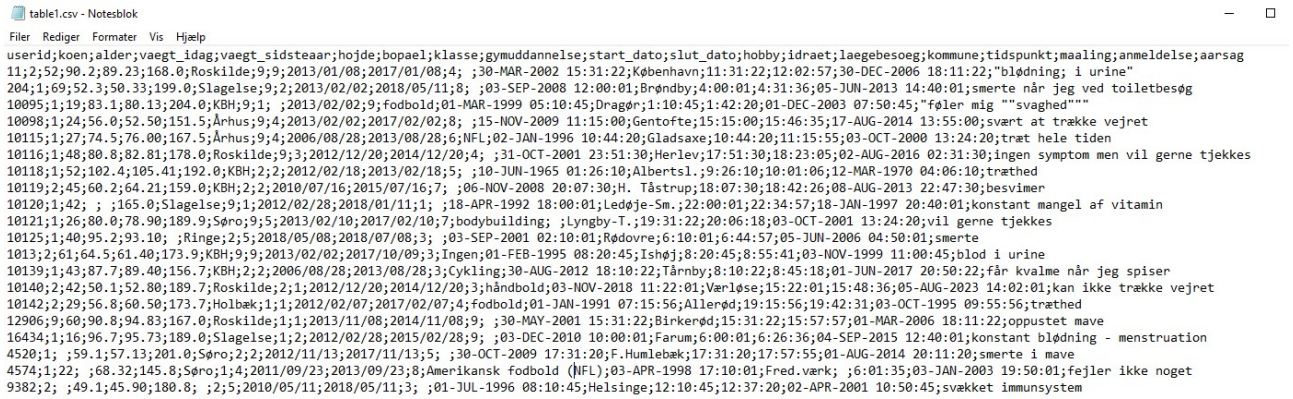


table1.csv - Notesblok

Filer Rediger Formater Vis Hjælp

```
userid;koen;alder;vaegt_idag;vaegt_sidsteaar;hojde;bopael;klasse;gymuddannelse;start_dato;slut_dato;hobby;idraet;laegebesoeg;kommune;tidspunkt;maaling;anmeldelse;aarsag
11;2;52;90.2;89.23;168.0;Roskilde;9;9;2013/01/08;2017/01/08;4; ;30-MAR-2002 15:31:22;København;11:31:22;12:02:57;30-DEC-2006 18:11:22;"blødning; i urine"
204;1;69;52.3;50.33;199.0;Slagelse;9;2;2013/02/02;2018/05/11;8; ;03-SEP-2008 12:00:01;Brøndby;4:00:01;4:31:36;05-JUN-2013 14:40:01;smerte når jeg ved toiletbesøg
10095;1;19;83.1;80.13;204.0;KBH;9;1; ;2013/02/02;9;fodbold;01-MAR-1999 05:10:45;Dragør;1:10:45;1:42:20;01-DEC-2003 07:50:45;"føler mig ""svaghed"""
10098;1;24;56.0;52.50;151.5;Århus;9;4;2013/02/02;2017/02/02;8; ;15-NOV-2009 11:15:00;Gentofte;15:15:00;15:46:35;17-AUG-2014 13:55:00;svært at trække vejret
10115;1;27;74.5;76.00;167.5;Århus;9;4;2006/08/28;2013/08/28;6;NFL;02-JAN-1996 10:44:20;Gladsaxe;10:44:20;11:15:55;03-OCT-2000 13:24:20;træt hele tiden
10116;1;48;80.8;82.81;178.0;Roskilde;9;3;2012/12/20;2014/12/20;4; ;31-OCT-2001 23:51:30;Herlev;17:51:30;18:23:05;02-AUG-2016 02:31:30;ingen symptom men vil gerne tjekkes
10118;1;52;102.4;105.41;192.0;KBH;2;2;2012/02/18;2013/02/18;5; ;10-JUN-1965 01:26:10;Albertsl.;9:26:10;10:01:06;12-MAR-1970 04:06:10;træthed
10119;2;45;60.2;64.21;159.0;KBH;2;2;2010/07/16;2015/07/16;7; ;06-NOV-2008 20:07:30;H. Tåstrup;18:07:30;18:42:26;08-AUG-2013 22:47:30;besvimer
10120;1;42; ; ;165.0;Slagelse;9;1;2012/02/28;2018/01/11;1; ;18-APR-1992 18:00:01;Ledøje-Sm.;22:00:01;22:34:57;18-JAN-1997 20:40:01;konstant mangel af vitamin
10121;1;26;80.0;78.90;189.9;Søro;9;5;2013/02/10;2017/02/10;7;bodybuilding; ;Lyngby-T.;19:31:22;20:06:18;03-OCT-2001 13:24:20;vil gerne tjekkes
10125;1;40;95.2;93.10; ;Ringe;2;5;2018/05/08;2018/07/08;3; ;03-SEP-2001 02:10:01;Rødovre;6:10:01;6:44:57;05-JUN-2006 04:50:01;smerte
1013;2;61;64.5;61.40;173.9;KBH;9;9;2013/02/02;2017/10/09;3;Ingen;01-FEB-1995 08:20:45;Ishøj;8:20:45;8:55:41;03-NOV-1999 11:00:45;blod i urine
10139;1;43;87.7;89.40;156.7;KBH;2;2;2006/08/28;2013/08/28;3;Cykling;30-AUG-2012 18:10:22;Tårnby;8:10:22;8:45:18;01-JUN-2017 20:50:22;får kvalme når jeg spiser
10140;2;42;50.1;52.80;189.7;Roskilde;2;1;2012/12/20;2014/12/20;3;håndbold;03-NOV-2018 11:22:01;Værløse;15:22:01;15:48:36;05-AUG-2023 14:02:01;kan ikke trække vejret
10142;2;29;56.8;60.50;173.7;Holbæk;1;1;2012/02/07;2017/02/07;4;fodbold;01-JAN-1991 07:15:56;Allerød;19:15:56;19:42:31;03-OCT-1995 09:55:56;træthed
12906;9;60;90.8;94.83;167.0;Roskilde;1;1;2013/11/08;2014/11/08;9; ;30-MAY-2001 15:31:22;Birkerød;15:31:22;15:57:57;01-MAR-2006 18:11:22;oppustet mave
16434;1;16;96.7;95.73;189.0;Slagelse;1;2;2012/02/28;2015/02/28;9; ;03-DEC-2010 10:00:01;Farum;6:00:01;6:26:36;04-SEP-2015 12:40:01;konstant blødning - menstruation
4520;1; ;59.1;57.13;201.0;Søro;2;2;2012/11/13;2017/11/13;5; ;30-OCT-2009 17:31:20;F.Humlebæk;17:31:20;17:57:55;01-AUG-2014 20:11:20;smerte i mave
4574;1;22; ;68.32;145.8;Søro;1;4;2011/09/23;2013/09/23;8;Amerikansk fodbold (NFL);03-APR-1998 17:10:01;Fred.værk; ;6:01:35;03-JAN-2003 19:50:01;fejler ikke noget
9382;2; ;49.1;45.90;180.8; ;2;5;2010/05/11;2018/05/11;3; ;01-JUL-1996 08:10:45;Helsinge;12:10:45;12:37:20;02-APR-2001 10:50:45;svækket immunsystem
```

9.G.1.a Første linje i datafilen skal altid angive alle variabelnavne, angivet i samme rækkefølge som i metadatafilen.

Første linje i datafilen skal angive variabelnavne på alle variable i den oprindelige statistikfil eller regneark, som dataudtrækket er lavet fra. Rækkefølgen af variabelnavne i datafilen skal stemme overens med rækkefølgen af variabelnavne angivet i metadatafilen under etiketten VARIABEL.

Variabelnavne skal i datafilen adskilles af semikolon, men der anvendes ikke semikolon efter sidste variabelnavn.

Variabelnavne skal overholde SQL99 standarden (ISO/IEC 9075:1999 - Database Language SQL) for angivelse af tabel- og kolonnenavne. Dvs. at variabelnavn må aldrig begynde med et tal, men må godt være en blanding af bogstaver, tal samt underscore på max 128 tegn. Hvis variabelnavnet er et reserveret ord i SQL:1999, skal det enten omdøbes eller omkranses af dobbeltapostrof.

