Høringsuttalelse til Nasjonal spesifikasjon for metadata om helsedata

Vi takker for muligheten til å komme med høringsuttalelse til oppdatert nasjonal spesifikasjon for metadata om helsedata.
Bidragsytere i vårt arbeid har vært bredt sammensatt og inkluderer statistikere, analytikere, forskere, helsetjenesteutviklere, virksomhetsarkitekter og deltakere med spisskompetanse innen kvalitetsregistre og EPJ. Arbeidsgruppen har hatt representanter fra Helse Bergen, Helse Stavanger, Helse Vest IKT og Helse Vest RHF.
Denne tilbakemeldingen er en samlet tilbakemelding fra Helse Vest.

Generell lesbarhet og innspill

Dokumentet i sin helhet er godt strukturert og gjennomarbeidet. Det har et godt formål som vi stiller oss bak. Et slikt felles dokument kan bidra til en god og helhetlig forståelse på tvers av registre og systemer og derfor ser vi på dette en fornuftig retning fremover.

Det er imidlertid noen betraktninger vi ønsker å gi tilbakemelding på, som etter vår mening kan bidra til at retningslinjen blir mer forståelig og at selve arbeidet med implementeringen av retningslinjen kan forenkles.

Hovedinntrykket vi sitter igjen med etter vår arbeid er at denne spesifikasjonen i veldig stor grad er rettet inn mot å tilfredsstille behov knyttet til Helseanalyseplattformen (HAP) og forsøk på å tilpasse seg mulighetsområdet knyttet til systemet Healthterm. Vi ser ingenting i veien for at mottak av data for HAP må harmoniseres slik at forskere lett kan søke opp variabler for å finne aktuelle datasett, men vi stiller oss undrende til at en retningslinje innsnevres for å tilfredsstille krav og føringer som i dette tilfellet skyldes systemet Healthterm. Det grenser opp til det å sette likhetstegn mellom Felles språk og SNOMED CT. På samme måte som at det er flere språkvarianter enn det som kan uttrykkes gjennom SNOMED CT, er det også mange helsedata som ikke passer inn i formatet som Healthterm (og HAP) krever. Vi mener at en spesifikasjon bør ha et langsiktig perspektiv og være uavhengig av temporære systemvalg. Da ville det være bedre å ha en egen kravspesifikasjon som henviste til hvordan man for øyeblikket stiller krav til import av metadata til Healthterm.

En annen refleksjon vi gjør oss er at dokumentet i sin helhet i enda tydeligere grad kunne kommunisert innretningen av retningslinjen – hva er det egentlige omfanget og hva faller utenom som andre (veiledere og enheter) har ansvaret for og som eventuelt sektoren i større grad må ta ansvar for.

Begreper

Dokumentet har et teknisk språk og det er også redegjort for hvilke målgrupper dokumentet retter seg mot. Men det kan likevel stilles spørsmål om det kunne og burde nådd ut til enda flere med et klarere språk uten at det gikk utover innholdets presisjon. Begrepsdefinisjoner er en teknikk for å gjøre innholdet mer forståelig, og det er bra at det vises en tabell over de viktigste begrepene tidlig i spesifikasjonen. Det er imidlertid viktig å presisere at Digitaliseringsdirektoratet har utarbeidet en

veileder for begrepsdefinisjoner med den hensikt å gi begrepene mer presist og entydig innhold, fordi termer i seg selv lett kan brukes om hverandre og forstås på ulik måte. Når begrepene i dette dokumentet blir tydeliggjort i form av forklaring fremfor definisjon, forstår vi det som uttrykk for forfatternes subjektive oppfatninger av begrepet fremfor en kollektiv forståelse av begrepene. Dette inntrykket forsterkes av at det ikke finnes noen kildehenvisninger til forklaringene. Dette er uheldig og vi mener at det burde vært gjort et større arbeid knyttet til harmonisering av begreper i forkant, slik at etablerte begrepsdefinisjoner ble benyttet, og at etablerte begreper og nye begreper var gjort tilgjengelig i begrepskatalogen på data.norge.no.

