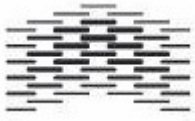
Hva er datakvalitet?

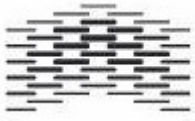
Hvordan skal arkivtjenesten forholde seg til det?

Thomas Sødring
Høyskolen i Oslo og Akershus
thomas.sodring@hioa.no
99570472



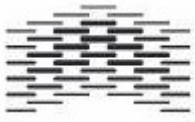
Bakgrunn

- Flere og flere depot institusjoner gjør seg klare til å ta imot elektroniske avleveringer
  - Hva vet depot om disse avleveringer?
  - Hvordan skal disse forvaltes? DIAS
  - Er det i det helt tatt mulig å bruke disse avleveringer?
- Flere stiller spørsmålet «*Kan vi måle kvaliteten på elektronisk materiale?*»



Datakvalitet

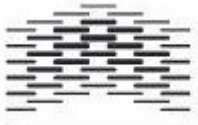
- Datakvalitet som et eget forskningsfelt har eksistert siden 1970 tallet
  - men det var etter 2000 tallet at flere og flere ble interessert i fagfeltet
    - Dette pga en eksplosjon i mengden av elektronisk data som ble generert
    - Hvordan dataene ble (og fortsatt blir) *håndtert på en ustrategisk* måte av mange selskaper



Hva er datakvalitet?

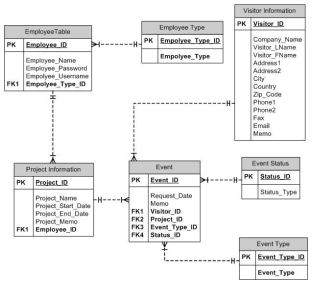
- Vi prøver på en definisjon

Datakvalitet angir i hvilken grad data i et system er i overensstemmelse med det virkelige scenarioet dataen representerer og er brukbar



Systemutvikling

modellering → implementasjon → bruk

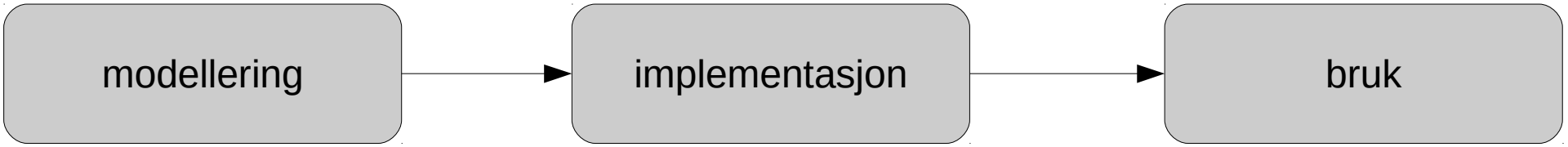


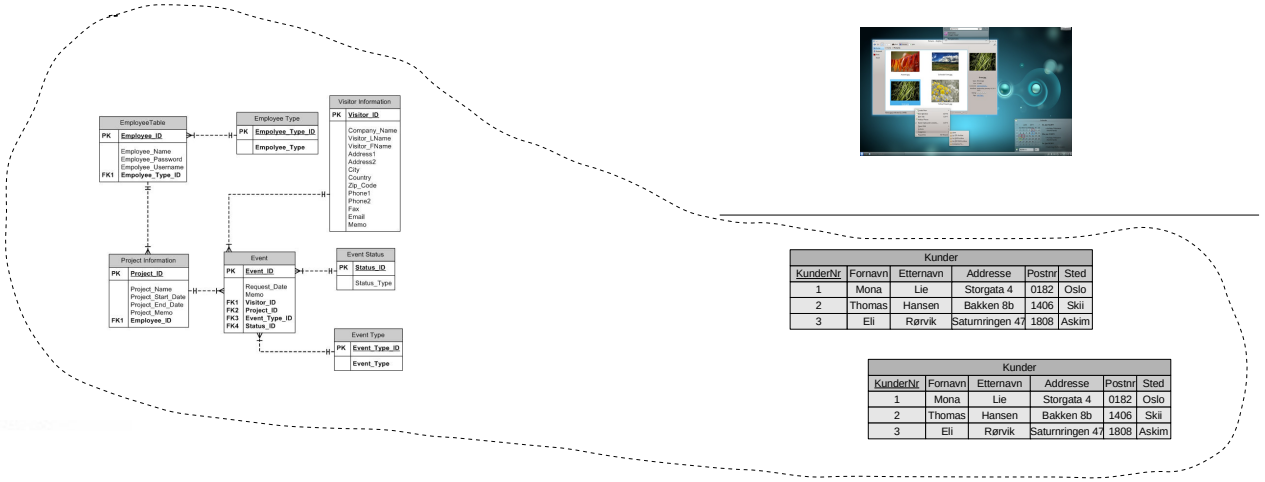
Kunder					
KunderNr	Fornavn	Etternavn	Addresse	Postnr	Sted
1	Mona	Lie	Storgata 4	0182	Oslo
2	Thomas	Hansen	Bakken 8b	1406	Skii
3	Eli	Rørvik	Saturnringen 47	1808	Askim