Reserverede ord i SQL 99 standarden

<reserved word> ::=

```
ABSOLUTE | ACTION | ADD | ADMIN | AFTER | AGGREGATE | ALIAS | ALL | ALLOCATE |
ALTER | AND | ANY | ARE | ARRAY | AS | ASC | ASSERTION | AT | AUTHORIZATION |
BEFORE | BEGIN | BINARY | BIT | BLOB | BOOLEAN | BOTH | BREADTH | BY
| CALL | CASCADE | CASCADED | CASE | CAST | CATALOG | CHAR | CHARACTER | CHECK |
CLASS | CLOB | CLOSE | COLLATE | COLLATION | COLUMN | COMMIT | COMPLETION |
CONNECT | CONNECTION | CONSTRAINT | CONSTRAINTS || CONSTRUCTOR | CONTINUE |
CORRESPONDING | CREATE | CROSS | CUBE | CURRENT | CURRENT_DATE | CURRENT_PATH |
CURRENT_ROLE | CURRENT_TIME | CURRENT_TIMESTAMP | CURRENT_USER | CURSOR | CYCLE
| DATA | DATE | DAY | DEALLOCATE | DEC | DECIMAL | DECLARE | DEFAULT |
DEFERRABLE | DEFERRED | DELETE | DEPTH | DEREF | DESC | DESCRIBE | DESCRIPTOR |
DESTROY | DESTRUCTOR | DETERMINISTIC | DICTIONARY | DIAGNOSTICS | DISCONNECT |
DISTINCT | DOMAIN | DOUBLE | DROP | DYNAMIC | EACH | ELSE | END | END-EXEC |
EQUALS | ESCAPE | EVERY | EXCEPT | EXCEPTION | EXEC | EXECUTE | EXTERNAL | FALSE
```

| FETCH | FIRST | FLOAT | FOR | FOREIGN | FOUND | FROM | FREE | FULL | FUNCTION | GENERAL | GET | GLOBAL | GO | GOTO | GRANT | GROUP | GROUPING | HAVING | HOST | HOUR | IDENTITY | IGNORE | IMMEDIATE | IN | INDICATOR | INITIALIZE | INITIALLY | INNER | INOUT | INPUT | INSERT | INT | INTEGER | INTERSECT | INTERVAL | INTO | IS | ISOLATION | ITERATE | JOIN | KEY | LANGUAGE | LARGE | LAST | LATERAL | LEADING | LEFT | LESS | LEVEL | LIKE | LIMIT | LOCAL | LOCALTIME | LOCALTIMESTAMP | LOCATOR | MAP | MATCH | MINUTE | MODIFIES | MODIFY | MODULE | MONTH | NAMES | NATIONAL | NATURAL | NCHAR | NCLOB | NEW | NEXT | NO | NONE | NOT | NULL | NUMERIC | OBJECT | OF | OFF | OLD | ON | ONLY | OPEN | OPERATION | OPTION | OR | ORDER | ORDINALITY | OUT | OUTER | OUTPUT | PAD | PARAMETER | PARAMETERS | PARTIAL | PATH | POSTFIX | PRECISION | PREFIX | PREORDER | PREPARE | PRESERVE | PRIMARY | PRIOR | PRIVILEGES | PROCEDURE | PUBLIC | READ | READS | REAL | RECURSIVE | REF | REFERENCES | REFERENCING | RELATIVE | RESTRICT | RESULT | RETURN | RETURNS | REVOKE | RIGHT | ROLE | ROLLBACK | ROLLUP | ROUTINE | ROW | ROWS | SAVEPOINT | SCHEMA | SCROLL | SCOPE | SEARCH | SECOND | SECTION | SELECT | SEQUENCE | SESSION | SESSION_USER | SET | SETS | SIZE | SMALLINT | SOME| SPACE | SPECIFIC | SPECIFICTYPE | SQL | SQLEXCEPTION | SQLSTATE | SQLWARNING | START | STATE | STATEMENT | STATIC | STRUCTURE | SYSTEM_USER | TABLE | TEMPORARY | TERMINATE | THAN | THEN | TIME | TIMESTAMP | TIMEZONE_HOUR | TIMEZONE_MINUTE | TO | TRAILING | TRANSACTION | TRANSLATION | TREAT | TRIGGER | TRUE | UNDER | UNION | UNIQUE | UNKNOWN | UNNEST | UPDATE | USAGE | USER | USING | VALUE | VALUES | VARCHAR | VARIABLE | VARYING | VIEW | WHEN | WHENEVER | WHERE | WITH | WITHOUT | WORK | WRITE | YEAR | ZONE

9.G.1.b Hvis semikolon indgår i en værdi for en variabel, skal hele værdien omsluttes med dobbelt apostrof »"« (U+0022). Hvis dobbelt apostrof indgår i en værdi for en variabel, skal dobbelt apostroffen foranstilles med en dobbelt apostrof, og hele værdien omsluttes med dobbelt apostroffor.

Hvis semikolon indgår i en værdi for en variabel i et tekstfelt, skal hele teksten omsluttes med dobbelt apostrof (dobbelte anførselstegn også kaldet gåseøjne).

Eksempel på udtræk af en variabelværdi der indeholder semikolon

Variabelværdien *Antal hjorte; 6* skal i datafilen angives som *"Antal hjorte; 6"*

Hvis dobbeltapostrof indgår i en værdi for en variabel i et tekstfelt, skal dobbeltapostroffen foranstilles med en dobbelt apostrof samtidig med at hele teksten omsluttes med dobbelt apostrof.

Dobbeltapostroffen der anvendes skal have unicode-værdien U+0022. Hvis der anvendes andre former for dobbeltapostrof (fx højrestillet dobbelt apostrof med unicode-værdien U+201D og UTF-8 hex værdien E2 80 9D eller venstrestillet dobbeltapostrof med unicode-værdien U+201C og UTF-8 hex-værdien E2 80 9C) bliver disse ekstra dobbelt apostroffer ikke fjernet igen, når arkivet konverterer data i afleveringspakken videre til bevaringsformat.

Eksempler på udtræk af variabelværdier der indeholder dobbelt apostrof

Variabelværdien *Antal "dyr" som kan tale* skal i datafilen angives som *"Antal ""dyr"" som kan tale"*
Variabelværdien *"Hurra"* skal i datafilen angives som *""""Hurra""""*

9.G.1.c Som linjeseparator skal anvendes en af følgende metoder for linjeskift: »CR+LF« (U+000D) samt (U+000A) eller »CR« (U+000D) eller »LF« (U+000A). Indhold i variable i datafilen må ikke indeholde linjeskift.

Rækker i datafilen skal adskilles med linjeskift. Når der angives et linjeskift i en tekstfil angives en usynlig linjeseparator med en af følgende unicode værdier »CR+LF« (U+000D) samt (U+000A) eller »CR« (U+000D) eller »LF« (U+000A). Hvilke metoder der anvendes for linjeskift i en tekstfil afhænger af det styresystem der arbejdes på.

I hver linje i datafilen skal angives alle variabelværdier, svarende til antallet af variabelnavne angivet i første linje, før der angives et linjeskift. Da variabelværdierne adskilles med semikolon, betyder det at der i alle linjer i datafilen skal være præcis lige mange semikolon anvendt før et linjeskift.

Variabelværdier i datafilen må ikke indeholde linjeskift, da dette vil medføre at linjen indeholder færre semikolon end tilladt.

Eksempel på datafil med korrekte linjeskift

```
ID;Alder;Vægt;Indkomst;Stilling
00001;45;70,4;45000;IT-medarbejder
00002;63;85,3;130000;Leder
00003;21;76,2;23000;Postbud
```

9.G.2 En manglende værdi kan være en af følgende tre typer: Manglende værdi (tom) (jf. 9.G.2.a), specialkode for manglende værdi (jf. 9.G.2.d) eller brugerdefineret kode for manglende værdi (jf. 9.I.7).

9.G.2.a Manglende værdier (tom) i datafilen skal enten repræsenteres som ingen værdi »«, eller et mellemrum » « (U+0020).

9.G.2.b I en datafil må der konsekvent kun anvendes enten specialkoder eller brugerdefinerede koder for manglende værdier.

9.G.2.c Specialkoder for manglende værdier må kun anvendes for kategoriske og numeriske variable.

9.G.2.d Specialkoder for manglende værdier må kun anvendes for heltal og decimaltal, og skal angives enten som en værdi fra A-Z eller .a-.z.

Hvad er koder for manglende værdier?
Koder for manglende værdier anvendes i statistikfiler og kaldes også missing values.