Det blir noe forvirrende med begrepsforklaringer knyttet til informasjonsmodellen i kapittel 2 og forklaringene i begrepslisten. Det er flere årsaker til dette. De fleste begrepene i kapittel 2 er også å finne i begrepslisten, men ikke alle. Siden det angis at en «komplett oversikt over alle fagbegrep er gjengitt i begrepslisten» er vår forventing at alle begrepene fra kapittel 2 burde være gjengitt. Det er ikke tilfelle. Om det er en forglemmelse eller om det er begrep i informasjonsmodellen som ikke kan regnes å være fagbegrep er uklart, men siden det er få begreper samlet sett, burde det ikke være problematisk å ha en komplett begrepsliste. Det er en smakssak hvor begrepslister bør plasseres i slike dokumenter, men flere av oss har uttrykt ønske om at alle begrepene var gjennomgått tidligere i dokumentet, for å forbedre lesbarheten og gjøre det enklere med navigering. Av de begrepene som finnes begge steder, er det enkelte med noe ulik forklaringstekst. Det virker unødvendig. Det er også begreper i dokumentet som med fordel kunne vært redegjort for; datakatalog, datasett, dcat, skos, Nasjonal variabelkatalog, objektstruktur. I tillegg kan det være problematisk at det gjennomgående brukes ordet *Concept* i spesifikasjonen, da man kan være av den oppfatning av at det symboliserer skos:concept. Det ser imidlertid ut til at *Concept* brukt i spesifikasjonen er løsningsspesifikk, og da bør denne distinksjonen kommenteres eksplisitt i forklaringen/ definisjonen. Spesielt viktig er det når spesifikasjonen forsøker å tilnærme seg internasjonale standarder som dcat og skos.

Det er nevnt i flere sammenhenger at vi mangler en omforent begrepsmodell innenfor helsesektoren. Vi er mange som ser behovet for det, samtidig er vi klar over at det vil innebære en stor innsats for mange aktører over lang tid. Vi tror imidlertid at dette arbeidet kan være behovsdrevet, og at en begrepsmodell gradvis kan utvikles. Vi ville sett det som en stor fordel om denne spesifikasjonen kunne være startpunktet for en slik begrepsmodell. Det innebærer blant annet å gjøre et valg om man skal ta utgangspunkt i en eksisterende internasjonal standard (f.eks. Contsys) eller om det skal utvikles noe helt nytt. En felles begrepsmodell vil kunne bidra til at miljøer som i dag ikke er en del av spesifikasjonens omfang, kan utvikle sine områder i tråd med et slikt felles overbygg.

Eksempler

Et annet grep for å gjøre innholdet lettere tilgjengelig er å bruke eksempler. Dette er gjort flere steder i dokumentet, men det kunne med fordel vært brukt enda flere steder (se kommentarer lenger ned i dokumentet)

Omfang

Med tittelen «Nasjonal spesifikasjon for metadata om helsedata» vil man umiddelbart tenke at dette dreier seg om standardisering av metadata for hele sektoren. Fokuset er imidlertid på eksport av data fra helseregistre (sekundærkilder) og import av disse dataene til hhv. HAP og Healthterm (tertiærkilder). Verdikjeden for helsedata fra den oppstår, blir registrert i en primærkilde, og rapportert til en sekundærkilde, er lite behandlet. Til tross for at dette er redegjort for i retningslinjen har det likevel skapt misforståelser. En figur kunne nok ha bidratt til å tydeliggjøre omfanget og avgrensninger.
I dokumentet benevnes omfanget som «sekundære formål som forskning, kvalitetsforbedring og annen analyse». Etter vår oppfatning dreier dette seg om sekundære formål via HAP og/eller helseregistre, men ikke fra primærkilder, som også kan være gjenstand for mange andre sekundære

