Veileder for beskrivelse av kvalitet på datasett – kvantifiserbar kvalitet

Digitaliseringsdirektoratet (Digdir) / The Norwegian Digitalisation Agency

Publisert: 2020-02-20
Oppdatert: 2023-02-14 (definisjonene er tatt ut fra veilederen og lenket til der de finnes)

Innmelding av feil og mangler:
Dersom du finner feil eller mangler i dokumentet, ber vi om at dette meldes inn på Github Issues. Dersom du ikke allerede har bruker på Github kan du opprette bruker gratis.

Innholdsfortegnelse

Innledning

Formål, omfang og avgrensninger

Formålet med dette dokumentet er å gi veiledning i felles definisjoner på og måleverktøy for kvantifiserbar kvalitet på datasett. Et felles sett med definisjoner og måleverktøy vil skape en unison beskrivelse og forståelse av datakvalitet på tvers av virksomheter. Hensikten er ikke å utelukke andre måter å beskrive datakvalitet på enn disse presentert i dette dokumentet. Man står fritt til å supplere med andre typer kvalitetsmål i kvalitetsbeskrivelsen.

I henhold til Spesifikasjon for beskrivelse av kvalitet på datasett, er det flere ulike måter å beskrive kvalitet på et datasett på:

- kvantifiserbar kvalitet
- ikke-kvantifiserbar kvalitet
- kvalitet i samsvar med gitt(e) standard(er)/spesifikasjon(er)
- brukertilbakemeldinger knyttet til kvalitet

Spesifikasjonen foreslo også å predefinere kvalitetsdimensjoner, kvalitetsdeldimensjoner og kvalitetsmål. Disse er publisert som egne kontrollerte vokabularer, ett for kvalitets(del)dimensjoner og ett for kvalitetsmål.

Dette er ikke en veileder for datakvalitetsarbeidet generelt. Veilederen forklarer hvordan datakvalitet beskrives, med fokus på å bruke predefinerte kvantifiserbare kvalitetsmål. Ved behov vil det bli utarbeidet tilsvarende veiledere for de andre måter å beskrive kvalitet på. Disse vil inngå i Rammeverk for informasjonsforvaltning.