Typisk skelnes mellem følgende tre typer af koder for manglende værdier i spørgeskemaundersøgelser:

- ***Irrelevant***: Spørgsmålet er ikke besvaret, fordi respondenten ikke skal svare grundet et filter i spørgeskemaet. Et filter i et spørgeskema anvendes, når en respondent ikke skal besvare det efterfølgende spørgsmål, hvis der svares noget specifikt på det foregående spørgsmål. Fx hvis respondenten svarer ”Mand” på spørgsmålet om hans køn, skal han efterfølgende ikke besvare spørgsmålet ”Hvor mange børn har du født”.
- ***Uoplyst***: Spørgsmålet er ikke besvaret, fordi respondenten ikke ønskede at svare.
- ***Deltager ikke*** Spørgsmålet er ikke besvaret, fordi respondenten ikke er en del af undersøgelsen (out of sample) selvom respondenten fremgår af statistikfilen.

Præcise informationer om hvorfor værdier mangler er vigtige for at kunne opnå pålidelige analyseresultater, når data genanvendes. Det er vigtigt at vide hvor godt besvaret et spørgeskema er. Hvis værdier mangler, fordi respondenterne ikke ønskede at besvare spørgsmålene, er data mangelfulde og ikke nær så pålidelige, som hvis værdierne blot mangler grundet spørgsmålsfiltre, fordi respondenterne ikke skulle svare.

Når koder for manglende værdier anvendes på ikke kategoriske variable, men reelle værdier (numeriske variable) som fx alder, antal eller pris, er det vigtigt at koden angives som en missing value i statistikfilen og ikke som en reel værdi. Hvis koden 999 i en variabel med alder ikke er opmærket som en kode for en manglende værdi, vil værdien 999 indgå i aldersgennemsnittet, når der i forbindelse med en statistisk analyse laves et gennemsnit at variablens indhold.

Hvordan angives manglende værdier i datafilen?
En manglende værdi kan angives på forskellige måder i forskellige statistikprogrammer.
I afleveringspakkens datafil kan manglende værdier dog kun angives på følgende 3 måder:

1. **Manglende værdi (tom)**
   Hvis feltet blot er tomt, er værdien fraværende og vi ved ikke hvorfor. Den tomme værdi angives enten som en helt tom værdi »« eller som et mellemrum » «. Det er således ikke tilladt at angive sysmis (punktum) ».« eller værdierne ”NULL”, ”n.a.” osv.

2. **Specialkode for manglende værdi**
   En specialkode for en manglende værdi er en kode der angiver hvorfor en værdi i en variable mangler.

   Brugeren af statistikprogrammet kan selv definere betydningen af specialkoderne, men kan kun anvende specialkoder med værdier der ligger inden for de tilladte udfaldsrum. Udfaldsrummet for specialkoder er A-Z i statistikprogrammet SAS og .a-.z i statistikprogrammet Stata.

   Eksempler på specialkoder for manglende værdier:

   .i = irrelevant
   .u = uoplyst
   .f = fejl

   Det særlige ved specialkoder er, at de i statistikprogrammerne SAS og Stata angives i en *special numeric* datatype, dvs. en datatype som er defineret som numerisk, men også kan indeholde tekst i form af disse specialkoder.

   Specialkoder må kun anvendes for kategoriske eller numeriske variable af datatypen heltal eller decimaltal. Dvs. at de ikke må anvendes i variable med datatyperne tekst, tidspunkt, dato og tidsstempel.

3. **Brugerdefineret kode for manglende værdi**
   En brugerdefineret kode for en manglende værdi er en kode der angiver hvorfor værdien er manglende. Modsat specialkoder, som skal vælges fra værdier i faste udfaldsrum, kan de brugerdefinerede koder frit defineres af brugeren med bogstaver, tal, specialtegn osv.

Brugerdefinerede koder anvendes typisk i statistikprogrammet SPSS.

Eksempler på brugerdefinerede koder for manglende værdier:

999 = irrelevant
100 = uoplyst
oos = out of sample
fejl = fejl i besvarelse

Læs mere om krav til brugerdefinerede koder for manglende værdier i punkt 9.I.7.

Hvad er forskellen på kategoriske og numeriske variable?

Kategorisk variabel
En kategorisk variabel indeholder værdier, som angiver koder for kategorier.
Til en kategorisk variabel er knyttet en række kodeforklaringer (value labels eller format). Dvs. værdien i variablen er en kode, hvis betydning fremgår af en kodeforklaring.
I spørgeskemaundersøgelser svarer disse koder og kodeforklaringer ofte til svarkategorier til et spørgsmål.

Eksempel på en kategorisk variabel

Variabelnavn: Ægteskab
Variabelbeskrivelse: Er du gift?

Kodeliste (value labels/format)
1 = Ja
2 = Nej
3 = Ved ikke

Numerisk variabel
En numerisk variabel kan være et heltal eller et decimaltal. En numerisk variabel indeholder reelle værdier. Fx alder, pris, indkomst eller antal.

9.G.3 Indholdet af de enkelte variable skal renses for eventuelle foran- og efterstillede blanktegn.

Indholdet af de enkelte variable skal renses for eventuelle foran- og efterstillede blanktegn (fx mellemrum).

Blanktegn vil altid blive fjernet i arkivets efterfølgende konvertering af afleveringspakken til bevaringsformat. Hvis blanktegnene er betydningsbærende skal de derfor erstattes af andre tilladte tegn. Dette gælder særligt for variable der indgår som nøgler i relationer mellem flere datafiler i afleveringspakken.

9.H. Datatyper

9.H.1 De seks standardiserede datatyper, som skal anvendes i datafilen, fremgår af figur 9.3.

9.H.2 Dataformatnotationer for anvendte dataformater skal angives i metadatafilen, jf. figur 9.3. Dataformatnotationerne er case sensitive.

9.H.2.a Værdier for bogstaverne »w« og »d« i figur 9.3 skal konsekvent angives. »w« angiver datatypens totale bredde. »d« angiver decimaler eller præcision i fraktioner af sekunder.

Figur 9.3 Tilladte datatyper

Datatype	**Dataformat i datafil**	Eksempler på data i datafil	**Dataformatnotation i metadatafil**	
Tekst	UTF-8 tegnsæt, jf. punkt 9.F.1.	"Antal hjorte; 6" "Hun sagde ""Hej""" Jeg spiller fodbold	xml	string
			Stata	%ws (fx %20s)
			SAS	$w. (fx $20.)
			SPSS	aw (fx a20)
Numerisk heltal	Repræsentation af et heltal med eller uden fortegn i henhold til DS/ISO 6093:1985 (NR1) standard, jf. syntaksregel i figur 9.6.	25 +45 -234 10000	xml	int
			Stata	%w.0f (fx %3.0f)
			SAS	fw. (fx f3.)
			SPSS	fw (fx f3)
Numerisk decimaltal	Repræsentation af et decimaltal med eller uden fortegn i henhold til DS/ISO 6093:1985 (NR2) standard, jf. syntaksregel i figur 9.7.	23,75 .10 -123.76 150000.25	xml	decimal
			Stata	%w.df eller %w.dg (fx %9.2f)
			SAS	fw.d (fx f9.2)
			SPSS	fw.d (fx f9.2)
Dato	Angivelse af kalenderdato i henhold til DS/ISO 8601:1993 udvidet format: CCYY-MM-DD	2019-11-15	xml	date
		2019-11-15	Stata	%tdCCYY-NN-DD
		2019-11-15	SAS	yymmdd10.
	Alternativt kan følgende format anvendes, jf. syntaksregel i figur 9.8: CCYY/MM/DD	2019/11/15	SPSS	sdate10
Tidspunkt	Angivelse af tidspunkt i henhold til DS/ISO 8601:1993 udvidet format, jf. syntaksregel i figur 9.9: hh:mm:ss	23:12:06 05:10.23	xml	time
			Stata	%tcHH:MM:SS
		4:22:59 05:10.23	SAS	time. eller time8.
			SPSS	time8
Tidsstempel	Angivelse af dato og tidspunkt i henhold til DS/ISO 8601:1993 udvidet format:	2019-11-15T08:10:23 Med fraktioner af sekunder: 2019-11-15T08:10:23.123456	xml	datetime

	CCYY-MM-DDThh:mm:ss.sss eller CCYY-MM-DD hh:mm:ss.sss Fraktioner af sekunder er valgfrit og tilladt med en præcision på op til 6 cifre. Tidszone i tidsangivelser er ikke tilladt, jf. syntaksregel i figur 9.10.	2019-11-15T08:10:23	Stata	%tcCCYY-NN-DD!THH:MM:SS
		2019-11-15T08:10:23.123	Stata	%tcCCYY-NN-DD!THH:MM:SS.sss
		2019-11-15T08:10:23.1		%tcCCYY-NN-DD!THH:MM:SS.s
		2019-11-15T08:10:23	SAS	e8601dt19.
		2019-11-15T08:10:23 .123456	SAS	e8601dtw.d (fx e8601dt25.6)
		2019-11-15T08:10:23 .12345		e8601dtw.d (fx e8601dt24.5)
		2019-11-15 08:10:23	SPSS	ymdhms19
		2019-11-15 08:10:23.123456	SPSS	ymdhmsw.d (fx e8601dt26.6)
		2019-11-15 08:10:23.1234		ymdhmsw.d (fx e8601dt24.4)
	Alternativt kan følgende format anvendes, jf. syntaksregel i figur 9.10: dd-mmm-yyyy hh:mm:ss	15-NOV-2019 08:10:23	SPSS	datetime20

9.I. Metadatafil

9.I.1 Metadata fra en datafil afleveres som en struktureret tekstfil, der overholder syntaksen beskrevet som EBNF i figur 9.11.