formål som forsking, kvalitetsforbedring og annen analyse lokalt, regionalt, nasjonalt og internasjonalt. Med primærkilder tenker vi alle journalsystemer der helsedata først oppstår/registreres.
Det bør også tydeliggjøres i hvilken grad denne retningslinjen påvirker andre deler av sektoren og leverandører som har sitt fokus på primærkilder, og overganger mellom primærkilder og sekundærkilder. Kanskje vil det kunne være forskjeller i grad av påvirkning på sektoren i dag og noen år frem i tid, og dette tror vi mange kunne hatt nytte av å få beskrevet tydeligere.
Vi er av den oppfatning at registrene må harmoniseres seg imellom på det semantiske nivået, da det er lite hensiktsmessig både for registrering i primærkildene og videre bruk i tertiærkilder dersom sekundærkildene definerer det samme datapunktet på mange ulike måter. For å kunne automatisere rapportering til helseregistre, må det også gjøres en form for harmonisering mellom primærkilder og sekundærkilder. Derfor er det problematisk å begrense omfanget til å gjelde overganger mellom sekundær- og tertiærkilder, uten å beskrive hvordan man ser for seg at denne retningslinjen kan tenkes å utvides i omfang i tiden som kommer. Automatisk høsting av data fra primærkilde (journal) og videre til sekundær og tertiærkilder vil best kunne løses ved at input til primærkilde (journal) struktureres. Det synes altså å være et behov for å se hele verdikjeden fra en datafødsel i journalen til nyttig bruk i registersammenheng under ett. Et alternativ til å ta for seg hele verdikjeden for helsedata i dette dokumentet kunne være å gi det et navn som beskriver begrensningene, for eksempel «Kravspesifikasjon for metadata til Helseanalyseplattformen (HAP) i innledende fase 2021-2022».

Internasjonale standarder

Det refereres til internasjonale standarder som HL7 FHIR, SNOMED CT, Contsys osv. vedrørende koordinerings-, harmoniserings- og berikingsformål uten at man kommer inn på hvordan dette er tenkt gjort eller hvilke muligheter det kan gi. Det hadde vært nyttig å vite mer om disse tankene og også hvordan det i større grad kan henge sammen med bruk av openEHR/ arketyper i primærkilder.

Detaljert tilbakemelding til kapittel 2 og 4

Logisk informasjonsmodell for metadata om helsedata

Den logiske informasjonsmodellen er omfattende og setter store krav til de som skal levere data på dette formatet. En bekymring er rigiditeten til importformatet og om det kan resultere i mange feilmeldinger ved import/eksport. I verste fall kan det innebære at en del data ikke blir lest inn og gjort tilgjengelig. Ekstra viktig blir dette om retningslinjen på sikt vil gjelde et større omfang, der man kan tenke seg andre behov og sekundærbruk av data er tilstede. I slike tilfeller kan det tenkes at man kun ønsker å harmonisere deler av spesifikasjonen med enkelte variabler, mens for øvrige deler har man en mer åpen og fleksibel tilnærming. Dette kan være viktig for blant annet maskinlæring der det er viktig å samle inn større datamengder, men der kravet til utfylling av alle variabelfelt ikke er like stort.

Generelt

- ER-diagrammet er godt dersom du har en teknisk forståelse, men ikke like nyttig for en som skal evaluere hvordan de skal sende inn data eller hvordan det skal representere ens entiteter, bilder, medikasjon osv.
- Modellen kan være nyttig om datakilden er helseregistre, men ikke for det som kommer inn til helseregistrene (primær kilde).
- Vurder bruk av flere eksempler for hvordan dette kan representeres (ref. informasjonsmodellen på side 11).
- Hvorfor er Variabelgruppe avgrenset til 2 nivå?

Informasjonsmodellen

- En Variabel kan ifølge figuren maksimalt bare være tilknyttet ett Instrument. Det finnes flere variabler (eks. kliniske målinger) som inngår i flere instrument (eks. skåringsverktøy). Dette passer da ikke inn i denne modellen. Kan modellen vurderes utvidet slik at en variabel kan være tilknyttet flere instrument?

Spørsmål angående variabler og versjoner

- I praksis kan en variabel endre seg over tid, ved for eksempel å få flere (eller færre) svaralternativ eller å gå fra å være ikke-obligatorisk til obligatorisk (eller omvendt). Er det slik at vi da må lage flere «kunstige» variabler i innsendingen, med unike variabelnavn?
- Dersom eksempelvis variabelen «KJONN» (kjønn) går fra ikke-obligatorisk i 2020 til obligatorisk i 2021, må vi da lage to variabler, eks. «KJONN_PRE2020» og «KJONN_POST2020», med ulike verdier for «Obligatorisk»? Og når det så går fra to svaralternativ (mann og kvinne) til tre (mann, kvinne og annet), må det lages enda en variant («KJONN_POST2020_TRESVARALTERNATIV»)? Vår erfaring er det er forekommer slike tilfeller i kvalitetsregistrene, der for eksempel en variabel er merket «obligatorisk», men som faktisk ikke er det (dvs. det mangler data for enkelte pasienter) da den på et tidligere stadium var ikke-obligatorisk.
- Er det et begrep for time-variabel? Kan du ha en variabel som endres underveis – f.eks. kjønn?
- Hva med kode for «missing data»?