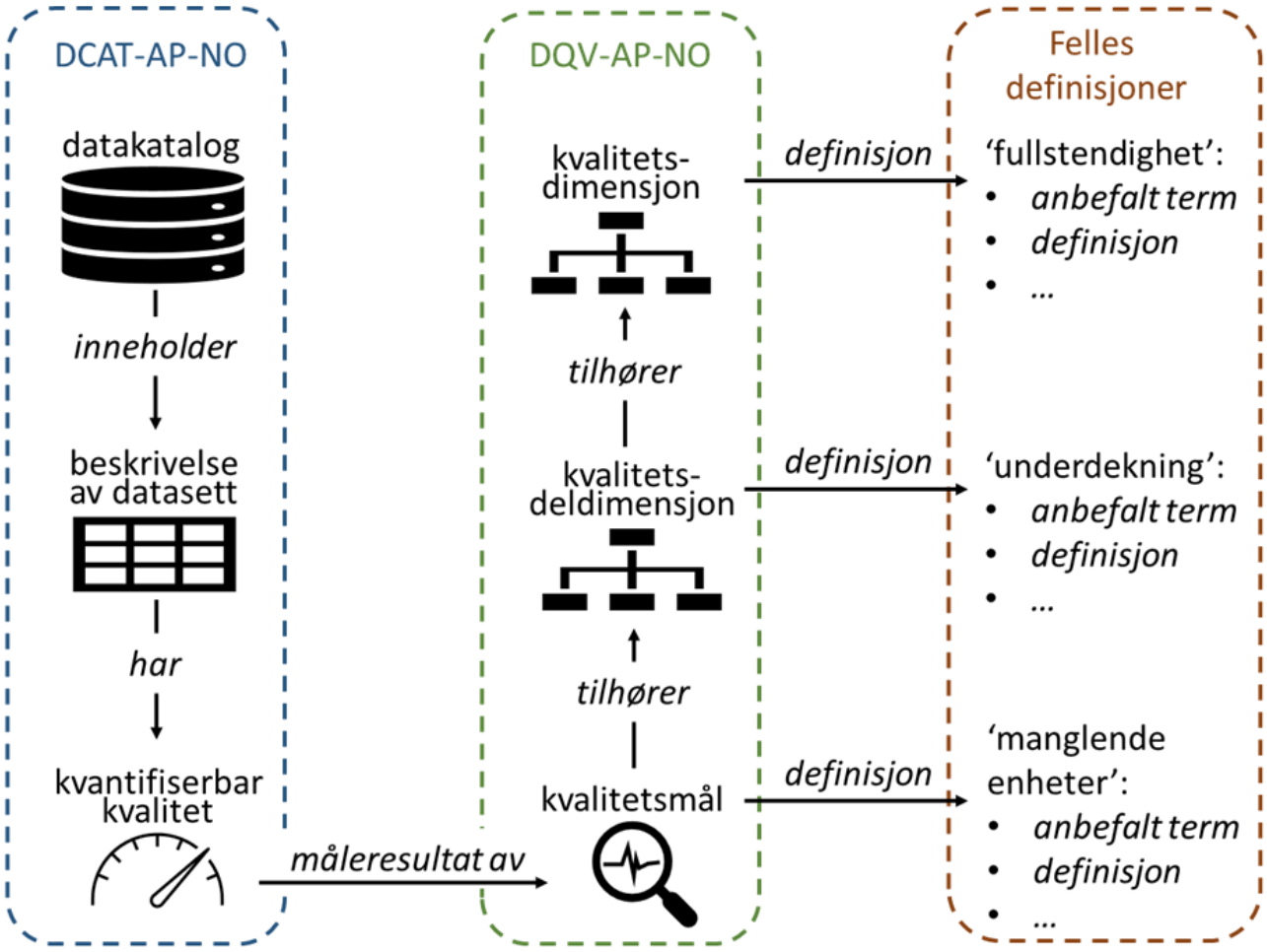
Målgruppe

Målgruppen for veilederen er:

- de som skal beskrive datasett og derunder kvantifiserbar kvalitet på datasett (primær målgruppe)
- de som leverer løsninger og verktøy for registrering og visning av denne type kvalitetsmål (primær målgruppe)
- de som skal forstå beskrivelse av kvantifiserbar kvalitet på datasett (sekundær målgruppe)

Sammenheng mellom relevante standarder, spesifikasjoner og denne veilederen

Veilederen må ses i sammenheng med DQV-AP-NO (Norsk applikasjonsprofil av DQV).



Tegningen ovenfor illustrerer sammenhengen mellom ulike standarder/spesifikasjoner og felles definisjoner:

- DCAT-AP-NO - Standard for beskrivelse av datasett, datatjenester og datakataloger: Denne spesifiserer bl.a. hvordan datasett beskrives, deriblant hvordan en beskrivelse av kvantifiserbar kvalitet knyttes til beskrivelsen av et datasett.
- DQV-AP-NO - Norsk applikasjonsprofil av DQV: Denne spesifiserer hvordan kvalitet på datasett beskrives, bl.a. å knytte kvalitetsbeskrivelse til kvalitetsdimensjon.
- Felles definisjoner: Hensikten med felles definisjoner er å skape et felles vokabular for kvalitetsbeskrivelse, slik at beskrivelsene forstås likt. Dette handler om felles definisjoner og forståelse av kvalitetsdimensjoner og kvalitetsdeldimensjoner, og også kvalitetsmål når det gjelder kvantifiserbar kvalitet.

Denne veilederen forklarer predefinerte kvalitetsmål for kvantifiserbar kvalitet, samt de kvalitetsdeldimensjonene og kvalitetsdimensjonene disse kvalitetsmålene tilhører.

Hensyn man bør ta når man beskriver kvantifiserbar kvalitet på datasett

Kvalitet kan måles på ulike nivåer

Datakvalitet kan måles på både enhetsnivå og egenskapsnivå. I eksempeldatasettet under vil hver bygning være en enhet (eksemplifisert med rød ramme), mens opplysningene som er knyttet til hver bygning er egenskaper (eksemplifisert med blå ramme for egenskapen «bruksareal»; «eier», «byggeår», «kommune» og «bruttoareal» er også egenskaper).

Der det er relevant, finnes det predefinere kvalitetsmål for både enhet- og egenskapsnivå. Et eksempel er kvalitetsmålene «antall manglende enheter» og «antall enheter med manglende verdi for en gitt egenskap» under kvalitetsdimensjonen «fullstendighet» som går på henholdsvis enhetsnivå og egenskapsnivå. I eksempeldatasettet under mangler det verdi for egenskapen «bruksareal» for Bygning nr. B00015.