9.I.1.a Metadatafilen udformes som anvist i figur 9.4, hvor otte etiketter opdeler metadata i specifikke kategorier.

9.I.1.b Hver etiket skal forekomme en gang i metadatafilen. Forekomsten af indholdet af etiketten fremgår af kolonnerne »Forekomst« og »Obligatorisk« i figur 9.4.

9.I.1.c Etiketnavne er reserverede ord og må ikke benyttes til navngivning af metadatafilens øvrige indhold.

Eksempel på metadatafil (table1.txt) der overholder EBNF i figur 9.11

table1_uddrag.txt - Notesblok

Filer Rediger Formater Vis Hjælp

```
SYSTEMNAVN
SPSS

DATAFILNAVN
Generationsundersøgelsen

DATAFILBESKRIVELSE
Generationsundersøgelse som har til formål at undersøge sammenhæng en mellem sundhed, IQ og arv. Indholdet af dette datasæt er oplysninger om studerende.

NØGLEVARIABEL
userid

REFERENCE
stata12345 'barn_userid' 'userid'
sas12345 'barnebarn_userid' 'userid'

VARIABEL
userid a765
koen f3 koen.
vaegt_idag f3.1
slut_dato sdate10
hobby f3 hobby.
laegebesoeg datetime20
maaling time8

VARIABELBESKRIVELSE
userid 'userid'
koen 'Er du mand eller kvinde?'
vaegt_idag 'Hvad er din vægt idag angivet i kilo?'
slut_dato 'Gymnasial uddannelse slut dato'
hobby 'Hvad er din hobby?'
laegebesoeg 'Læge besøg - dato og tidspunkt'
maaling 'Blodtryksmåling tidspunkt'

KODELISTE
hobby
'1' 'Sport (cykling, svomning, løb, fitness træning...)'
'2' 'Film'
'3' 'Rejse'
'4' 'Kunst, musik'
'5' 'Bøger'
'6' 'Håndarbejde'
'7' 'Puslespil'
'8' 'andet'
'9' 'uoplyst'
koen
'1' 'Mand'
'2' 'Kvinde'
'9' 'uoplyst'

BRUGERKODE
hobby '9'
koen '9'
```

Figur 9.4 Metadatafilens struktur

Etiket	Beskrivelse af etiketindhold	Udfaldsrum	Forekomst	Obligatorisk
SYSTEMNAVN	Navn på det program, som data udtrækkes fra, eller datas oprindelige format.	SPSS SAS Stata Excel eller fritekst	1	Ja
DATAFILNAVN	Navnet på datafilen, som den benævnes i brugssammenhæng. Datafilnavn skal overholde krav i punkt 9.I.3.	ISO/IEC 9075:1999 - Database Language SQL (SQL-99)	1	Ja
DATAFILBESKRIVELSE	Beskrivelse af datafilens indhold.	Fritekst	1	Ja
NØGLEVARIABEL	Datafilens unikke nøglevariabel angivet med	ISO/IEC 9075:1999 -	0-1	Ja, hvis unik nøglevariabel

	navne på den/de variable, nøglevariablen består af.	Database Language SQL (SQL-99)		findes
REFERENCE	Referencer til andre datafiler i afleveringspakken angivet som anvist i punkt 9.I.4.	ISO/IEC 9075:1999 - Database Language SQL (SQL-99)	0-m	Ja, hvis reference til anden datafil i afleveringspakken findes
VARIABEL	Variable i datafilen. En variabel angives som et sæt bestående af variabelnavn efterfulgt af notation for variablens dataformat, jf. figur 9.3, samt en eventuel kodelistereference, jf. 9.I.6.f. Variable skal overholde krav i punkt 9.I.5.	Navngivning af variable skal overholde ISO/IEC 9075:1999 - Database Language SQL (SQL-99)	1-m	Ja
VARIABELBESKRIVELSE	Beskrivelser af variablenes indhold.	Fritekst	1-m	Ja
KODELISTE	Kodelister angivet med kodelistens navn efterfulgt af flere sæt bestående af kode og kodebeskrivelse. Kodelister skal overholde krav i punkt 9.I.6.	Navngivning af kodelister skal overholde ISO/IEC 9075:1999 - Database Language SQL (SQL-99)	0-m	Ja, hvis der findes kodelister
BRUGERKODE	Brugerdefinerede koder for manglende værdier angivet med navnet på den variabel, hvor koderne anvendes, efterfulgt af de brugerdefinerede koder. Brugerdefinerede koder for manglende værdier skal overholde krav i punkt 9.I.7.	Navngivning af variable skal overholde ISO/IEC 9075:1999 - Database Language SQL (SQL-99)	0-m	Ja, hvis der findes brugerdefinerede koder for manglende værdier

Hvordan skabes metadatafilen?
Metadatafilen kan udtrækkes automatisk fra den oprindelige statistikfil fra statistikprogrammerne SAS, Stata og SPSS ved brug af Rigsarkivets værktøj ASTA.

Laves udtræk fra et regneark, kan metadatafilen udfyldes manuelt i en teksteditor. Se Rigsarkivets vejledning om aflevering af statistikdata fra regneark og csv-filer på www.sa.dk.

SYSTEMNAVN
Under etiketten SYSTEMNAVN angives navn på det program, som data udtrækkes fra, eller datas oprindelige format.
Angiv enten SPSS, SAS, Stata, Excel eller andet i fritekst.

Eksempel:
```
SYSTEMNAVN
SPSS
```

DATAFILNAVN
Under etiketten DATAFILNAVN angives navnet på datafilen, som den benævnes i brugssammenhæng. Datafilnavne skal være unikke inden for samme afleveringspakke, jf. 9.I.3.
Bemærk at datafilnavn *ikke* skal indeholde filens extension (fx .sav)

Eksempel:
```
DATAFILNAVN
Generationsundersøgelsen
```

DATAFILBEKSRIVELSE
Under etiketten DATAFILBESKRIVELSE angives en beskrivelse af datafilens indhold, fx den fulde titel for statistikfilen, en eksisterende projektbeskrivelse eller et abstract, der præcist dækker statistikfilens indhold.

Bemærk at datafilbeskrivelsen ikke må indeholde tegn for linjeskift.

Eksempel:
```
DATAFILBESKRIVELSE
Generationsundersøgelse som har til formål at undersøge sammenhæng en mellem
sundhed, IQ og arv. Indholdet af dette datasæt er oplysninger om studerende.
```

NØGLEVARIABEL
Under etiketten NØGLEVARIABEL angives datafilens unikke nøglevariabel med navne på den/de variable, nøglevariablen består af. Nøglevariablen skal entydigt identificere en række i datafilen. Dvs. to værdier i variablen må ikke have samme værdi.

En nøglevariabel kan være sammensat af flere variable. Flere variable i en nøglevariabel adskilles af mellemrum.

Eksempel:
```
NØGLEVARIABEL
fødselsdato lbnr
```

REFERENCE
Under etiketten REFERENCE angives referencer til andre datafiler i afleveringspakken, som anvist i punkt 9.I.4.

Hvis der findes flettevariable som kan koble to datasæt (datafiler) i afleveringspakken sammen skal disse angives under etiketten REFERENCE i metadatafilen.

Der skal kun angives de relationer, hvor sammenfletningen resulterer i, at variable tillægges datasættet (horisontal sammenfletning). Dvs. i det tilfælde hvor der er to datasæt med forskellige variable, men samme individer, som over en fælles flettevariabel kan samles til et datasæt.

