

Digitalt filarkiv gjort søkbart

Sikker formidling av filbaserte privatarkiv

Rolf Holte, Vestfoldarkivet

Urpremierie !

For første gang i verdenshistorien er KDRS søk og vis
brukt til formidling av fil/mappe arkiv fra et privatarkiv (og
vist frem for itarkivarer)

Utfordringer

- Vestfoldarkivet fikk utviklingsmidler til å lage «veileder for digitale privatarkiv» som er i sluttfasen. En rapport ment for arkivar hvermannsen. Hovedfokus er hvordan selge inn og få tak i digitale arkiv for arkivarer til Næringslivet.
- Bedrifter er svært skeptiske til digitale avleveringer og er veldig opptatt av sikkerhet.
- Selges inn med at det er en sikkerhetskopi/snapshot og at bedriften selv er pliktig å holde orden i eget hus. Denne sikkerhetskopien er den historien bedriften velger skal være deres stemme i fremtiden – Helhetligsamfunnsdokumentasjon.

Utfordringer

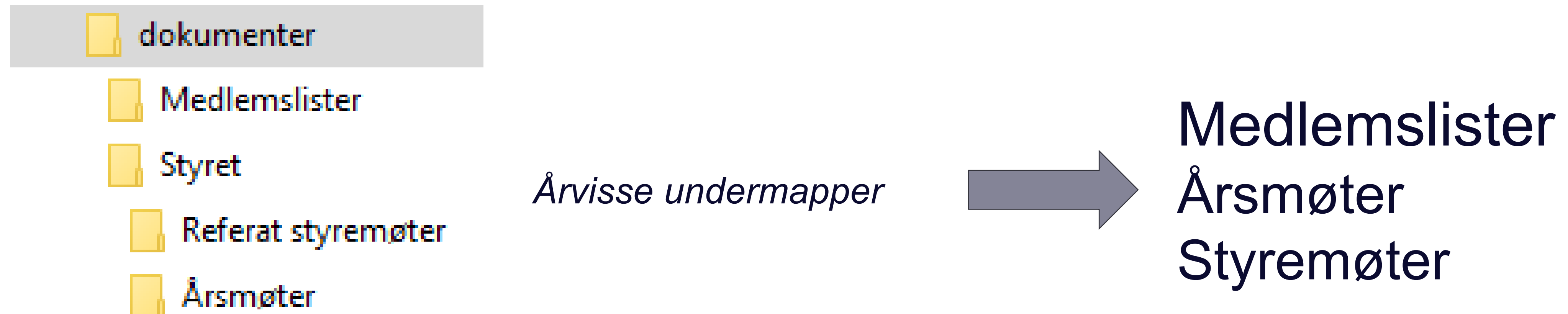
- **Rapportuttrekk** (excel/pdf) selges inn i stedet for databaseuttrekk fordi bedrifter ikke vil slippe inn fremmende for å ta «kompliserte uttrekk» /føler ikke at de har kontroll informasjonsutveksling ved for eksempel siard. Ønsker heller ikke at eksterne skal ha all info, men spesifikk info de har valgt ut.
- Vi kommer til å få filbaserte digitale arkiv med **sensitive data**, som ikke bør slenges på filserveren vår
- Vi IT arkivarer vet at filbaserte digitale privatarkiv etter foreninger og lag, bedrifter kommer til å måtte **formidles før eller senere**.
- **Hvilke løsninger i dag kan gi sikker, kontrollert og sporbar tilgang?**

Hva hvis!

Hvorfor fant jeg på dette?

Case: Vestviken blandakor

- Kun digitalt. Arkivet er samlinger av dokumenter i en mappestruktur som på overordnet nivå består av:



- Ca 300 filer (84 Mb) hovedsakelig en blanding av PDF og Word og Excel. Bildefiler er kun skannede signaturer (godkjenning av årsmøtereferat)

Valgmuligheter Vestfoldarkivet har for digitale privatarkiv?

- Håndtere selv i egen organisasjon
- Digitalarkivet
 - Startfasen med bevaring (digitalt sikringsmagasin)
 - Har formidling for Noark
 - Tredjeparter som KDRS vil kunne lage tjenester for formidling
- KDRS
 - Det digitale sikringsmagasinet har vært operativt i mange år
 - Har innsynsløsning for **Noark 5**
 - Har innsynsløsning for siard (**KDRS søk og vis**)
 - Skal inngå som **tredjepartsleverandør til Digitalarkivet**

Hvilke valg har **Vi** for digitale privatarkiv?

- Digitalarkivet
- KDRS
- Begge disse vil ha innsyn basert på Siard

KDRS Søk og vis - Hva trenger vi for å vise filer?