Kunder					
KunderNr	Fornavn	Etternavn	Addresse	Postnr	Sted
1	Mona	Lie	Storgata 4	0182	Oslo
2	Thomas	Hansen	Bakken 8b	1406	Skii
3	Eli	Rørvik	Saturnringen 47	1808	Askim

Hva er datakvalitet?

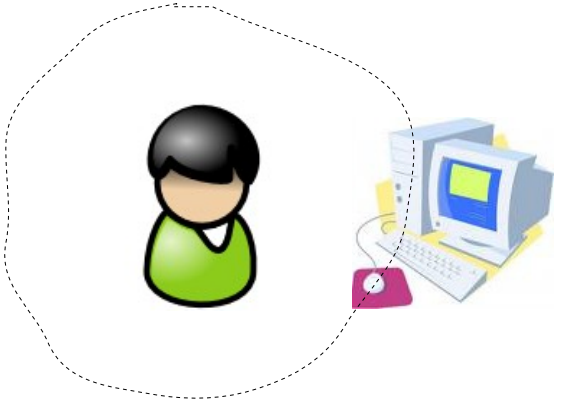
Datakvalitet angir i hvilken grad ***data i et system*** er i ***overensstemmelse*** med det virkelige ***scenarioet dataen representerer*** og er ***brukbar***





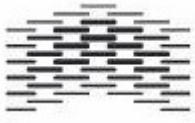
Kunder					
KunderNr	Fornavn	Etternavn	Addresse	Postnr	Sted
1	Mona	Lie	Storgata 4	0182	Oslo
2	Thomas	Hansen	Bakken 8b	1406	Skii
3	Eli	Rørvik	Saturnringen 47	1808	Askim

Kunder					
KunderNr	Fornavn	Etternavn	Addresse	Postnr	Sted
1	Mona	Lie	Storgata 4	0182	Oslo
2	Thomas	Hansen	Bakken 8b	1406	Skii
3	Eli	Rørvik	Saturnringen 47	1808	Askim



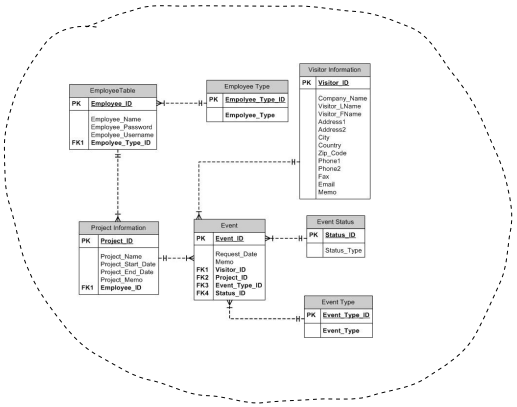
Hvordan oppstår dårlig datakvalitet

- Manglende datakvalitet er ofte et resultat av
  - *dårlig system design* eller
    - Scenarioet system representerer er ikke tilstreklig modellert
  - forbundet med *dårlige prosedyrer* ved data innførsel
    - Da er datakvalitet er en form av ***god arkivdanning***!



Hva er datakvalitet?