Eksempeldatasett: Bygninger

Identifikator	Eier	Byggeår	Kommune	Bruksareal	Bruttoareal
B00013	Rex Eiendom	1974	Halden	510 m^2	650 m^2
B00014	Oslo kommune	2003	Oslo	2300 m^2	3000 m^2
B00015	Eiendomsutvikling AS	1995	Bergen		1200 m^2

Figur 1. Eksempeldatasett: Bygninger

Predefinerte kvalitetsmål bruker negativt ladede ord

I predefineringen av kvalitetsmål er det valgt å bruke såkalte negativt ladede ord for flere av kvalitetsmålene. Negativt ladede ord tydeliggjør feil og mangler i datakvaliteten.

Et eksempel er kvalitetsmålet «andel manglende enheter» som handler om mangel. Det vil være for eksempel 2 % *mangel* (negativt ladet) istedenfor 98 % *fullstendig* (positivt ladet) som oppgis. Det er viktig å være klar over dette, både ved angivelse av verdier til kvalitetsmål og ved visning av verdiene i et sluttbrukergrensesnitt. I et konkret sluttbrukergrensesnitt kan man godt presentere det positivt (f.eks. regne om «2 % mangel» til «98 % fullstendig» og presentere resultatet positivt).

Fritekst kan brukes som supplerende forklaring

DQV-AP-NO (Norsk applikasjonsprofil av DQV) tillater bruk av fritekst-kommentarer som supplerende forklaringer til et kvantifiserbart kvalitetsmåleresultat. For eksempel, til vårt eksempeldatasett «Bygninger», hvis resultatet på kvalitetsmålet «andel enheter med manglende verdi for en gitt egenskap» er «2 %», kan man i fritekst-kommentaren spesifisere hvilken egenskap mangelen gjelder, for eksempel: «Dette gjelder egenskap 'byggeår'».

For de aller fleste brukstilfeller antok arbeidsgruppen som utarbeidet disse definisjonene, at det burde holde med supplerende fritekst-kommentarer. For avanserte kvalitetsbeskrivelser, for

eksempel der det er behov for å avgi resultater for hver enkelt egenskap, er det i henhold til DQV mulig å oppgi slike resultater som egne datasett (dqv:QualityMeasurementDataset). For eksempel en «tabell» som sier «2 % mangel» for egenskap «byggeår», «3 % mangel» for egenskap «bruksareal» og «0 % mangel» for alle de andre egenskapene:

Egenskap	Eier	Byggeår	Kommune	Bruksareal	Bruttoareal
Andel enheter med manglende verdi for en gitt egenskap	0%	2%	0%	3%	0%

Kvalitetsmål som erfaringsmessig kan være viktige for brukerne å vite

Ikke alle de predefinerte kvalitetsmålene er relevant å måle i enhver sammenheng. Man står fritt til å velge ut de kvalitetsmålene som er aktuelle for datasettet. I noen tilfeller er det «nøyaktighet» som er viktigst for brukerne av datasettet, i andre tilfeller kan det være «konsistens», eller begge. I mange tilfeller må man også ta høyde for at det er flere ulike typer brukere. Man bør derfor velge de kvalitetsmålene som erfaringsmessig er viktige for mange brukere av det aktuelle datasettet.

Samme prinsipp gjelder for kvalitetsmålets verditype. Flere av kvalitetsmålene kan måles på inntil tre forskjellige måter: boolsk (ja/nei), heltall (antall) og prosent (andel). Dette er for å legge til rette for nivået av innsikt i kvaliteten på datasettet. For eksempel kan det være at man vet at datasettet mangler noen enheter, men ikke hvor mange eller hvor stor andel, bl.a. fordi man i utgangspunktet ikke vet hvor mange og hvilke som skal være med. I slike tilfeller benytter man seg av den boolske verditypen («ja, noen enheter mangler»). I de tilfellene der man vet hvor mange enheter som mangler, benytter man seg av heltall («fire enheter mangler») og/eller prosent («5% av enhetene mangler») hvis man også vet hvor mange enheter som skulle vært med i datasettet.