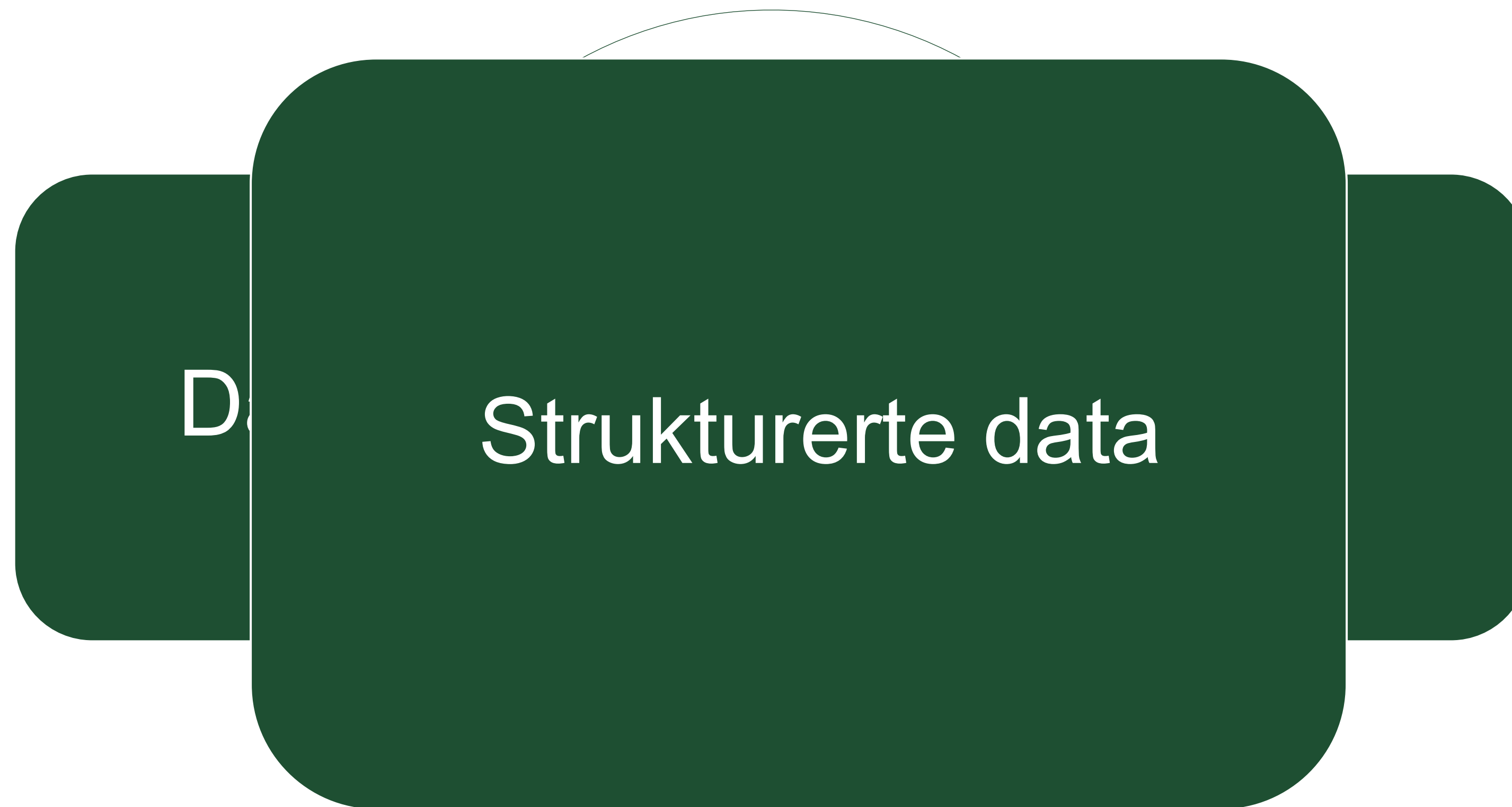


Siard metoden

Kan siard brukes for fil og mappe arkiv?

Hva hvis?

Hva trenger vi for å lage en siard?



Hva om filarkiv var del av ett siard uttrekk?

Filsystem er egentlig database, men **brukere flest ser ikke dette**

- Filsystem har metadata som kan lese og lagres som en tabell
- Filanalyse kan gi opplisting av egenskaper som kan lagres som tabeller

Siard filer lages som oftest av databaser.

- Databaser kan lages av csv filer og excel m.m.
- Kommersielt verktøy kan lage siard av Excel filer
- Excel filer kan ha flere tabeller



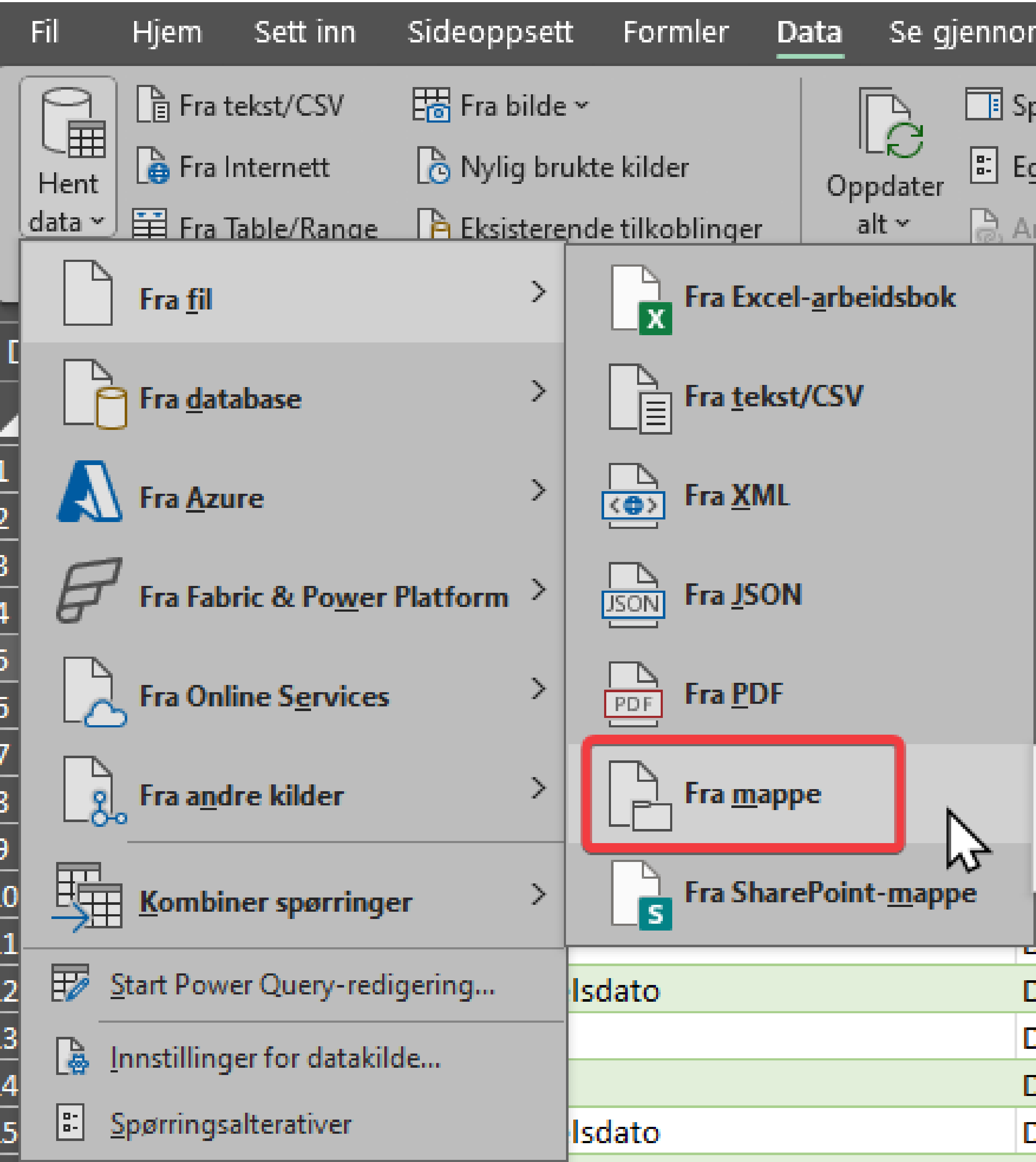
Excel er motoren

- Excel kan med få tastetrykk lese og dokumentere filer i en mappe og underliggende mapper
- Excel kan importere data
 - mappe på server
 - csv filer
 - databaser
 - Andre Excel filer
- Excel kan manipulere data – PowerQuery
- Tekstinnhold er generert med en blanding av libreOffice, copilot på håndskrift, Adobe Acrobat samt PowerAutomate for å automatisere lagring til txt filer
- TXT filene er slått sammen med metadata om filene i arkivet – via PowerQuery i Excel