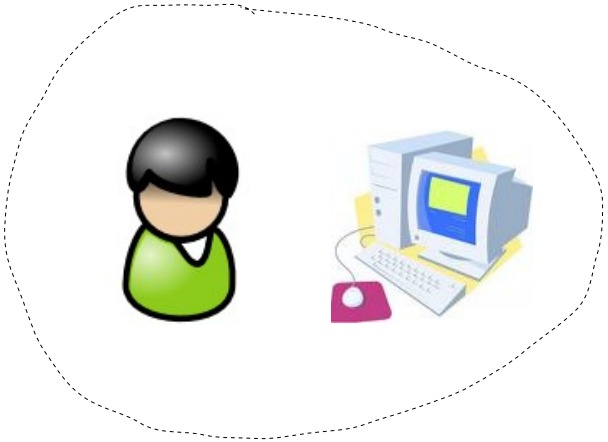
- *dårlig system design*
- *dårlige prosedyrer* ved data innførsel



dårlig system design

Kunder					
KunderNr	Fornavn	Etternavn	Addresse	Postnr	Sted
1	Mona	Lie	Storgata 4	0182	Oslo
2	Thomas	Hansen	Bakken 8b	1406	Skii
3	Eli	Rørvik	Saturnringen 47	1808	Askim

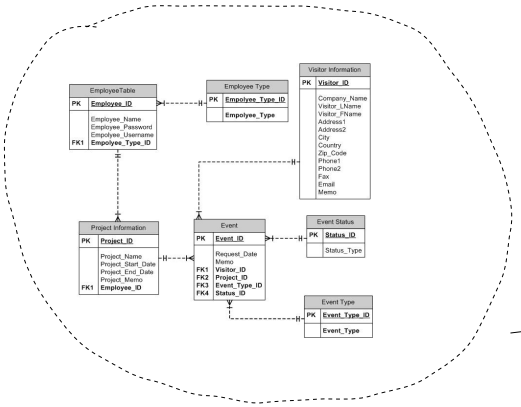
Kunder					
KunderNr	Fornavn	Etternavn	Addresse	Postnr	Sted
1	Mona	Lie	Storgata 4	0182	Oslo
2	Thomas	Hansen	Bakken 8b	1406	Skii
3	Eli	Rørvik	Saturnringen 47	1808	Askim



dårlige prosedyrer

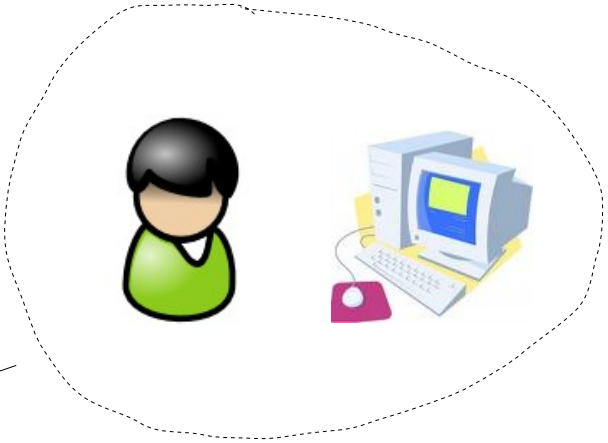
Hva er datakvalitet?

- *dårlig system design*
- *dårlige prosedyrer* ved data innførsel



Kunder					
KunderNr	Fornavn	Etternavn	Addresse	Postnr	Sted
1	Mona	Lie	Storgata 4	0182	Oslo
2	Thomas	Hansen	Bakken 8b	1406	Skii
3	Eli	Rørvik	Saturnringen 47	1808	Askim

Kunder					
KunderNr	Fornavn	Etternavn	Addresse	Postnr	Sted
1	Mona	Lie	Storgata 4	0182	Oslo
2	Thomas	Hansen	Bakken 8b	1406	Skii
3	Eli	Rørvik	Saturnringen 47	1808	Askim



dårlig system design

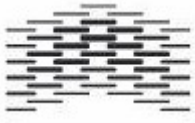
dårlige prosedyrer

Datakvalitetsproblemet manifesterer seg her

måles også her!
(til en hvis grad)

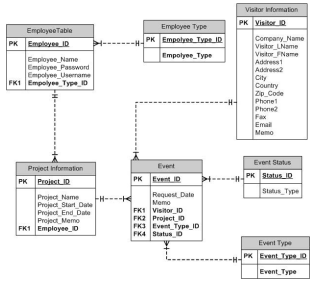
DK ved danning eller bevaring?