Når man beskriver et datasett som gjøres tilgjengelig for andre, beskriver man som regel kvalitet ut fra sin egen brukskontekst. Om kvaliteten som er beskrevet fra datatilbyderens ståsted er god (nok) eller ikke for brukerne av datasettet, er avhengig av brukskontekst og bruksformål. «2 % mangel» kan være bra for noen og ikke bra nok for andre.

Når man beskriver kvalitet på datasett som er løpende oppdatert (f.eks. direkte oppslag i et register som løpende ajourholdes), vil det være umulig på forhånd å vite nøyaktig hvilken kvalitet datasettet kommer til å ha. Man vil derfor som regel basere seg på erfaringene man har med datasettet, f.eks. «Statistisk sett er det 2% etterregistrering» som fritekstkommentar til «2% mangel».

Oversikt over predefinerte kvalitetsdimensjoner, kvalitetsdeldimensjoner og kvalitetsmål

Denne veilederen forklarer bruk av predefinerte kvalitetsdimensjoner, kvalitetsdeldimensjoner og kvalitetsmål. Der det er hensiktsmessig kan man referere til andre definisjoner som ligger til grunn for kvalitetsmål.

Tabellen nedenfor gir en oversikt over de kvalitetsdimensjoner, kvalitetsdeldimensjoner og kvalitetsmål som så langt er predefinerte. De er også lenket til der definisjonene finnes.

Tabell 1. Oversikt over vokabularet for kvalitetsdimensjoner, kvalitetsdeldimensjoner og kvantifiserbare kvalitetsmål.

Kvalitetsdimensjon	Kvalitetsdeldimensjon	Kvalitetsmål
fullstendighet	underdekning	manglende enheter
		antall manglende enheter
		andel manglende enheter
		antall enheter med manglende verdi for en gitt egenskap
		andel enheter med manglende verdi for en gitt egenskap
	overdekning	overflødige enheter
		antall overflødige enheter
		andel overflødige enheter
	imputering	antall enheter med imputert verdi for en gitt egenskap
		andel enheter med med imputert verdi for en gitt egenskap
aktualitet	tidsdifferanse	samlet tidsdifferanse
konsistens	konsistens innad i datasett	andel enheter med inkonsistente egenskaper
		andel enheter med inkonsistens mellom gitte egenskaper
nøyaktighet	identifikatorriktighet	antall enheter med identifikatorfeil
		andel enheter med identifikatorfeil
	klassifikasjonsriktighet	antall feilklassifiserte enheter for en gitt egenskap
		andel feilklassifiserte enheter for en gitt egenskap

Kvalitetsdimensjonen «fullstendighet»

Kvalitetsdimensjonen fullstendighet handler både om mangel på elementer i datasettet (kvalitetsdeldimensjon underdekning) og overflødige elementer i datasettet (kvalitetsdeldimensjon overdekning). Disse har kvalitetsmål på både enhets- og egenskapsnivå. Videre kan fullstendighet måles ut fra tre ulike verdityper, boolsk (ja/nei) på enhetsnivå, antall og andel på både enhets- og egenskapsnivå. Fritekst-feltet kan brukes til å opplyse om hvilken gitt egenskap som det mangler verdier for (for eksempel «bruksareal»).

Eksempel: I vårt eksempeldatasett over byggninger mangler det 2 bygninger, og dette utgjør ca. 0,02%.

Eksemplet uttrykt i RDF Turtle i henhold til DQV-AP-NO blir (eksemplet er ikke komplett slik at det kan mangle verdier for obligatoriske felter):

```
:dsBuildings a dcat:Dataset ; # et datasett
   dqv:hasQualityMeasurement # har måleresultat
      :qmMissingObjects , # manglende enheter
      :qmNumberMissingObjects , # antall manglende enheter
      :qmRateMissingObjects . # andel manglende enheter

:qmMissingObjects a dqv:QualityMeasurement ; # et måleresultat
   dqv:isMeasurementOf <https://data.norge.no/vocabulary/quality-metric#qm-
completeness-1001> ; # manglende enheter
   dqv:value "true"^^xsd:boolean ;
   rdfs:comment "Ja, noen bygninger mangler i datasettet."@nb ,
      "Yes, some buildings are missing in the dataset."@en .

:qmNumberMissingObjects a dqv:QualityMeasurement ; # et måleresultat
   dqv:isMeasurementOf <https://data.norge.no/vocabulary/quality-metric#qm-
completeness-1002> ; # antall manglende enheter
   dqv:value "2"^^xsd:nonNegativeInteger ;
   rdfs:comment "To bygninger mangler i datasettet."@nb ,
      "Two buildings are missing in the dataset."@en .

:qmRateMissingObjects a dqv:QualityMeasurement ; # et måleresultat
   dqv:isMeasurementOf <https://data.norge.no/vocabulary/quality-metric#qm-
completeness-1003> ; # andel manglende enheter
   dqv:value "0.02"^^xsd:double ;
   rdfs:comment "0,02% av bygninger mangler i datasettet."@nb ,
               "0.02% of buildings are missing in the dataset."@en .
```