Dataimport i Excel

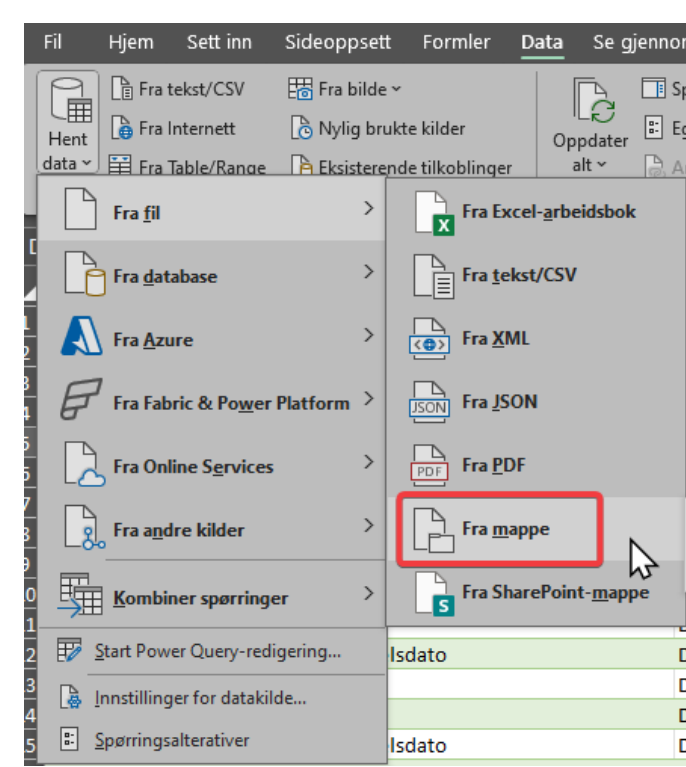
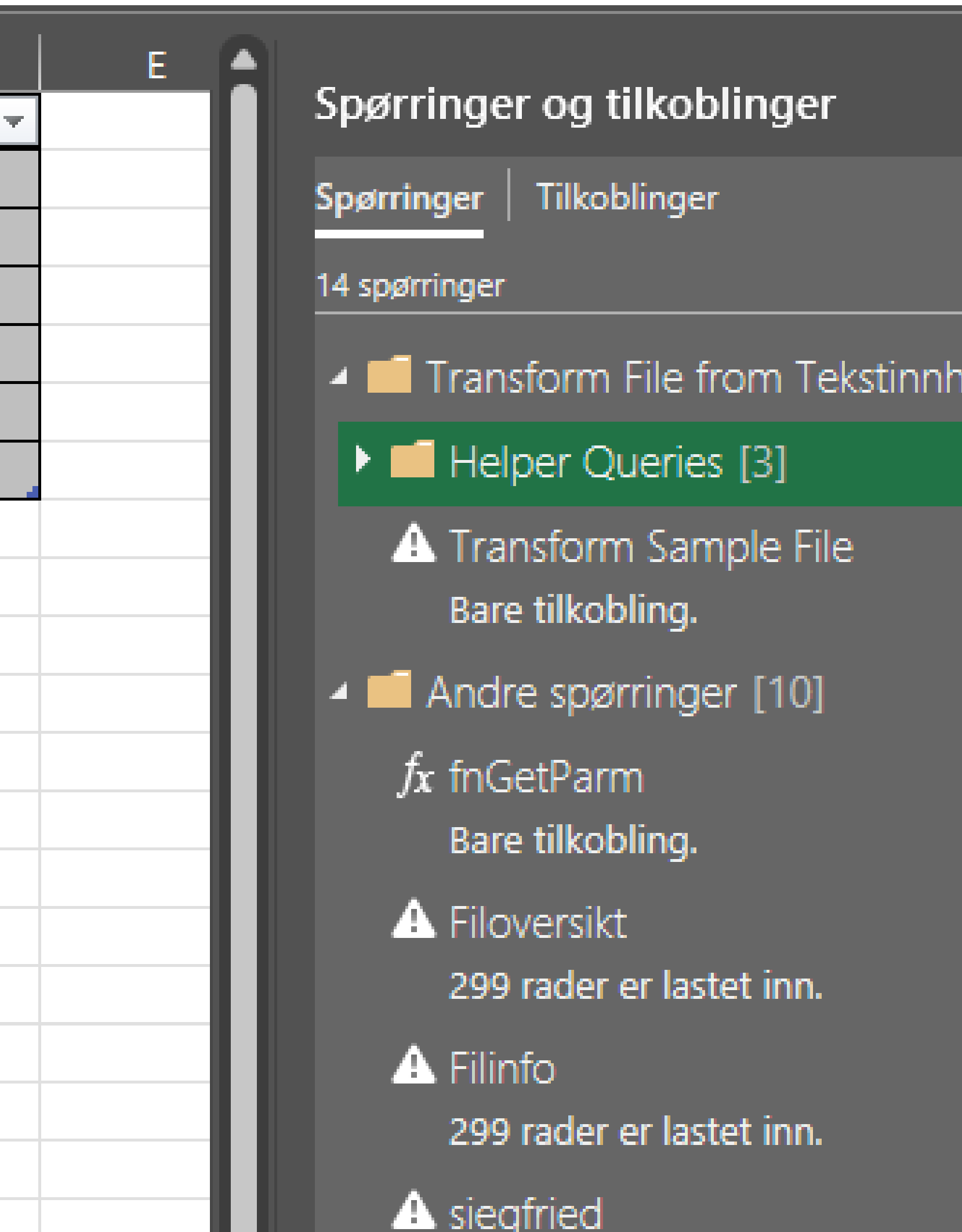
- Dette generer PowerQuery (script språket M)

	A	B	C	D	E
1	Parameter	Value	Kommentar		
2	GbFaktor	1073741824	<---- IKKE rør !		
3	MbFaktor	8388608	<---- IKKE rør !		
4	FiINavn	FiInfo.xlsx	<---- IKKE rør !		
8	Analyse	siegfried.csv	Navn på analyse fil (som unntas oversikt)		
9	Script	KjørSiegfried.cmd	Navn på script fil (som unntas oversikt)		
10	Root	D:\local\Temp\A-2713 - Pa 1456 - Vestviken blandakor\	<---- Sti med \ tilslutt		
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					



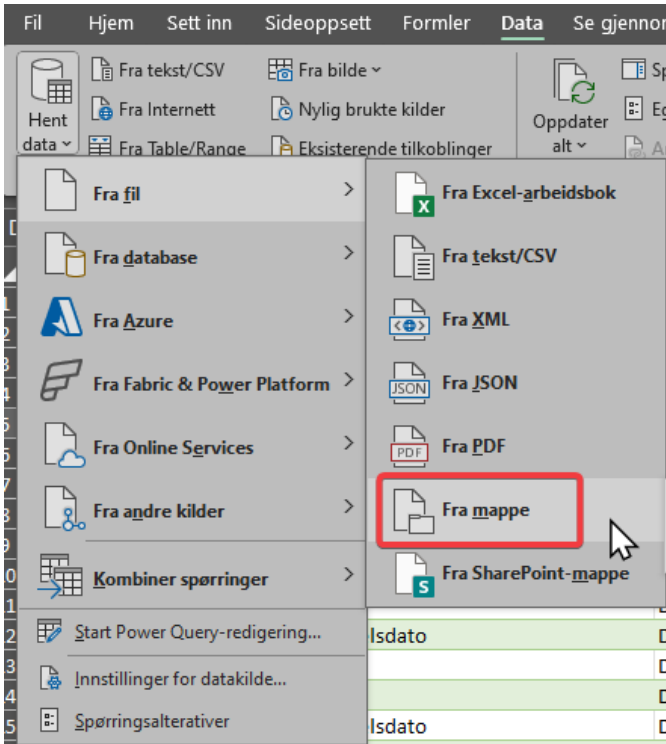
Dataimport i Excel

- Dette generer PowerQuery script (språket M)