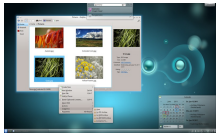
- Er datakvalitet noe som vi skal være opptatt av under danning eller bevaring?
- Hvordan oppleves datakvalitets problematikken i hver av fasene?
- Når depot overtar et uttrekk er det praktisk talt umulig å rette på kvaliteten
- Hvorfor tar ingen ansvar for datakvalitet?
  - Det står ikke noe om det i Noark standarden
  - Data er låst i leverandørenes systemer



DK ved danning eller bevaring?

modellering → implementasjon → bruk





Kunder

KunderNr	Fornavn	Etternavn	Addresse	Postnr	Sted
1	Mona	Lie	Storgata 4	0182	Oslo
2	Thomas	Hansen	Bakken 8b	1406	Skii
3	Eli	Rørvik	Saturnringen 47	1808	Askim

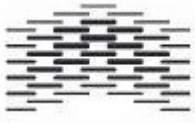
Kunder

KunderNr	Fornavn	Etternavn	Addresse	Postnr	Sted
1	Mona	Lie	Storgata 4	0182	Oslo
2	Thomas	Hansen	Bakken 8b	1406	Skii
3	Eli	Rørvik	Saturnringen 47	1808	Askim

Et uttrekk blir skapt fra data i databasen

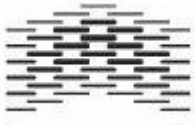
DK ved danning eller bevaring?

- Depot vil arve DK fra danningsystemene
  - DK er noe depot bør være opptatt av
  - Danning har jo bare en 5 års perspektiv på kvalitet. Bevaring har ??
- Ja takk begge deler!
  - Mål og fiks datakvalitet underveis, la DK inngå i avleveringer
    - Kan depot nekte å ta imot et uttrekk pga dårligkvalitet?
    - Hvis den validerer mot DTD/XSD?



Nekte å ta i mot avleveringer?

- Statlig nivå
  - Riksarkivet vil sjekke at filene valideres mot en skjema
  - Teste med arkn4 / URD
  - Hvilken hjemmel kan Riksarkivet bruke til å nekte å ta imot en avlevering med dårlig datakvalitet
- IKA nivå
  - IKA er eid av de som skal levere, vanskeligere å nekte å ta imot
  - IKA bør måle kvalitet



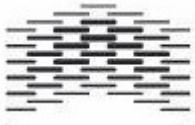
Datakvalitet

- Datakvalitet måles i datakvalitetsdimensjoner
- Det finnes mange datakvalitetsdimensjoner. Ofte grupperer de:
  - Subjektive
  - Objektive
  - Prosesss (ser ikke på det)
- De kan også grupperes på forskjellige måter*

*2 artikkler
http://www.emeraldinsight.com/journals.htm?articleid=841292&show=html
http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.144.4212