Innrapportering av metadata til Nasjonal variabelkatalog

4.2.1 Filtyper for importfiler

- Det står at alle importfiler skal være i .xls- eller .xlsx? .xls er et veldig gammelt, ustandardisert format, som få bruker. Kan det avgrenses til XLSX?
- Det bør spesifiseres for hvert felt / hver Excel-kolonne hvilken variabeltype som skal benyttes. Skal for eksempel alle kolonner være tekstkolonner? At kolonner som DataFra og DataTil skal skrives som YYYYMMDD i steden for som Excel-datoer, tyder på det. Men hva med kolonner som MIN, MAX og AVG? Eller KodeverkLokalID (denne er ofte heltall, men for noen variabler i noen registre er det tekst).
- Vi har dårlige erfaringer med (ugyldige/rare) Excel-filer generert av ulike verktøy, og slike filer er vanskelige å feilsøke i. Vi mener det bør være CSV for alle filer.
- Det står at «Mapping-fil skal være i CSV-format (*.csv) med komma (‘,’) som separasjonstegn, ikke semikolon (‘;’).». Dette er veldig upresist. CSV kan jo (dessverre) være så mangt. Hvordan skal for eksempel "-tegn eller linjeskift inni tekststrenger skrives/skapes? Det bør være en referanse til en standard også her, for eksempel RFC 4180. Det bør også stå eksplisitt hvilken tegnkode CSV-filene skal ha (UTF-8?).

4.2.2 Navnestandard for importfiler

- Det bør stå noe om hva som er gyldige tegn som (ikke) kan brukes i de ulike komponentene av filnavnene. Det virker for eksempel rimelig at en ikke kan bruke understreking (da blir filnavnene tvetydige).
- I eksempelet på filnavn på side 17 er det et mellomrom som trolig ikke skal være der (før _2018): DAR_1a-Kildemetadata1 _2018_v128_20190130

4.2.3 Format for importfiler

- I første linje står det en feilmelding («Error! Reference source not found»).
- Det er ikke helt enkelt å forstå Tabell 2, side 19. Her kan det være nyttig med et ekstra eksempel.

4.2.4 Formatering av tekst

- Det står at enkelte Properties kan formatere med Markdown-syntaks. Dette bør være «skal». Innholdet i en tekst vil tolkes forskjellig avhengig om det regnes som ren tekst eller Markdown-tekst. Dersom en for eksempel har tekst som nevner formelen «konsentrasjon*vekt*tid», blir det tolket som en referanse til tre variabler som multipliseres dersom det er ren tekst.
- Felttyper/kolonnetyper bør eksplisitt spesifiseres. Det er ikke åpenbart hvilket felt en referer til med teksten «Properties med beskrivende tekst». Kan for eksempel ha en ekstra kolonne (i alle tabeller i dokumentet) som sier om hvilket felt skal skrives som ren tekst, Markdown-tekst, heltall, desimaltall e.l.
- Oversikt over Markdown-syntaks er inkonsistent og forvirrende. Av og til viser kolonne tre til eksempel på inndata (**Bold tekst**) og av og til på utdata (<h3>). (Dessuten er vel #, ## og ### heller noe som bør være <h1>, <h2> og <h3> enn <h3>, <h4> og <h5>.)
- Markdown-koden for uordnet liste er tankestrek (i enkle hermetegn) i kolonne 2, men bindestreker i kolonne 3.
- Hva skjer med ugyldig markup (eks. som inneholder «[»)? Blir det det validert?