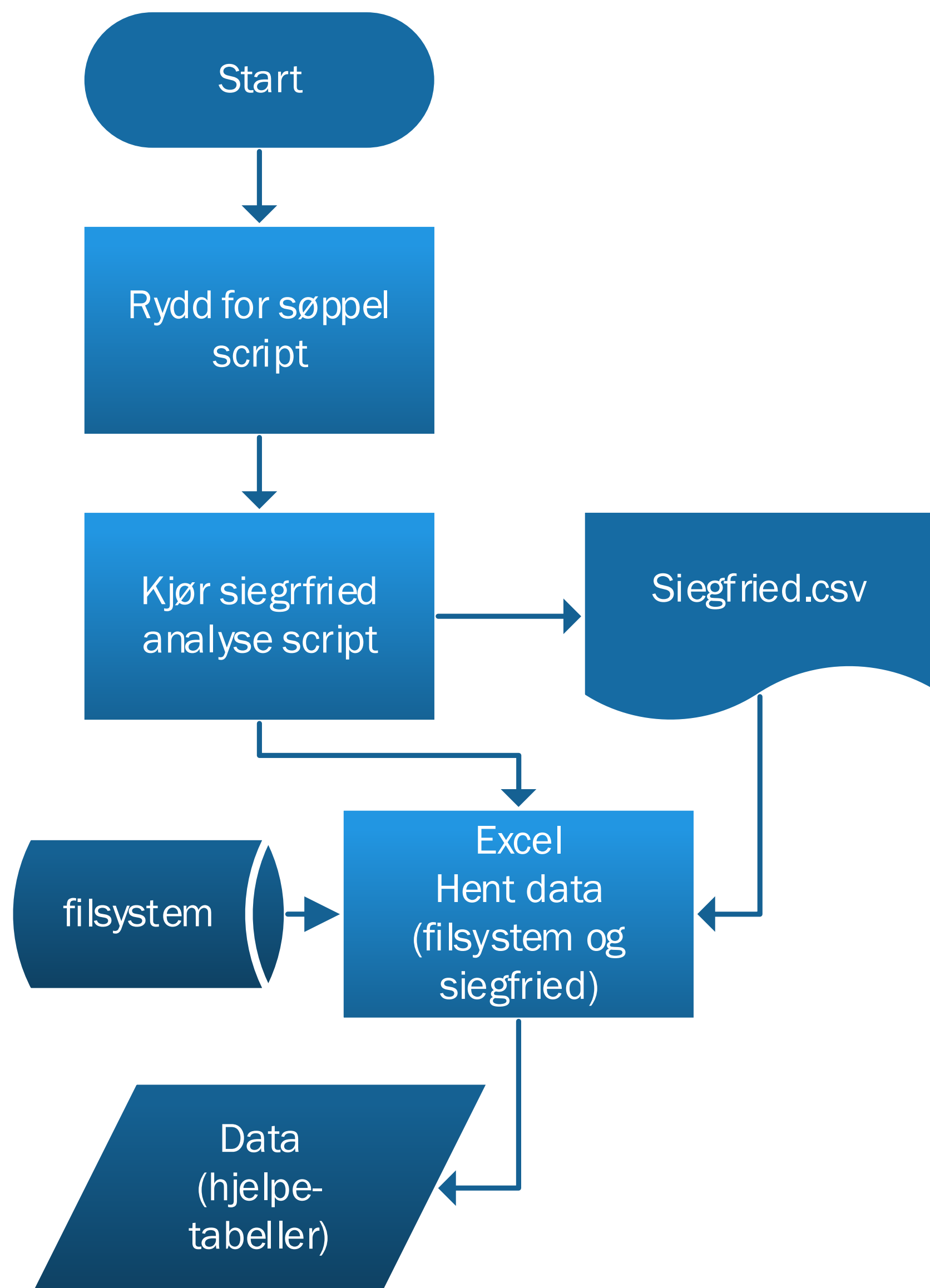
Dataimport i Excel

- Dette generer PowerQuery script (språket M)