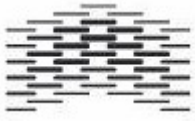
Subjektivt Datakvalitet

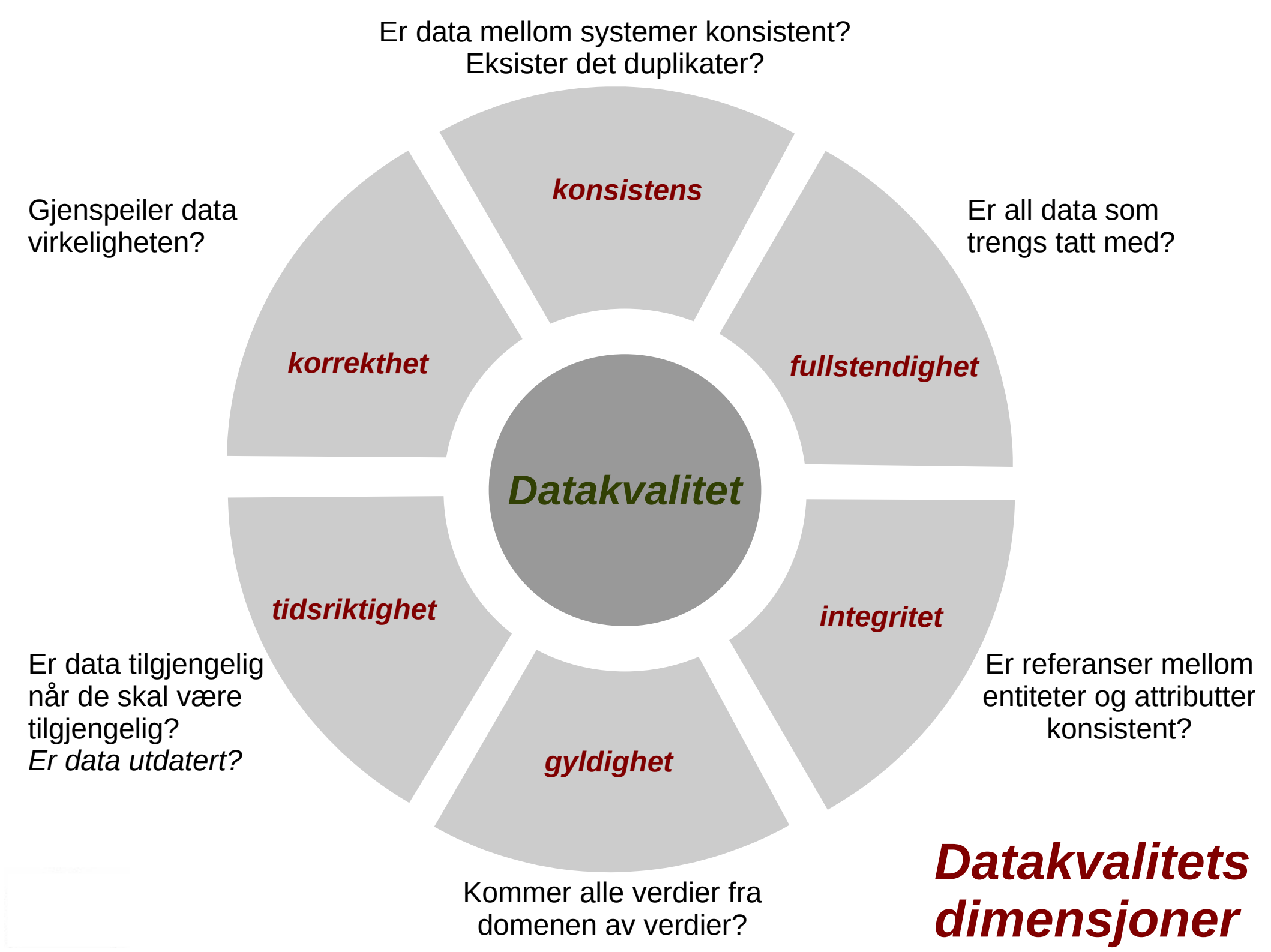
- Subjektive datakvalitets dimensjoner ofte basert på intervjuteknikker og svarene lar seg ofte ikke generalisere
- En saksbehandler kan ha en forståelse at dataen i saksbehandlingsystemet er av høy kvalitet utifra sitt perspektiv som bygger på mange års erfaring og bruk av systemet
  - En annen saksbehandler kan se på de samme dataene og mene at kvaliteten er dårlig
- Et eksempel av denne typen datakvalitet er forståelighets dimensjonen, dvs hvor enkelt det er å forstå informasjonen