Relationer til eventuelle nøglefiler skal også angives her. Hvis du fx afleverer et datasæt, hvor personfølsomme oplysninger fra datasættet er udtrukket til en nøglefil, skal relationen mellem datasættet og nøglefilen defineres i en reference.

Eksempel:

```
REFERENCE
stata12345 'barn_userid' 'userid'
sas12345 'barnebarn_userid' 'userid'
```

VARIABEL

Under etiketten VARIABLE angives alle variable i datafilen.

Variabelnavne skal være unikke inden for samme metadatafil., jf. 9.I.5.

En variabel angives som et sæt bestående af variabelnavn efterfulgt af notation for variablens dataformat, jf. figur 9.3, samt en eventuel kodelistereference, jf. 9.I.6.f.

For hver variabel angives således 2-3 oplysninger.

1) ***Første oplysning*** er angivelse af variabelnavn. Variabelnavne skal overholde SQL99 standarden (ISO/IEC 9075:1999 - Database Language SQL) for angivelse af tabel- og kolonnenavne. Dvs. at variabelnavn må aldrig begynde med et tal, men må godt være en blanding af bogstaver, tal samt underscore på max 128 tegn. Hvis variabelnavnet er et reserveret ord i SQL:1999, skal det enten omdøbes eller omkranses af dobbeltapostrof.
2) ***Anden oplysning*** er angivelse af variablens datatype. Tilladte datatyper for variable fremgår af figur 9.3. Det er dataformatnotationerne angivet i kolonne ”Dataformatnotation i metadatafil”, som skal angives som variablens datatype. Datatypen *f10* i eksemplet nedenfor angiver, at variablen *id* er af datatypen heltal med en længde på 10 cifre.
3) ***Tredje oplysning*** er angivelse af en eventuel kodelistereference. Tilladte notationer for angivelse af kodelistereferencer fremgår af 9.I.6.g og 9.I.6.h.

Eksempel:

```
VARIABEL
id f10
køn f1 køn.
fødselsdato sdate10
fødselstidspunkt time8
ægteskab f2 ægteskab.
nationalitet a2 $nationalitet.
kommentar a200
```

VARIABELBESKRIVELSE

Under etiketten VARIABELBEKSRIVELSE angives beskrivelser af variablenes indhold.

Alle variable skal have en variabelbeskrivelse, der udfyldes så udførligt som muligt, således at en fremtidig bruger har mulighed for at forstå hvilken oplysning der er registreret i variablen. Alle koder i beskrivelsen skal forklares og eventuelle måleenheder skal angives (fx meter, centimeter, kilo). Findes datoer, skal det angives hvad datoen omhandler, fx bryllupsdato eller fødselsdato. Findes oplysninger om flere individer i et datasæt, fx både mor og barn, skal det tydeligt af variabelbeskrivelsen fremgå hvilke variabeloplysninger der er knyttet til hvilke individer, fx ’*Barnets køn’* eller *’Respondentens køn’*.

For spørgeskemaundersøgelser anbefaler Rigsarkivet, at alle spørgsmålsteksterne indsættes i variabelbeskrivelserne (variable labels i statistikfilen). Hermed bliver det lettere at søge i og anvende data for fremtidige brugere. Spørgsmålsteksterne bør indsættes i deres oprindelige formulering og fulde længde. Hvis spørgsmålsteksterne er for lange til at blive udtrukket automatisk fra statistikfilen, kan de efterfølgende indsættes direkte i metadatafilen (fx table1.txt) under etiketten VARIABELBESKRIVELSE.

Eksempel:

```
VARIABELBESKRIVELSE
id ’Respondentens cpr-nummer’
køn ’Respondentens køn’
ægteskab ’Er du gift?’
dato ’Hvilken dato blev du gift?’
nationalitet ’Hvilket statsborgerskab har du?’
vægt ’Hvad vejer du? (Angivet i hele kilo)’
kommentar ’Har du yderligere kommentarer kan du angive dem her.’
```

KODELISTE

Under etiketten KODELISTE listes datasættets kodelister angivet med kodelistens navn efterfulgt af flere sæt bestående af kode og kodebeskrivelse.

Kodelister skal desuden overholde kravene i punkt 9.I.6.

I spørgeskemaundersøgelser svarer kodelister til svarkategorier til et spørgsmål.
I statistikprogrammet SPSS svarer kodelister til de *value labels* der er knyttet til variablene.
I statistikprogrammet SAS svarer kodelister til formaterne i format-filen (katalog-filen).

Et kodesæt under etiketten KODELISTE i metadatafilen består af et kodelistenavn hvorunder der angives rækker med koder og tilhørende kodebeskrivelser. Hver række skal bestå af en kode og en kodebeskrivelse. Begge oplysninger skal hver især omkranses af enkeltapostroffer og adskilles med mellemrum.

Eksempel:

```
KODELISTE
Nationalitet
’DA’ ’Danmark’
’GB’ ’England’
’SE’ ’Sverige’
’9’ ’uoplyst’
’10’ ’irrelevant’
Køn
’1’ ’Mand’
’2’ ’Kvinde’
```

BRUGERKODE

Under etiketten BRUGERKODE angives brugerdefinerede koder for manglende værdier med navnet på den variabel, hvor koderne anvendes, efterfulgt af de brugerdefinerede koder omkranset af enkeltapostroffer. Bemærk at betydningen af de brugerdefinerede koder angives i kodelisten knyttet til den variabel, hvor koderne anvendes, under etiketten KODELISTE.

Brugerdefinerede koder for manglende værdier må kun anvendes i kategoriske og numeriske variable med datatyperne heltal, decimaltal, og tekst jf. 9.I.7.

Eksempel:

```
BRUGERKODE
Nationalitet ’9’ ’10’
```

9.I.3 Datafilnavne skal være unikke inden for samme afleveringspakke.

Hvis der udtrækkes data fra flere statistikfiler til en afleveringspakke skal disse navngives forskelligt. Navnet på en datafil angives i metadatafilen til den pågældende datafil under etiketten DATAFILNAVN

9.I.4 Reference

9.I.4.a En reference til en anden datafil i afleveringspakken angives med navnet på den datafil, der refereres til (fremmeddatafil), efterfulgt af variabelnavnet for fremmeddatafilens nøglevariabel (fremmedvariabel) efterfulgt af variabelnavnet for den variabel (referencevariabel) i datafilen, der refererer til fremmedvariablen i fremmeddatafilen.

Eksempel på angivelse af reference i metadatafilen til Datasæt 1

I dette eksempel er angivet en reference over de to flettevariable *nøglefil_id* og *id*.

Flettevariablen *id* findes i hoveddatasættet med datafilnavnet *Generationsundersøgelsen*.

Flettevariablen *nøglefil_id* findes i datasættet med datafilnavnet *Nøglefil*.

Referencen defineres under etiketten REFERENCE i metadatafilen for hoveddatasættet *Generationsundersøgelsen*, som indeholder flettevariablen (referencevariablen) *id*, som peger på den unikke flettevariabel/nøglevariabel *nøglefil_id* (fremmedvariablen) i datasættet *Nøglefil*.

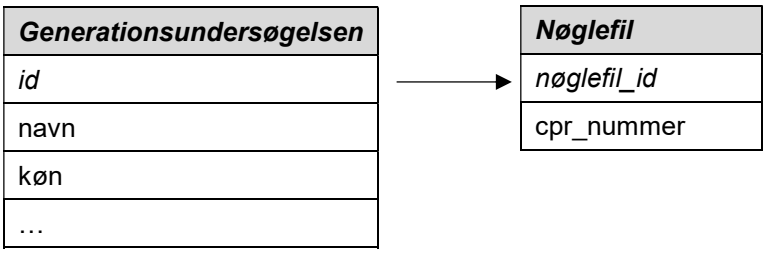
Datasæt 1:

```
…
DATAFILNAVN
Generationsundersøgelsen
…
REFERENCE
Nøglefil 'nøglefil_id' 'id'

VARIABEL
id f8
navn a200
køn f2 køn.
…
```

Datasæt 2:

```
…
DATAFILNAVN
Nøglefil
…
REFERENCE


VARIABEL
nøglefil_id f8
cpr_nummer f10
```

Illustration af referencen angivet ovenfor:



9.I.4.b Der skal være fuld overensstemmelse mellem datatype og længde i nøglevariablene, der indgår i referencen.

9.I.5 Variabelnavne skal være unikke inden for samme metadatafil.