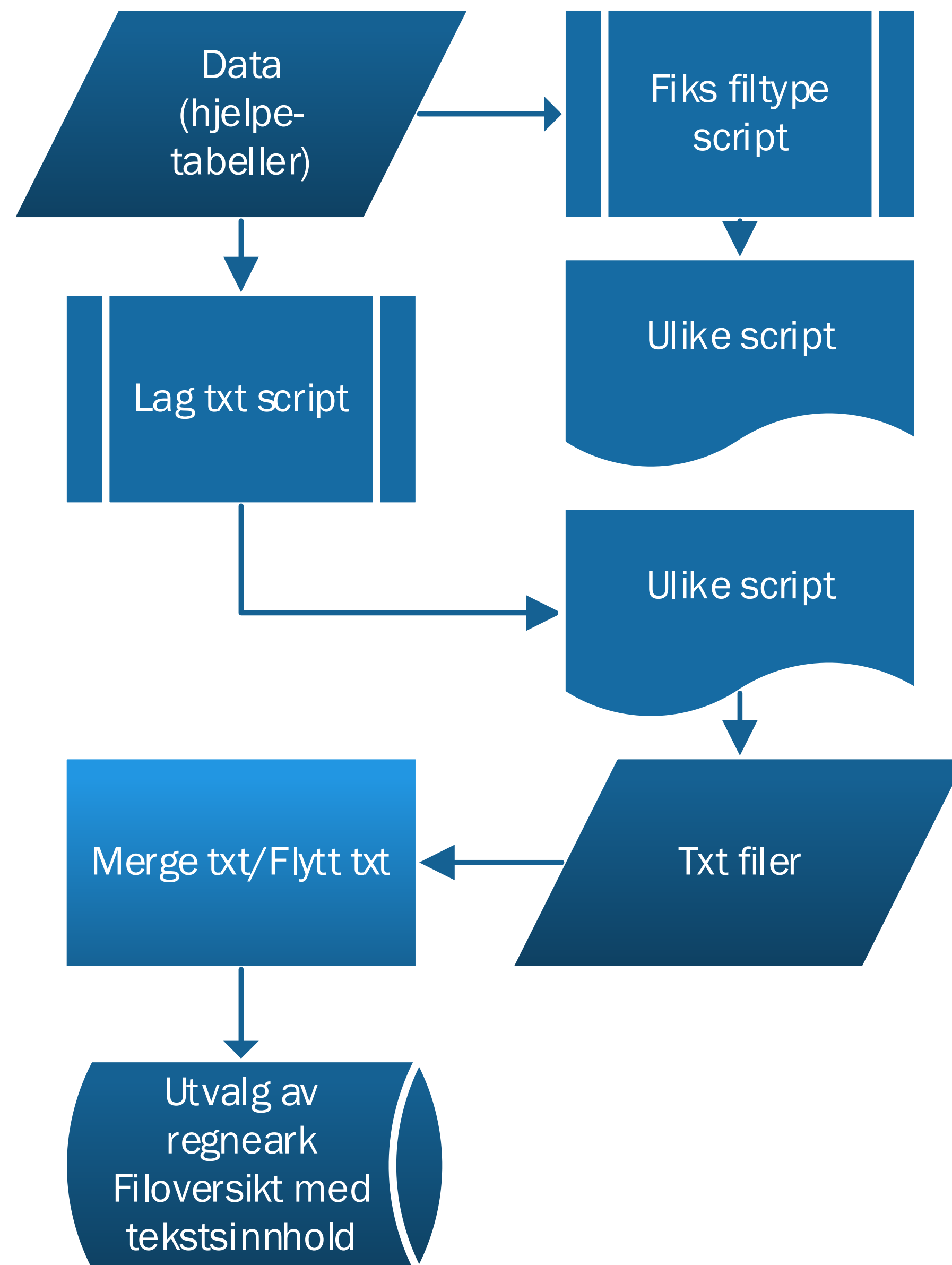
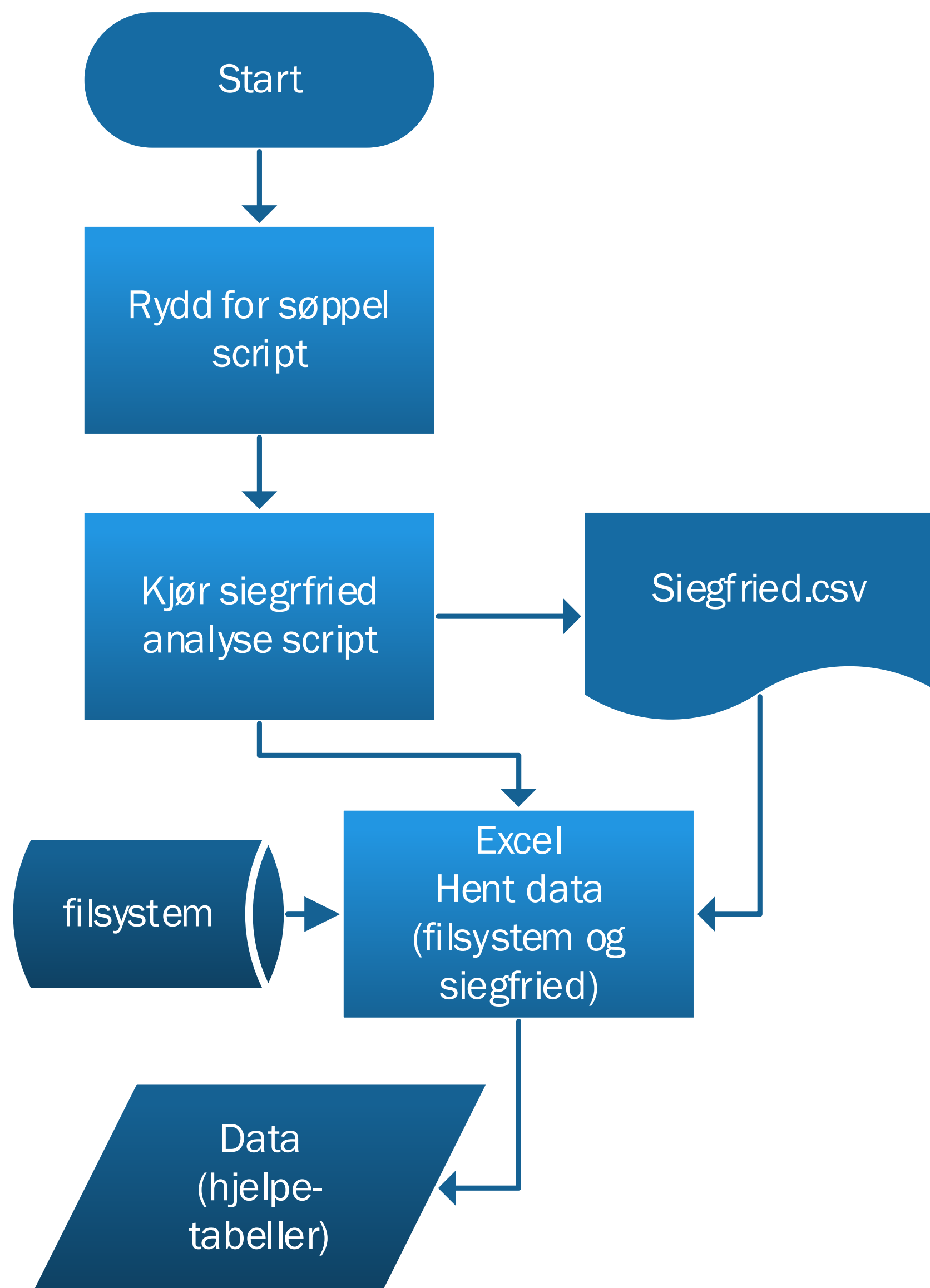