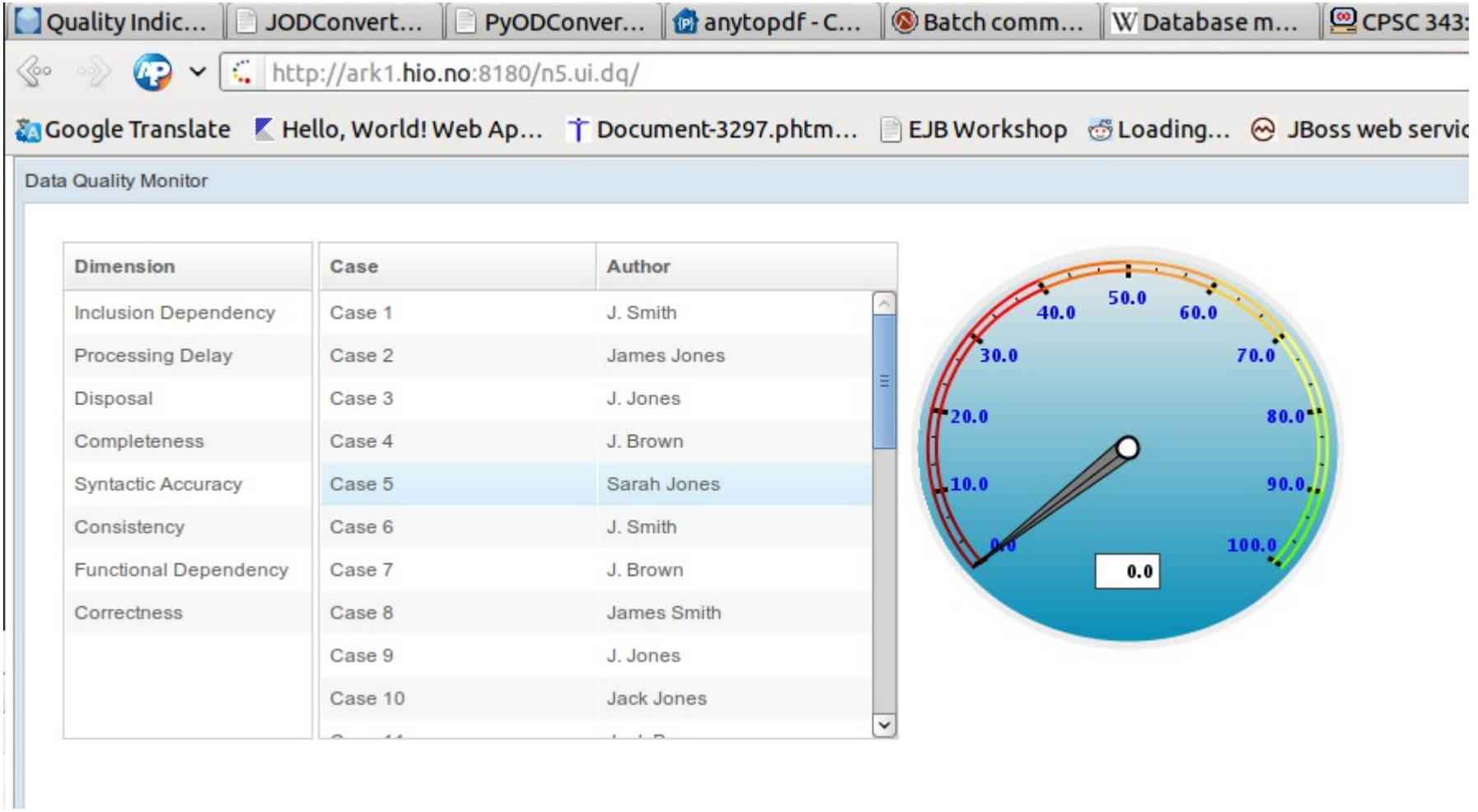
Objektivt Datakvalitet

- Objektive datakvalitets dimensjoner vurderes gjennom en uavhengig analyse av dataen i systemet (databasen) og er ofte basert på programvare som går inn i databasen og måler kvaliteten i tabellene
  - Et eksempel på en slik dimensjon er fullstendighet eller om all relevant data er innhentet og lagret
  - Med uavhengighet menes hvis målingen kjøres to ganger over samme database vil du få samme resultat



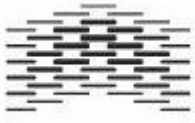

Er data mellom systemer konsistent?
Eksister det duplikater?
konsistens
Gjenspeiler data virkeligheten?
korrekthet
Er all data som trengs tatt med?
fullstendighet
Datakvalitet
tidsriktighet
Er data tilgjengelig når de skal være tilgjengelig?
Er data utdatert?
integritet
Er referanser mellom entiteter og attributter konsistent?
gyldighet
Kommer alle verdier fra domenen av verdier?
Datakvalitets dimensjoner

Prototype datakvalitets GUI



Datakvalitet og Noark

- Noark indirekte sørger for et minimum av datakvalitet
  - Det er ingen eksplosjon av elektronisk informasjon brukt på en ustrategisk måte i offentlig sektor
    - Eller?



Prosjekt om datakvalitet

- Kan vi *kvantifisere* «kvalitet» på en *objektiv måte*
  - Kan arkivfeltet kan dra nytte av datakvalitet
  - Subjektiv kan også være nyttig