4.2.5: Importfil 1a-Kildemetadata (Tabell 3)

- Det må finnes en spesifikasjon av navnekonvensjon for datakilder (det ser ikke ut til at vi kan ha understrek eller punktum).
- Generelt for Code i de ulike tabellene: det bør være mer presis spesifisering av hva som er gyldige navn. Det står nå «Store bokstaver, uten mellomrom og æ, ø, å.» Er greske bokstaver (μ og π) tillatt?! Er mellomrom tillatt? Det ser også ut til at punktum vil føre til tvetydigheter. Forslag: Bare de store bokstavene A–Z, sifrene 0–9 og understrek (_) er tillatt, og første tegn kan ikke være et siffer. Sistnevnte unntak fordi variabel-/feltnavn som begynner med siffer er ulovlige i noen databaser og statistikkprogram, og det fører generelt til problemer.
- Hva som er gyldige navn kan godt nevnes et sted i dokumentet (ikke i hver tabell), og så kan man ha med referanse til hovedregelen der det er aktuelt (dvs. i hver tabell).
- PreferredTerm vs. KortNavn: Førstnevnte skal være et kort navn, og sistnevnte en (obligatorisk!) forkortelse av navnet? Kan vises med et eksempel?
- Beskrivelse: Ha heller med et eksempel enn å skrive at en kan finne et eksempel et eller annet sted på helsedata.no.
- Tilslutningsgrad, Dekningsgrad: Hva betyr «Oppgis som desimaltall.»? Skal for eksempel 85,7 % skrives som «85,7», «0,857», «0.857», «.857» eller på en annen måte?
- Lovverk, Forskrift, HjemmelTilgjengeliggjøring: Kan man bare ha en URL, eller kan vi ha flere? Det er behov for flere. I så fall, skiller vi URL'ene med mellomrom?
- Hjemmeside: Er «Lenke» noe annet enn URL? (En kan ha ekte, klikkbare «lenker» i Excel-filer.)
- Kontaktinformasjon: Er det noe spesielt format de ulike opplysningene her skal ha, eller er det fritt fram (fritekst)?

4.2.6: Importfil 1b-Variabelmetadata

- Nr 6. PresentationOrder (her og andre plasser): det er ikke definert hvilke verdier disse kan ha (tallene 1, 2, 3 osv., eller fritekst), og hva verdiene betyr.
- Nr. 23 DataType: 7: Datetime: Det bør stå noe om hvilken tidssone datoene skal være i? Norsk (sommartid/vintertid?) eller UTC? Dersom Excel ikke lagrer noe informasjon om tidssoner, kan det bli utelatt fra importfilene.

- Nr. 24: Lengde: det er uklart hva dette betyr og hvordan det skal skrives. Gjelder det bare for String-variabler eller også for eksempelvis Integer og Date-variabler?
- Nr. 25: Presisjon: Maks tall på desimaler, minimum eller nøyaktig hvor mange desimaler alle tallene skal ha?
- Nr. 26 og 27: GrenseLav og Grensehoy: det bør presiseres hva dette er (minimums og maksimumsverdier som verdiene i datasettet kan ha). Gjelder dette bare tall eller også datoer og tidspunkt? Dersom ja, hvordan skal de skrives?)
- Vi har også erfart at «manglende verdier» for noen variabler kan registreres på ulikt vis, for eksempel at verdien «-1» (eller «NA» eller «NULL» eller «""») betyr «ikke utfylt». Er det noen måte å støtte dette på, enten her eller i 4.2.11? Eller skal datafilene vi sender fra oss standardiseres slik at manglende / ikke utfylt verdi skal normaliseres til «tom verdi». Da bør dette stå eksplisitt.

4.2.8 Importfil 1d-Statistikk

- Nr. 21: STD: Det finnes to vanlige definisjoner av standardavvik (i Excel kalla STDAV.S og STDAV.P). Det bør defineres hvilken av de to som skal benyttes. Den vanligste i bruk i statistikkprogram er STDAV.S og vi anbefaler derfor bruk av denne.
- Nr. 23: KodeverkLokalID: Hvilket tegn kan denne inneholde?

Begrepsliste (tabell 12)

- I skildringen av «Kode» er «Kodebeskrivelse» nevnt som et begrep. Men det er ikke brukt noe annet sted i dokumentet. Se for øvrig kommentarer om begreper i den generelle tilbakemeldingen.