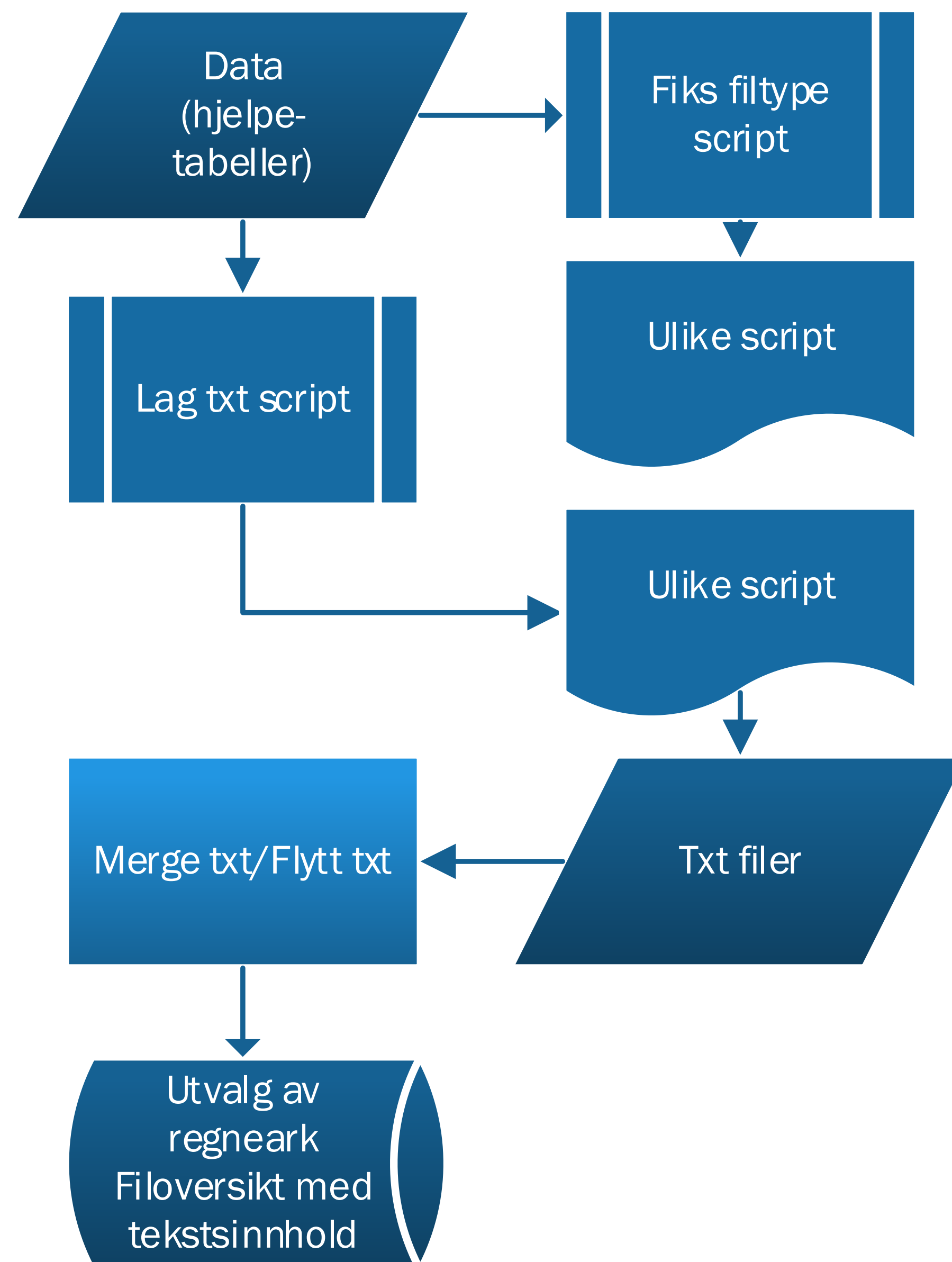
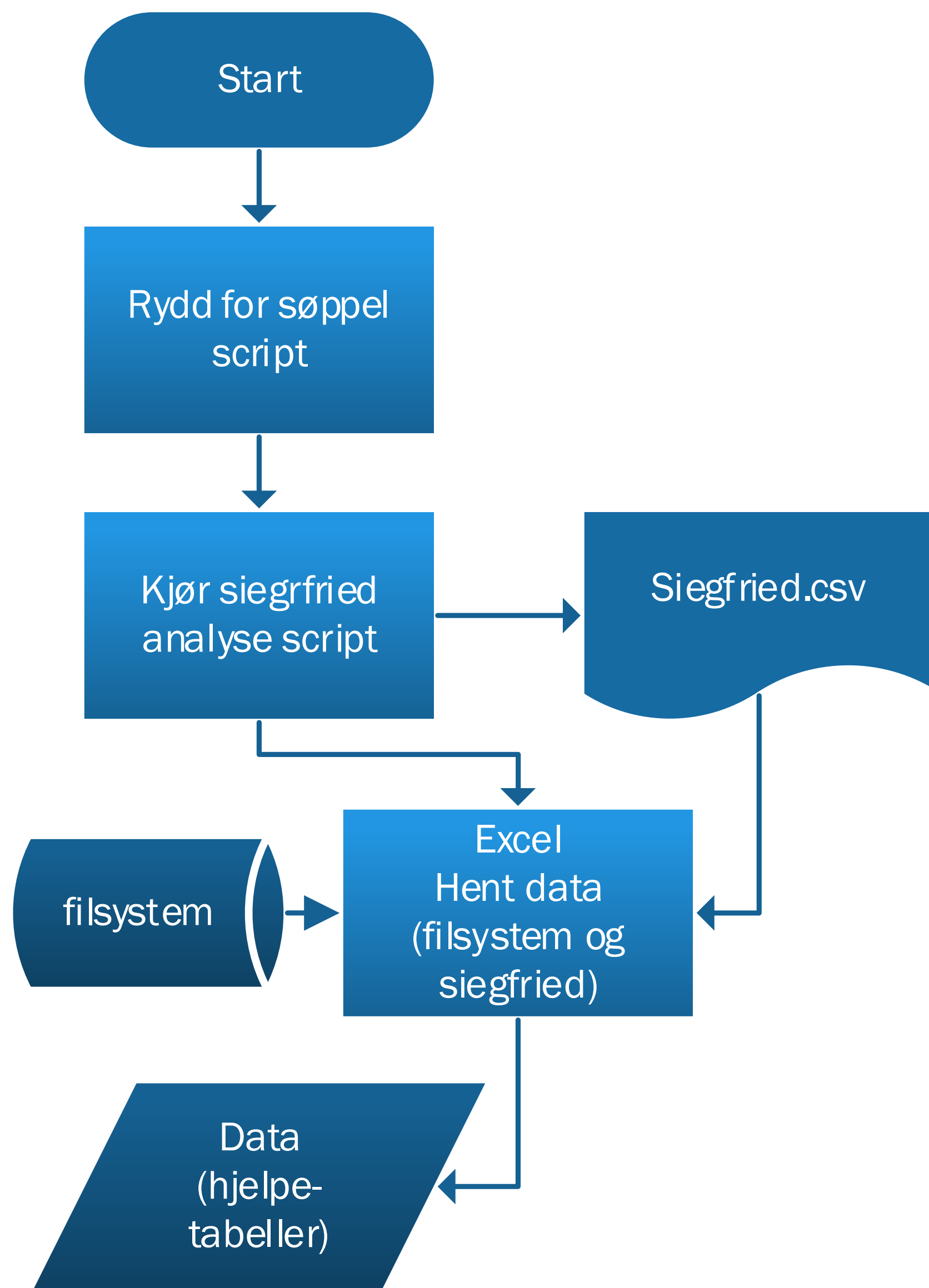
	A	B	C	D	E
1	Parameter ▼	Value ▼	Kommentar ▼		
2	GbFfaktor	1073741824	<---- IKKE rør !		
3	MbFaktor	8388608	<---- IKKE rør !		
4	FilNavn	FilInfo.xlsx	<---- IKKE rør !		
8	Analyse	siegfried.csv	Navn på analyse fil (som unntas oversikt)		
9	Script	KjørSigfried.cmd	Navn på script fil (som unntas oversikt)		
10	Root	D:\local\Temp\A-2713 - Pa 1456 - Vestviken blandakor\	<---- Sti med \ tilslutt		
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					