To variable under etiketten VARIABEL i metadatafilen knyttet til en datafil må ikke have samme navn.

9.I.6 Kodeliste

9.I.6.a Kun kategoriske og numeriske variable må have henvisninger til en kodeliste.

9.I.6.b Det er kun tilladt at medtage kodelister for variable, der indeholder data af datatypen numerisk heltal, decimaltal eller tekst, jf. datatyper i figur 9.3.

Det er ikke tilladt at anvende koder i og knytte kodelister til variable med datatyperne dato, tidspunkt og tidsstempel.

En kategorisk variabel indeholder værdier, som angiver koder for kategorier. Datatyperne for en kategorisk variabel kan være enten heltal, decimaltal eller tekst.

Eksempler på værdier i kategoriske variable

Heltal
1 = Ja
2 = Nej
3 = Ved ikke

Decimaltal
1.00 = Ingen smerte
2.00 = Lidt smerte
3.00 = Meget smerte

Tekst
DA = Danmark
GB = England
SE = Sverige

En numerisk variabel indeholder reelle værdier. Fx alder, pris, indkomst eller antal. Numeriske variable kan have datatyperne heltal eller decimaltal.

Eksempler på værdier i numeriske variable

Heltal
1
2
3
…

Decimaltal
150013,50
12420,67
34880,54
…

9.I.6.c Alle koder i data skal defineres i en kodeliste og forklares med en kodebeskrivelse.

9.I.6.d Hvis koder er defineret som et interval, hvor ikke alle koder har kodebeskrivelser, skal dette dokumenteres i variabelbeskrivelsen.

For variable der anvender koder der er defineret som et interval tilføjes følgende eller lignende dokumentation i variabelbeskrivelsen: ”Skala anvendt. Ikke alle koder har kodebeskrivelse.”

Eksempel på koder defineret som et interval

Et eksempel på koder defineret som et interval er en Likert skala, hvor der fx stilles følgende spørgsmål:

På en skala fra 1-5, hvor enig er du i følgende udsagn: ”Jeg er altid glad”?

Alle besvarelserne 1, 2, 3, 4, 5 findes i datasættet, men der findes kun kodebeskrivelserne 1 = Meget uenig og 5 = Meget enig. Kodelisten angives på følgende måde i metadatafilen:

```
KODELISTE
humør
’1’ ’Meget uenig’
’5’ ’Meget enig’
```

9.I.6.e Koder skal være unikke inden for samme kodeliste.

To eller flere koder i en kodeliste må ikke have samme værdi.

9.I.6.f Referencen mellem en kodeliste og den variabel, som refererer til kodelisten, angives under etiketten VARIABEL, jf. figur 9.4, som anvist i 9.I.6.g og 9.I.6.h.

9.I.6.g Hvis datatypen er numerisk heltal eller decimaltal, angives referencen med det valgte kodelistenavn og et efterstillet punktum ».« (U+002E).

Eksempel på en kodelistereference når variablen er af datatypen heltal eller decimaltal:

```
VARIABEL
ægteskab f2 ægteskab.
```

9.I.6.h Hvis datatypen er tekst, angives referencen med det valgte kodelistenavn med et foranstillet dollartegn »$« (U+0024) og et efterstillet punktum ».« (U+002E).

Eksempel på en kodelistereference når variablen er af datatypen tekst:

```
VARIABEL
nationalitet a2 $nationalitet.
```

9.I.7 Brugerkode

Brugerkode er en kort betegnelse for brugerdefinerede koder for en manglende værdier.

En brugerdefineret kode for en manglende værdi er en kode der angiver hvorfor værdien er manglende. Modsat specialkoder, som skal vælges fra værdier i faste udfaldsrum, kan de brugerdefinerede koder frit defineres af brugeren med bogstaver, tal, specialtegn osv.

Brugerdefinerede koder anvendes typisk i statistikprogrammet SPSS.

Eksempler på brugerdefinerede koder for manglende værdier

9 = irrelevant
10 = uoplyst
11 = deltager ikke
? = uafklaret
fejl = fejl i besvarelse

9.I.7.a En brugerdefineret kode for en manglende værdi er kun tilladt for kategoriske og numeriske variable.

Brugerdefinerede koder kan i statistikprogrammet SPSS anvendes på alle variable, blot de overholder variablens datatype. Jf. bilag 9 er det dog i datafilen kun tilladt at anvende koder for manglende værdier for kategoriske variable med datatyperne heltal, decimaltal eller tekst eller numeriske variable med datatyperne heltal eller decimaltal.

En kategorisk variabel indeholder værdier, som angiver koder for kategorier. Til en kategorisk variabel er skal altid knyttet kodeforklaringer, der tydeliggør betydningen af koderne.

En numerisk variabel indeholder reelle værdier. Fx alder, pris, indkomst eller antal.

Eksempler på værdier i kategoriske variable inkl. eksempler på brugerkoder

Heltal
1 = Ja
2 = Nej
3 = Ved ikke
9 = uoplyst
10 = irrelevant

Decimaltal
1.00 = Ingen smerte
2.00 = Lidt smerte
3.00 = Meget smerte
9 = uoplyst
10 = irrelevant

Tekst
DA = Danmark
GB = England
SE = Sverige
9 = uoplyst
8 = irrelevant

Eksempler på værdier i numeriske variable samt eksempel på brugerkoder

Heltal (Alder)
1
2
3
…

999 = uoplyst
1001 = irrelevant

Decimaltal (Indkomst)
150000,00
12420,50
34880,00

Det er ikke tilladt at anvende brugerdefinerede koder for manglende værdier i variable med datatyperne dato, tidspunkt og tidsstempel.

Hvis en dato, fx 9999-12-31, anvendes som en kode for en manglende værdi for en variabel i den statistikfil udtrækket laves fra, bør denne kode omkodes til en manglende værdi (tom), jf. 9.G.2.a, før udtræk til datafil. Er betydningen af denne værdi ikke forklaret fyldestgørende ved en tom værdi eller findes der flere datoer anvendt som koder for manglende værdier med forskellige betydninger i samme variabel, kan disse medtages i udtrækket til datafilen men uden angivelse af brugerkoderne i metadatafilen. Betydningerne af brugerkoderne skal da i stedet dokumenteres i variabelbeskrivelsen for variablen, hvor koderne anvendes, eller i et kontekstdokument.

9.I.7.b En brugerdefineret kode for en manglende værdi angivet i metadatafilen, skal altid fremgå af kodelisten, som den tilhører.

Eksempel på brugerdefinerede koder for manglende værdier angivet i kodelisten

```
KODELISTE
Nationalitet
'DA' 'Danmark'
'GB' 'England'
'SE' 'Sverige'
'9' 'uoplyst'
'10' 'irrelevant'

BRUGERKODE
Nationalitet '9' '10'
```

Hvad er EBNF?
Figur 9.6 til 9.12 herunder anviser EBNF-syntakser for de tilladte datatyper i datafiler samt for opbygning af en data- og metadatafil i afleveringspakken. EBNF er en notationsform der meget kort og præcist ved brug af særlige tegn definerer hvad der er tilladt. En tegnforklaring til hvordan EBNF-syntaksen skal læses fremgår af figur 9.5.

Et heltal er fx defineret som [FORTEGN] CIFFER {CIFFER}, hvilket betyder at et heltal må bestå af et [FORTEGN], hvor de kantede parenteser angiver at fortegn kan forekomme 0-1 gange, dvs. det er valgt frit. Herefter skal altid komme et CIFFER efterfulgt af {CIFFER}, hvor tuborg-klammerne angiver at dette andet ciffer kan forekommer 0 eller flere gange.