Fullstendighet handler også om imputering. Imputering er å fylle inn verdi for en gitt egenskap der verdien mangler eller er ubrukbar. Dette gjøres for å håndtere manglende verdier for egenskaper (tomme celler) i et datasett der disse manglende verdiene skaper problemer for, blant annet, analysen av dataene. Imputerte verdier som kvalitetsmål gir datatilbyderen mulighet til å informere brukerne av datasettet at det er egenskaper i datasettet som ikke er hentet fra virkeligheten.

Eksempel: Egenskapen «byggeår» for fire av de eldre bygningene i eksempeldatasettet vårt har fått impurterte verdier.

Eksemplet uttrykt i RDF Turtle i henhold til DQV-AP-NO blir (eksemplet er ikke komplett slik at det kan mangle verdier for obligatoriske felter):

```
:dsBuildings a dcat:Dataset ; # et datasett
   dqv:hasQualityMeasurement # har måleresultat
      :qmNumberImputedValues . # antall impurterte verdier

:qmNumberImputedValues a dqv:QualityMeasurement ; # et måleresultat
   dqv:isMeasurementOf <https://data.norge.no/vocabulary/quality-metric#qm-
completeness-3001> ; # andel enheter med imputert verdi for en gitt egenskap
   dqv:value "4"^^xsd:nonNegativeInteger ;
   rdfs:comment "Fire bygninger har fått imputert verdi for egenskapen 'byggeår'."@nb
,
      "Four buildings have got imputed value for the property 'year of
construction'."@en .
```

Kvalitetsdimensjonen «aktualitet»

Det er predefinert ett kvalitetsmål i kvalitetsdimensjonen aktualitet – samlet tidsdifferanse.

Eksempel: For bygninger i eksempeldatasettet vårt tar det statistisk sett ca. 24 dager fra en bygning kontraktsmessig skifter eier til eierskiftet blir meldt inn. Medregnet intern saksbehandlingstid setter datatilbydren «30 dager» som «samlet tidsdifferanse».

Eksemplet uttrykt i RDF Turtle i henhold til DQV-AP-NO blir (eksemplet er ikke komplett slik at det kan mangle verdier for obligatoriske felter):

```
:dsBuildings a dcat:Dataset ; # et datasett
   dqv:hasQualityMeasurement # har måleresultat
      :qmOverallDelay . # samle tidsdifferanse

:qmOverallDelay a dqv:QualityMeasurement ; # et måleresultat
   dqv:isMeasurementOf <https://data.norge.no/vocabulary/quality-metric#qm-
currentness-1001> ; # samlet tidsdifferanse
   dqv:value "P30D"^^xsd:duration ;
   rdfs:comment "Det tar i gjennomsnitt 24 dager fra en bygning står ferdig eller er
revet til den er innlemmet i eller tatt ut fra datasettet. Medregnet intern
saksbehandlingstid blir den samlede tidsdifferansen 30 dager."@nb ,
      "On average there will be 24 days from a building is completed or demolished,
to it is included in or excluded from the dataset. With internal processing time
included, the overall time difference is 30 days."@en .
```

Kvalitetsdimensjonen «konsistens»

Kvalitetsdimensjonen konsistens gjelder konsistens innad i ett og samme datasett, og ikke konsistens mellom datasett. Om datasettet er i samsvar med gitte standarder og krav er ikke definert på nytt som et eget kvalitetsmål ettersom dette er dekket av DCAT-AP-NO (Datasett: i samsvar med). Et eksempel på slik innbyrdes inkonsistens er når bruksareal er større enn bruttoareal for en bygning.