Prosesen

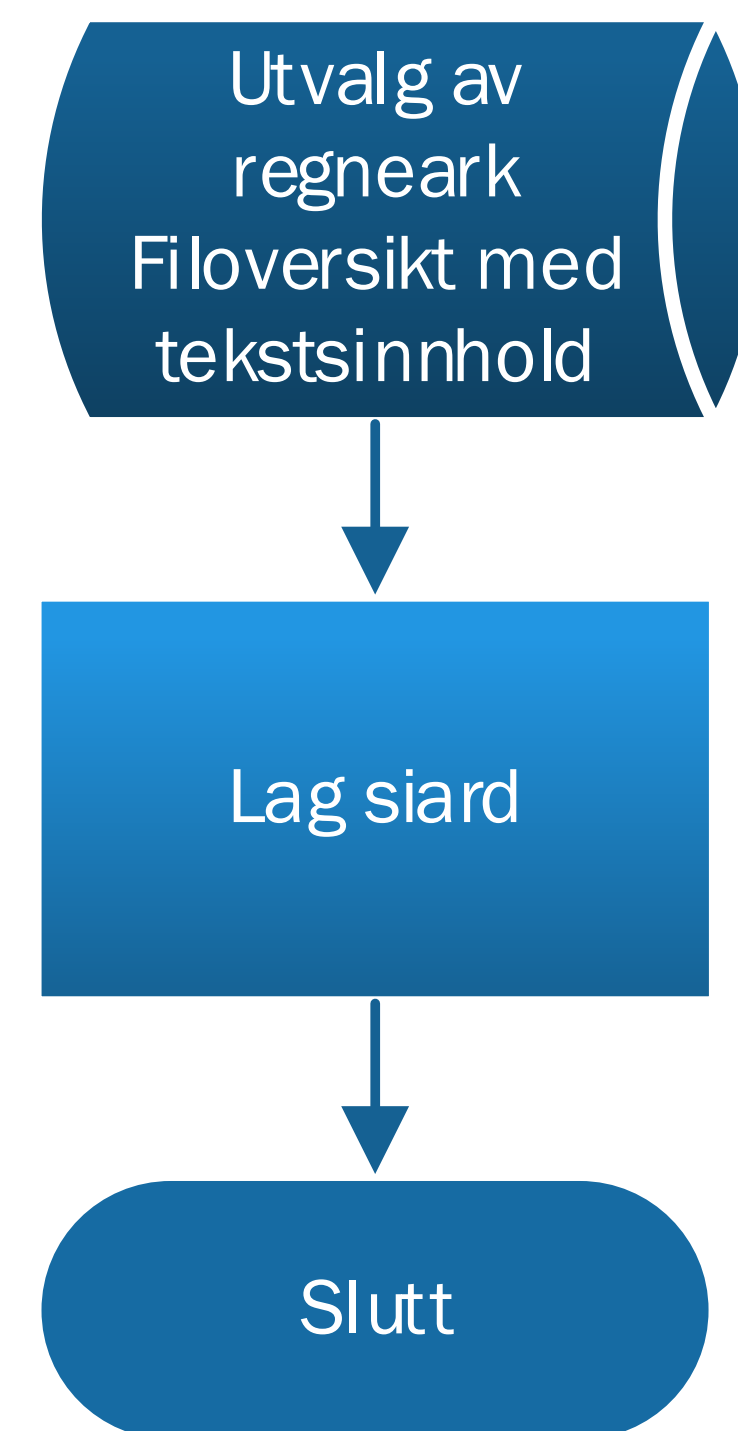


Prosesen





Prosessen



Metadata – filsystem (PowerQuery)

Name	Første medlemsliste vår 2014.pdf	Medlemsliste høst 2014.pdf	Medlemsliste 28.01.15.pdf
Extension	.pdf	.pdf	.pdf
Date accessed	04.09.2024 10:45	04.09.2024 10:45	04.09.2024 10:45
Date modified	12.10.2021 09:35	12.10.2021 09:58	12.10.2021 10:04
Date created	12.10.2021 09:35	12.10.2021 09:58	12.10.2021 10:04
Content Type	application/pdf	application/pdf	application/pdf
Mb	0,595811844	0,259535551	0,023075104
ReadOnly	FALSE	FALSE	FALSE
Hidden	FALSE	FALSE	FALSE
System	FALSE	FALSE	FALSE
Archive	TRUE	TRUE	TRUE

Metadata – lagt til

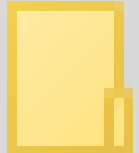
- Innhold – Innhold fra txt fil
- Filstørrelse (kb -> Mb / Gb)
- Splitter sti og filnavn i egne kolonner
- (Soketekst – utvalgte felt kjedesammen)

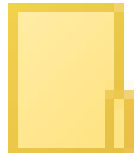
PowerAutomate – erfaringer

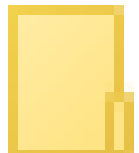
Dersom det finnes dedikerte komponenter i selve PowerAutomate prøv disse først.

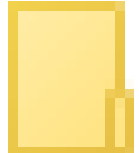
Man kan «fjernstyre» UI elementer. Adobe Acrobat er ett herk å fjernstyre. RPA er litt ustabilt (i hvert fall på Adobe Acrobat)

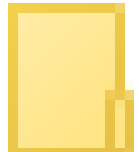
Case: Vestviken blandakor

 dokumenter

 Medlemslister

 Styret

 Referat styremøter

 Årsmøter

Medlemslister

Styremøter

Årsmøter

(Årvisse undermapper)

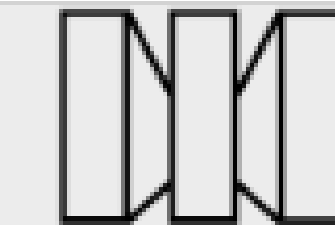
KDRS Søk & Vis

NO SE EN

Rolf (arkivar) ▼

Arkiver

Vestviken Blandakor > Filarkivet



Vestfold-
arkivet

@tc["title"]



Alle filer

Medlemslister

Styret

Årsmøter

↓ Last ned

Utskrift

2014

Søk

☐ Navn

Dato

Mb

Dokument

☐ 20140107 Referat.pdf

04.04.2025 10:09 0.011

Åpne

☐ 20140122 Referat.pdf

04.04.2025 10:09 0.009

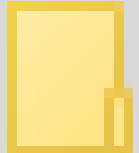
Åpne

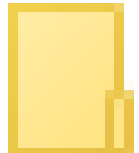
☐ 20140305 Referat.pdf

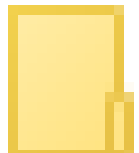
04.04.2025 10:09 0.008

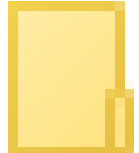
Åpne

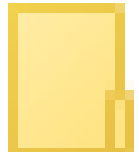
Case: Vestviken blandakor

 dokumenter

 Medlemslister

 Styret

 Referat styremøter

 Årsmøter

Medlemslister

Styremøter

Årsmøter

(Årvisse undermapper)

SV - Nerding

Bruker rubyview, eget søkefelt som søker i 3 ulike kolonner (Navn, Sti og Innhold)

```
Method: table(table, criteria, fields:, filter:, rows:)
```

```
parent =:@tc["title"]&.sub("alle filer","")+ "*"
```

```
navn_hits = table("filoversikt",  
  name: query,  
  parent:parent,  
  fields: fields,  
  rows: 10)
```

```
sti_hits = table("filoversikt",  
  relativpath: query,  
  parent:parent,  
  fields: fields,  
  rows: 10)
```

```
innhold_hits = table("filoversikt",  
  innhold: query,  
  parent:parent,  
  fields: fields,  
  rows: 10)
```

Feltet parent er overliggende sti (til fil i tabellen filoversikt). Formålene med anene skiller ut data fra tabellen filoversikt og gir subsett av data. @tc["title"] er navn på fanen, bruker denne til å filtrere data (criteria)

SV - Nerding

Bruker rubyview, eget søkefelt som søker 3 ulike kolonner (Navn, Sti og Innhold)

```
Method: table(table, criteria, fields:, filter:, rows:)
```

```
parent =:@tc["title"]&.sub("alle filer","",)+"*"
```

```
navn_hits = table("filoversikt",  
  name: query,  
  parent:parent,  
  fields: fields,  
  rows: 10)
```

Feltet parent er overliggende sti (til fil i tabellen filoversikt). Formålene med anene skiller ut data fra tabellen filoversikt og gir subsett av data. @tc["title"] er navn på fanen, bruker denne til å filtrere data (criteria)

SV - Nerding

Disse slås sammen og kun unike verdier lagres i @docs
sorteres etter navn (som senere vises)

```
@docs = (navn_hits + sti_hits + innhold_hits).uniq  
@docs = @docs.sort_by { |doc| doc["navn"] }
```