Figur 9.5 EBNF tegnforklaring

::= defineret som
() angiver en gruppering, der skal udføres samlet
[] angiver muligheden for (0 eller 1)
{} angiver mulig gentagelse (0 eller flere)
| angiver et valg (enten eller)
… angiver et fortløbende interval
!! angiver en beskrivende forklaring på alm. dansk
”” omslutter faktiske værdier, der skal skrives uden fortolkning

Figur 9.6 EBNF for heltalstype DS/ISO 6093:1985 (NR1) standard

Nonterminal	**Terminal**	Eksempel
INT ::=	NR1	25 +45 -234 10000
NR1 ::=	[FORTEGN] CIFFER {CIFFER}	
FORTEGN ::=	”+” | ”-” *!! Det er valgfrit at anvende fortegn for positive heltal*	
CIFFER ::=	”0” | ”1” | … | ”9”	

Figur 9.7 EBNF for decimaltalstype DS/ISO 6093:1985 (NR2) standard

Nonterminal	**Terminal**	Eksempel
DECIMAL ::=	NR2	23,75 .10 -123.76 +123.76 150000.25
NR2 ::=	[FORTEGN] CIFFER {CIFFER} | [FORTEGN] CIFFER {CIFFER} DECIMALMÆRKE CIFFER {CIFFER} | DECIMALMÆRKE CIFFER {CIFFER}	
FORTEGN ::=	”+” | ”-” *!! Det er valgfrit at anvende fortegn for positive decimaltal !! Det er ikke tilladt at anvende negativt fortegn foran værdien 0*	
CIFFER ::=	”0” | ”1” | … | ”9”	
DECIMALMÆRKE ::=	”.” | ”,”	

Figur 9.8 EBNF for datotyper

Nonterminal	Terminal	Eksempel
DATO ::=	ISO-8601-DATE | ALTERNATIV-DATE	
ISO-8601-DATE ::=	CC YY BINDESTREG MM BINDESTREG DD	2019-11-15
CC ::=	CIFFER CIFFER *!! årtusinde angivet med to heltal*	
BINDESTREG ::=	”-” *!! bindestreg (U+002D)*	
YY ::=	CIFFER CIFFER *!! årstal angivet med to heltal*	
MM ::=	CIFFER CIFFER *!! måned angivet med to heltal*	
DD ::=	CIFFER CIFFER *!! dag angivet med to heltal*	
CIFFER ::=	”0” | ”1” | … | ”9”	
ALTERNATIV-DATE ::=	CC YY SKRÅSTREG MM SKRÅSTREG DD	2019/11/15
SKRÅSTREG ::=	”/” *!! skråstreg U+002F*	

Figur 9.9 EBNF for tidstyper

Nonterminal	Terminal	Eksempel
TIDSPUNKT ::=	TIME KOLON MINUT KOLON SEKUND	05:10.23 4:22:59
TIME ::=	CIFFER [CIFFER] *!! time angivet med et eller to heltal !! udfaldsrum er 0-23*	
KOLON ::=	”:” *!! kolon (U+003A)*	
MINUT ::=	CIFFER CIFFER *!! minut angivet med to heltal !! udfaldsrum er 00-59*	
SEKUND ::=	CIFFER CIFFER *!! sekund angivet med to heltal !! udfaldsrum er 00-59*	
CIFFER ::=	”0” | ”1” | … | ”9”	

Figur 9.10 EBNF for datetimetyper

Nonterminal	Terminal	Eksempel
DATETIME ::=	ISO-8601-DATETIME | ALTERNATIV-DATETIME | IBM-DATETIME	
ISO-8601-DATETIME ::=	CC YY BINDESTREG MM BINDESTREG DD TIDSTEMPELMÆRKE TIME KOLON MINUT KOLON SEKUND [PUNKTUM MILLISEKUNDER]	2019-11-15T08:10:23 2019-11-15T08:10:23.123456 2019-11-15T08:10:23.123
CC ::=	CIFFER CIFFER *!! årtusinde angivet med to heltal*	

CIFFER ::=	”0” | ”1” | … | ”9”	
BINDESTREG ::=	”-” *!! bindestreg (U+002D)*	2019-11-15 08:10:23 2019-11-15 08:10:23.123456 2019-11-15 08:10:23.12
YY ::=	CIFFER CIFFER *!! årstal angivet med to heltal*	
MM ::=	CIFFER CIFFER *!! måned angivet med to heltal*	
DD ::=	CIFFER CIFFER *!! dag angivet med to heltal*	
TIDSTEMPELMÆRKE ::=	”T” | MELLEMRUM	
MELLEMRUM ::=	” ” *!! mellemrumstegn (U+0020)*	
TIME ::=	CIFFER CIFFER *!! time angivet med to heltal !! udfaldsrum er 00-23*	
KOLON ::=	”:” *!! kolon (U+003A)*	
MINUT ::=	CIFFER CIFFER *!! minut angivet med to heltal !! udfaldsrum er 00-59*	
SEKUND ::=	CIFFER CIFFER *!! sekund angivet med to heltal !! udfaldsrum er 00-59*	
PUNKTUM ::=	”.” *!! punktum (U+002E)*	
MILLISEKUNDER ::=	CIFFER {CIFFER} *!! millisekunder angivet med op til max 6 cifre*	
IBM-DATETIME ::=	DD BINDESTREG MÅNED BINDESTREG ÅRSTAL MELLEMRUM TIME KOLON MINUT KOLON SEKUND	20-May-2019 11:05:48 15-OCT-2019 08:10:23
MÅNED ::=	”Jan” | ”Feb” | ”Mar” | ”Apr” | ”May” | ”Jun” | ”Jul” | ”Aug” | ”Sep” | ”Oct” | ”Nov” | ”Dec” *!! der skal anvendes engelske forkortelser for månedsnavn, og de er ikke case sensitive.*	
ÅRSTAL ::=	CIFFER CIFFER CIFFER CIFFER *!! årstal angivet med fire heltal*	

Figur 9.11 EBNF Syntaksregler for metadatafil

Nonterminal	**Terminal**	Eksempel

METADATAFIL ::=	SYSTEMNAVN
	DATAFILNAVN
	DATAFILBESKRIVELSE
	NØGLEVARIABEL
	REFERENCE
	VARIABEL
		`SYSTEMNAVN`
		`DATAFILNAVN`
		`DATAFILBESKRIVELSE`
		`NØGLEVARIABEL`
		`REFERENCE`

	VARIABELBESKRIVELSE KODELISTE BRUGERKODE	VARIABEL VARIABELBESKRIVELSE KODELISTE BRUGERKODE
SYSTEMNAVN::=	"SYSTEMNAVN" LINJESKIFT "SAS" | "Stata" | "SPSS" | "Excel" | FRITEKST LINJESKIFT LINJESKIFT {LINJESKIFT}	SYSTEMNAVN SAS SYSTEMNAVN R statistikfil
DATAFILNAVN ::=	"DATAFILNAVN" LINJESKIFT TITEL LINJESKIFT LINJESKIFT {LINJESKIFT}	DATAFILNAVN Generationsundersøgelsen DATAFILNAVN "Aggregate"
LINJESKIFT ::=	(CR LF) | CR | LF	
CR ::=	*!! vognretur (U+000D)*	
LF ::=	*!! linjeskift (U+000A)*	
TITEL ::=	(BOGSTAV {BOGSTAV | CIFFER}) | (DOBBELTAPOSTROF BOGSTAV {BOGSTAV | CIFFER} DOBBELTAPOSTROF) *!! En TITEL må aldrig begynde med et tal, men må godt være en blanding af bogstaver og tal på max 128 tegn, og hvis titel er et reserveret ord i SQL:1999, skal titel omkranses af dobbeltapostrof, jf. ISO/IEC 9075:1999 - Database Language SQL (SQL:1999)*	
DOBBELTAPOSTROF ::=	"""" *!! dobbeltapostrof (U+0022)*	
BOGSTAV ::=	"A" |"B" | … | "Z" | "a" | "b" | … | "z" | *"_"!! Samt andre nationale karakterer, der er tilladte i constraint names i standarden SQL:1999 !! underscore (U+005F)*	
CIFFER ::=	"0" | "1" | … | "9"	
DATAFILBESKRIVELSE ::=	"DATAFILBESKRIVELSE" LINJESKIFT FRITEKST LINJESKIFT LINJESKIFT {LINJESKIFT}	DATAFILBESKRIVELSE Generationsundersøgelse som har til formål at undersøge sammenhæng en mellem sundhed, IQ og arv. Indholdet af dette datasæt er oplysninger om studerende.
FRITEKST ::=	BOGSTAV | CIFFER | ANDRETEGN {BOGSTAV | CIFFER | ANDRETEGN}*!! FRITEKST kan være en blanding af bogstaver, tal og andre tegn, så længe det giver semantisk mening*	
ANDRETEGN ::=	*!! alle tilladte tegn i UTF-8, som ikke er bogstaver og tal jf.*	

	punkt 9.F	

NØGLEVARIABEL ::=	”NØGLEVARIABEL” LINJESKIFT {VARIABELNAVN MELLEMRUM} LINJESKIFT LINJESKIFT {LINJESKIFT}	`NØGLEVARIABEL`
		`fødselsdato lbnr`
		