Kvalitetsdimensjonen konsistens kan i mange tilfeller lett forveksles med kvalitetsdimensjonen nøyaktighet. Det som bl.a. skiller nøyaktighet og konsistens er at når det gjelder konsistens *vet man ut fra vurdering av flere egenskaper at det er feil, men ikke hvilken eller hvilke egenskaper som er feil* i datasettet. I eksemplet over er det ikke mulig å avgjøre om det er bruksareal eller bruttoareal (eller begge) som er feil. Når det gjelder nøyaktighet, *vet man hvilken egenskap som er feil* (for eksempel feil identifikator).

Det første kvalitetsmålet i konsistens (andel enheter med inkonsistente egenskaper) måles på enhetsnivå. Her måles andel enheter som har en form for inkonsistens knyttet til seg. Det andre kvalitetsmålet (andel enheter med inkonsistens mellom gitte egenskaper) går mer i dybden og brukes der man har innsikt i hva inkonsistensen gjelder på egenskapsnivå. Kvalitetsmålene oppgis i prosentandel; fritekstfeltet kan brukes til å forklare for hvilke egenskaper inkonsistensen gjelder.

Eksempel: For vårt eksempeldatasett er det ca. 0,03% av bygningene som har inkonsistens mellom egenskapene "bruksareal" og "bruttoareal".

Eksemplet uttrykt i RDF Turtle i henhold til DQV-AP-NO blir (eksemplet er ikke komplett slik at det kan mangle verdier for obligatoriske felter):

```
:dsBuildings a dcat:Dataset ; # et datasett
   dqv:hasQualityMeasurement # har måleresultat
      :qmRateInconsistencyGivenProperties . # andel enheter med inkonsistens mellom
gitte egenskaper

:qmRateInconsistencyGivenProperties a dqv:QualityMeasurement ; # et måleresultat
   dqv:isMeasurementOf <https://data.norge.no/vocabulary/quality-metric#qm-
consistency-1002> ; # andel enheter med inkonsistens mellom gitte egenskaper
   dqv:value "0.03"^^xsd:double ;
   rdfs:comment "0,03% av bygningene i datasettet står oppført med 'bruksareal' som
er høyere enn 'bruttoareal'."@nb ,
      "0.03% of the buildings in the dataset have 'usable area' larger than 'gross
area'."@en .
```

Kvalitetsdimensjonen «nøyaktighet»

I kvalitetsdimensjonen nøyaktighet måles i hvilken grad dataene korrekt representerer virkeligheten.

Nøyaktighet av en dataverdi er ofte avhengig av type data, og kvalitetsmål for nøyaktighet blir fort svært fag- og sektorspesifikke. De mest generelle nøyaktighetsmålene er derfor plukket ut i denne

sammenheng: identifikatorriktighet som går på identifikasjonsnøkler, og klassifikasjonsriktighet som går på bruk av klassifikasjoner og kodeverk.

Eksempel: For vårt eksempeldatasett er det ca. 0,01% av bygningene som har fått feil identifikator.

Eksemplet uttrykt i RDF Turtle i henhold til DQV-AP-NO blir (eksemplet er ikke komplett slik at det kan mangle verdier for obligatoriske felter):

```
:dsBuildings a dcat:Dataset ; # et datasett
   dqv:hasQualityMeasurement # har måleresultat
      :qmRateIncorrectIdentifier . # andel enheter med identifikatorfeil

:qmRateIncorrectIdentifier a dqv:QualityMeasurement ; # et måleresultat
   dqv:isMeasurementOf <https://data.norge.no/vocabulary/quality-metric#qm-accuracy-
1002> ; # andel enheter med inkonsistens mellom gitte egenskaper
   dqv:value "0.01"^^xsd:double ;
   rdfs:comment "0,01% av bygningene i datasettet har fått feil identifikator."@nb ,
      "0.01% of the buildings in the dataset have wrong identifiers."@en .
```

Navnerom som er brukt i veilederen

Tabell 2. Oversikt over navnerom som er brukt i denne veilederen.

Prefiks	Navnerom	Forklaring/navn
dcat	http://www.w3.org/ns/dcat#	Data Catalog Vocabulary
dqv	http://www.w3.org/ns/dqv#	Data Quality Vocabulary
rdfs	http://www.w3.org/2000/01/rdf-schema#	RDF Schema 1.1
xsd	http://www.w3.org/2001/XMLSchema#	XML Schema Part 2: Datatypes Second Edition

Eksempel på prefiksene ovenfor uttrykt i RDF Turtle:

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dqv: <http://www.w3.org/ns/dqv#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
```