Hvor mye kode må til for å få opp data?

```
?xml version="1.0"?>
<!-- windows filesystem via PowerQuery in Excel -->
<views>
  <version>0.0.2</version>
  <description>views for filsystem</description>
  <view>
    <name>Filarkivet</name>
    <table>
      <name>filoversikt</name>
      <parent>top</parent>
      <search>>false</search>
      <title>alle filer</title>
      <primarykey>relativpath,name</primarykey>
      <fields>name as navn,modified as dato,mb,relativpath as sti,soketxt as tekst,innhold</fields>
      <rubyview>filer_sok</rubyview>
      <rows>10</rows>
      <export>print,download</export>
    </table>
    <table>
      <name>filoversikt</name>
      <parent>top</parent>
      <search>>false</search>
      <title>medlemslister</title>
      <rubyview>filer_sok</rubyview>
      <rows>10</rows>
      <export>print,download</export>
    </table>
  </view>
</views>
```

2 av fire faner

1/2 av koden

46 linjer «html» kode +

51 linje ruby kode

Hvor mye kode må til for å få opp data?

```
<table>
  <name>filoversikt</name>
  <parent>top</parent>
  <search>>false</search>
  <title>alle filer</title>
  <primaryKey>relativpath,name</primaryKey>
  <fields>name as navn, modified as dato, mb, relativpath as sti, soketxt as tekst, innhold</fields>
  <rubyview>filer_sok</rubyview>
  <rows>10</rows>
  <export>print,download</export>
</table>
```

```
<table>
  <name>filoversikt</name>
  <parent>top</parent>
  <search>>false</search>
  <title>medlemslister</title>
  <rubyview>filer_sok</rubyview>
  <rows>10</rows>
  <export>print,download</export>
</table>
```

Siard svarer ut behov i næringslivet?

- Sikkerhet
 - Bruker tilganger
 - Views (Malstyrt visning av utvalgte data)
 - Logging av bruk
- Bærekraft
- Enkelt
 - Metakatalog (med eller uten innhold)

Siard metoden

Kan siard brukes for fil og mappe arkiv?

Hva hvis?

Søk: «2014» (alle filer)

- Jeg har **valgt** at **søk** skal skje på følgende metadata:
 - filnavn
 - katalognavn
 - innhold.
- Alle filer med tekst er forsøkt mediekonvert til tekst og dermed gjort søkbare
- Velger å vise:
 - filnavn
 - fildato
 - størrelse
- Utskrift viser søkeresultat
- Nedlastning for avkryssede filer

The screenshot shows the 'KDRS Søk & Vis' web application. The header is blue with the title 'KDRS Søk & Vis' and a user profile 'Rolf (arkivar)' with a dropdown arrow. Below the header, there's a breadcrumb trail: 'Arkiver' > 'Vestviken Blandakor' > 'Filarkivet'. To the right of the breadcrumb is the 'Vestfold-arkivet' logo. A navigation bar contains links: '<', 'Alle filer' (underlined), 'Medlemslister', 'Styret', and 'Årsmøter'. On the right side of the navigation bar are two green buttons: 'Last ned' and 'Utskrift'. Below the navigation bar, there's a search input field containing '2014' and a 'Søk' button. A table displays the search results with columns: 'Navn', 'Dato', 'Mb', and 'Dokument'. The table lists three PDF files from 2014, each with a checkbox and an 'Åpne' button. A mouse cursor is pointing at the 'Åpne' button for the first file.

☐	Navn	Dato	Mb	Dokument
☐	20140107 Referat.pdf	04.04.2025 10:09	0.011	Åpne
☐	20140122 Referat.pdf	04.04.2025 10:09	0.009	Åpne
☐	20140305 Referat.pdf	04.04.2025 10:09	0.008	Åpne

Søk: «2014» (alle filer)

Mer sikkert?

- Data er mer skjermet siden de ligger på en dedikert server og filene er ikke direkte tilgjengelig via nettverk
- «Alle filer» viser ikke noe før man har søkt på noe. Filene eksponeres i mindre grad.
- «Åpne» gir PDF dokumenter i nettleser laster ned om filtype ikke har «visning» i nettleser.
- 3 serier:
 - Medlemslister
 - Styre(referat)
 - Årsmøter

The screenshot shows the 'KDRS Søk & Vis' web application. At the top, there's a blue header with the title 'KDRS Søk & Vis' and language options 'NO SE EN'. Below the header, a grey bar contains 'Arkiver' and a breadcrumb 'Vestviken Blandakor > Filarkivet'. On the right, a user profile 'Rolf (arkivar)' is shown. Below this, a navigation bar has tabs: 'Alle filer' (selected), 'Medlemslister', 'Styret', and 'Årsmøter'. To the right of the tabs are two green buttons: 'Last ned' and 'Utskrift'. A search bar contains the text '2014' and a 'Søk' button. Below the search bar is a table with columns: 'Navn', 'Dato', 'Mb', and 'Dokument'. The table has one row with the entry '20140107 Referat.pdf', date '04.04.2025 10:09', size '0.011', and an 'Åpne' button. A mouse cursor is pointing at the 'Åpne' button.

Navn	Dato	Mb	Dokument
20140107 Referat.pdf	04.04.2025 10:09	0.011	Åpne

All bruk logges

Arkiver

Logger

Time	UserID	Ip	Action	Level	Collection Id	Collection Title	Table	Row	Message
2025-04-28 02:33	rolf.holte@vestfoldmuseene.no	10.0.2.2	view	info	b293beb6-4944-4bd4-95c8-c1f9a0ba4a51	Vestviken Blandakor	alle filer	ruby view filer_sok	
2025-04-28 02:33	rolf.holte@vestfoldmuseene.no	10.0.2.2	download	info	b293beb6-4944-4bd4-95c8-c1f9a0ba4a51	Vestviken Blandakor			params: {"download_info"=>#<ActionCont permitted: false>, "query"=>"*xls*", "contr "id"=>"3", "vi"=>"0", "ti"=>"0"}
2025-04-28 02:33	rolf.holte@vestfoldmuseene.no	10.0.2.2	view	info	b293beb6-4944-4bd4-95c8-c1f9a0ba4a51	Vestviken Blandakor	alle filer	ruby view filer_sok	
2025-04-28 02:33	rolf.holte@vestfoldmuseene.no	10.0.2.2	view	info	b293beb6-4944-4bd4-95c8-c1f9a0ba4a51	Vestviken Blandakor	alle filer	ruby view filer_sok	

Bærekraftperspektiv - Veien videre?

- Om man dokument innholdet er bevart som ren tekst, trenger vi pene «store filer» i PDF, eller produksjonsformater ?
- Mediekonvertering bør bare være ren tekst?
- Ved å gi slipp på utseende og deler av konteksten reduseres lagringsbehov og strømforbruk etc
- OBS Det er ingen garanti at OCR eller håndskriftsgjenkjenning blir korrekt

Veien videre - Kortsiktigmål

- Lagre maler
 - KDSRS søk og vis
 - Decom
- Forenkle og dokumentere prosessene
 - Ha en oppskrift som alle kan følge (uten IT bakgrunn).
- Standardisere slik at excel arket gir enkle oppgaver
 - Oppgaver (script) som kan utføres av alle (Foranalyse + enkel feil retting)
- Mere automasjon og KI?

MÅL: Finne bedre produksjonsløype for privatarkiv, slik at «vanlige arkiver» bidrar