		`NØGLEVARIABEL`
		`cprnummer`

VARIABELNAVN ::=	TITEL	
MELLEMRUM ::=	*!! mellemrumstegn (U+0020)*	
REFERENCE ::=	”REFERENCE” LINJESKIFT {FREMMEDDATAFIL MELLEMRUM FREMMEDVARIABEL MELLEMRUM REFERENCEVARIABEL LINJESKIFT} LINJESKIFT	`REFERENCE` `Nøglefil 'nøglefil_id' 'hovedfil_id'`
FREMMEDDATAFIL ::=	TITEL *!! navn på den datafil, der refereres til*	
FREMMEDVARIABEL ::=	APOSTROF VARIABELNAVN {MELLEMRUM VARIABELNAVN} APOSTROF *!! navn på nøglevariablen i den datafil, der refereres til*	
REFERENCEVARIABEL ::=	APOSTROF VARIABELNAVN {MELLEMRUM VARIABELNAVN} APOSTROF *!! navn på variablen i datafilen, der refereres fra, som refererer til nøglevariablen i den datafil, der refereres til*	
APOSTROF ::=	”””” *!! enkelt apostrof (U+0027)*	

VARIABEL ::=	”VARIABEL” LINESKIFT VARIABELSÆT {VARIABELSÆT} LINJESKIFT {LINESKIFT}	`VARIABEL`
		`id f10`
		`køn f1 køn.`
		`fødselsdato sdate10`
		`fødselstidspunkt time8`
		`ægteskab f2 ægteskab.`
		`nationalitet a2 $nationalitet.`
		`kommentar a200`

VARIABELSÆT ::=	VARIABELNAVN MELLEMRUM DATAFORMATNOTATION MELLEMRUM [KODELISTEREFERENCE] LINJESKIFT	
KODELISTEREFERENCE ::=	[DOLLAR] TITEL PUNKTUM	`køn.` `$nationalitet.`
DATAFORMATNOTATION ::=	INTEGERNOTATION | DECIMALNOTATION | DATONOTATION | DATETIMENOTATION | TIMENOTATION | TEKSTNOTATION	
INTEGERNOTATION ::=	*!! se figur 9.3 for dataformatnotationer for numerisk heltal*	`int`

		%3.0f (%w.0f) f3. (fw.) f3 (fw)
DECIMALNOTATION ::=	*!! se figur 9.3 for dataformatnotationer for decimaltal*	decimal %9.2f (%w.df eller %w.dg) f9.2 (fw.d) f9.2 (fw.d)
DATONOTATION ::=	*!! se figur 9.3 for dataformatnotationer for datotyper*	date %tdCCYY-NN-DD yymmdd10. sdate10
DATETIMENOTATION ::=	*!! se figur 9.3 for dataformatnotationer for tidsstempler*	datetime %tcCCYY-NN-DD!THH:MM:SS %tcCCYY-NN-DD!THH:MM:SS.sss %tcCCYY-NN-DD!THH:MM:SS.s e8601dt19. e8601dt25.6 (e8601dtw.d) e8601dt24.5 (e8601dtw.d) ymdhms19 e8601dt26.6 (ymdhmsw.d) e8601dt24.4 (ymdhmsw.d) datetime20
TIMENOTATION ::=	*!! se figur 9.3 for dataformatnotationer for tidspunkter*	time %tcHH:MM:SS time. eller time8. time8
TEKSTNOTATION ::=	*!! se figur 9.3 for dataformatnotationer for tekst*	string %20s (%ws) $20. ($w.) a20 (aw)
DOLLAR ::=	”$” *!! dollartegn (U+0024)*	$
PUNKTUM ::=	”.” *!! punktum (U+002E)*	.
VARIABELBESKRIVELSE ::=	”VARIABELBESKRIVELSE” LINJESKIFT BESKRIVELSE {BESKRIVELSE} LINJESKIFT {LINJESKIFT}	VARIABELBESKRIVELSE id ’Respondentens cpr-nummer’ køn ’Respondentens køn’

BESKRIVELSE ::=	VARIABELNAVN MELLEMRUM APOSTROF FRITEKST APOSTROF LINJESKIFT	ægteskab 'Er du gift?' dato 'Hvilken dato blev du gift?' nationalitet 'Statsborgerskab' vægt 'Hvad vejer du? (kilo)'
KODELISTE ::=	"KODELISTE" LINJESKIFT {KODESÆT} LINJESKIFT {LINJESKIFT}	KODELISTE Nationalitet 'DA' 'Danmark' 'GB' 'England' 'SE' 'Sverige' Køn '1' 'Mand' '2' 'Kvinde' '.u' 'Uoplyst'
KODESÆT ::=	KODELISTENAVN LINJESKIFT {KODE MELLEMRUM KODEBESKRIVELSE LINJESKIFT}	
KODELISTENAVN ::=	TITEL	
KODE ::=	APOSTROF HELTAL | DECIMAL | FRITEKST APOSTROF	
KODEBESKRIVELSE ::=	APOSTROF FRITEKST APOSTROF	
HELTAL ::=	*!! se figur 9.6 EBNF for heltalstype*	25 +45 -234 10000
DECIMAL ::=	*!! se figur 9.7 EBNF for decimaltalstype*	23,75 .10 -123.76 +123.76 150000.25
BRUGERKODE ::=	"BRUGERKODE" LINJESKIFT {VÆRDISÆT} LINJESKIFT {LINJESKIFT}	BRUGERKODE Nationalitet '9' '10'
VÆRDISÆT ::=	VARIABELNAVN MELLEMRUM VÆRDI {MELLEMRUM VÆRDI} LINJESKIFT	
VÆRDI ::=	APOSTROF HELTAL | DECIMAL | FRITEKST APOSTROF	

Figur 9.12 EBNF Syntaksregler for datafil

Nonterminal	**Terminal**	Eksempel
DATAFIL ::=	DATAINDHOLD	id;køn;fødselsdato;fødselstidspunkt 0101013333;2;2016-12-15;17:23:04 0202024444;1;1980-02-21;05:15:42
DATAINDHOLD ::=	OVERSKRIFT LINJESKIFT RÆKKE {RÆKKE}	
OVERSKRIFT ::=	VARIABELNAVN {SEPARATORTEGN	

	VARIABELNAVN}*!! Der anvendes ikke et separatortegn efter det sidste variabelnavn i overskriftslinjen*	0303035555;1;1079-11-04;03:33:08
VARIABELNAVN ::=	TITEL	
TITEL ::=	(BOGSTAV {BOGSTAV | CIFFER}) | (DOBBELTAPOSTROF BOGSTAV {BOGSTAV | CIFFER} DOBBELTAPOSTROF) *!! En TITEL må aldrig begynde med et tal, men må godt være en blanding af bogstaver og tal, på max 128 tegn, og hvis titel er et reserveret ord i SQL:1999 skal titel omkranses af dobbeltapostrof, jf. ISO/IEC 9075:1999 - Database Language SQL (SQL:1999)*	
BOGSTAV ::=	”A” |”B” | … | ”Z” | ”a” | ”b” | … | ”z” | *”_”!! Samt andre nationale karakterer, der er tilladte i constraint names i standarden SQL:1999 !! underscore (U+005F)*	
CIFFER ::=	”0” | ”1” | ”…” | ”9”	
SEPARATORTEGN ::=	”;” *!! semikolon (U+003B)*	
LINJESKIFT ::=	(CR LF) | CR | LF	
CR ::=	*!! vognretur (U+000D)*	
LF ::=	*!! linjeskift (U+000A)*	
RÆKKE ::=	VÆRDI | (DOBBELTAPOSTROF VÆRDI DOBBELTAPOSTROF) {SEPARATORTEGN VÆRDI | (DOBBELTAPOSTROF VÆRDI DOBBELTAPOSTROF)} LINJESKIFT *!! Der anvendes ikke et separatortegn efter den sidste værdi i en række. Se 9.G.1.b for omslutning af en værdi, som indeholder separatortegn og/eller dobbeltapostrof*	
VÆRDI ::=	BOGSTAV | CIFFER | ANDRETEGN {BOGSTAV | CIFFER | ANDRETEGN}*!! VÆRDI kan være en blanding af bogstaver tal og andre tegn !!Foran- og efterstillede blanktegn er ikke tilladt i værdier*	
ANDRETEGN ::=	*!! alle tilladte tegn i UTF-8, som ikke er bogstaver og tal, jf. 9.F.*	
DOBBELTAPOSTROF ::=	”””” *!! dobbeltapostrof (U+0022)*	