URI-er for begreper og data i norsk offentlig sektor

Semicolon rapport

Status: under arbeid.

Versjon 1.0

November 2011

Endringskontroll

Versjon	Dato	Forfatter	Kommentar
0.1	31.09.2011	Jens K. Mjelva (ed.)	Initiell versjon (wiki)
0.2	23.10.2011	David Norheim	Overført til dokument
0.3	28.10.2011	David Norheim (ed.)	Konsolidert versjon etter kommentarer fra semicolon partnere (Audun Stolpe, UiO)
0.4	03.11.2011	David Norheim (ed.)	Konsolidert versjon etter kommentarer fra semicolon partnere (Per Myrseth, DNV)
0.5	13.11.2011	David Norheim (ed.)	Konsolidert versjon etter kommentarer fra semicolon partnere (Dumitru Roman, Sintef)
0.6	23.11.2011	David Norheim (ed.)	Innhentet kommentarer fra Steinar Skagemo, Difi og Mediarena
1.0	14.12.2011	David Norheim (ed.)	Endelig leveranse

Innholdsfortegnelse

Innledning

Dette dokumentet er en del av et sett av anbefalinger relatert til bruk og publisering av gjenbrukbare begreper og data for norsk offentlig sektor fra Semicolon.

Dokumenter vil omfatter emner som:

- Oppbygging av URI-er for begreper og data i norsk offentlig sektor (dette dokumentet)
- Publisering av Linked Open Data for norsk offentlig sektor
- Bruk av Linked Open Data
- Kvalitetskarakteristikker for data identifisert ved et URI-sett

I tillegg bør dokumentet sees i sammenheng med:

- Livssyklusprosess for å utvikle og kvalitetssikrede begrepsystemer [1] (publisert ifbm. Semicolon I)
- Åpne og lenkede data. En informasjonsinfrastruktur for elektronisk samhandling. [2] (publisert i fbm. Semicolon I)

Dette dokumentet definerer designvurderinger og anbefalinger på hvordan Universal Uniform Resource Identifiers (URI-er) for begrepsapparater og data bør bygges opp og utvikles i norsk offentlig sektor. Dokumentet er ment for:

- Etater som selv publiserer begrepsapparater
- Etater som selv publiserer data (dataeiere)
- Etater og andre som refererer (lenker) til andres begreper eller data
- Leverandører som ønsker å bygge løsninger for eller på data fra norsk offentlig sektor.

Dokumentet søker å etablere et sett av prinsipper detaljert i designvalg, konsekvenser av designvalg og beste praksis for oppbygging av Linked Data URI-er. Det er viktig å merke seg at anbefalingene gjelder uavhengig av om data kun benyttes internt i en etat eller om de viderebrukes utenfor etaten i åpne eller lukkede nett.

Aktuelle prinsipper er

- Domene og opphav for URI-er
- Valg av sti-struktur for URI-er
- Livstidsutfordringer med tanke på endringer
- Oppslag på URI-er
- Kvalitetskarakteristikker
- Maskinlesbarhet og menneskelesbarhet
- Eierstyring (Governance)

Dokumentet har ikke til hensikt å dekke innholdet, strukturer og vokabularer på hva som returneres av et oppslag på en URI, f.eks. hvis URI-en limes inn i nettleserens adressefelt. Dette vil dekkes av fremtidige rapporter.

Om bakgrunn for dokumentet

Dokumentet er et resultat av arbeid i samhandlingsprosjektet Semicolon II. Det er blitt til gjennom å studere praksis nasjonalt og internasjonalt, og vil ved godkjenning ha vært tema for en rekke workshop-er i regi av Semicolon II.

Prosessen med godkjenning, bestemt i styringsgruppemøtet, er:

a. Intern gjennomgang blant Semicolon II LOD deltakere
b. først til to eksterne DIFI og Mediaarena for teknisk gjennomgang,
c. retting av kommentarer.
c. gjennomgang/godkjent leveranse av styringsgruppen (denne leveransen).
d. sendes videre til DIFI for innspill til standard for forvaltningen

Eierskap til prosessen og dette dokumentet

Versjon 1.0 av dokumentet har fulgt Semicolon IIs prosess for godkjenning. I videre arbeid med dokumentet bør prinsippene nevnt i dette dokumentet gjennomgås med en målsetning å fremmes som en forvaltningsstandard av Difi (Standardiseringsrådet) som en eventuell anbefalt eller obligatorisk standard hjemlet i referansekatalogen/IT-standardforskriften. Den videre prosessen vil da følge Standardiseringsrådets prosess.

Aktiviteter i andre miljøer

Arbeid rundt design av URI-er til dette formålet er tidligere gjort i andre (herunder UK government og i norske samhandlingsprosjekter), og de siste årene har gitt nyttig erfaring rundt disse prinsippene som vi også tar med her.

Inspirasjonskilder inkluderer

- UK Governments arbeid rundt generelle offentlig sektor URI-er og spesifikke regimer for lokasjoner. Se spesielt [3][4].
- Brønnøysundregisterenes arbeid rundt Semantikkregisteret for offentlig sektor, SERES [15].

I referansekapittelet er det angitt andre konkrete dokumenter som tar opp deler av problemstillingene som også er nevnt i dette dokumentet.

Motivasjon

Hva betyr det å etablere et system for URI-er for begreper og data? Kort fortalt betyr det at man skal ha mulighet til å entydig peke på en definisjon, uavhening av hvilken term som er brukt f.eks.

> «Bruttoinntekt – slik Skatteetaten definerer det»
> «Bruttoinntekt – slik Nav definerer det»

ettersom disse er forskjellige selv om man i dagligtale vil kalle begge «bruttoinntekt».

Tilsvarende for en konkret «ting»

> «Bilen, som Vegdirektoratet har gitt kjennetegn DN 19690» og
> «Enheten, som Brønnøysundregisteret har gitt organisasjonsnummer 986352325»

I en global skala er det URI-er en standard for denne type referanser.

Hvorfor trenger vi globale referanser?

I dag benyttes normalt lokale primærnøkler til dette formålet i databaser. Nasjonalt benyttes nasjonale standardiserte identifikatorer for å referere mellom organisasjoner (fødselsnummer, organisasjonsnummer).

En av de viktigste årsakene til å benytte URI-er er muligheten for kunne referere og gjøre oppslag på disse nøklene. Dette er best forklart gjennom en sammenlikning med en HTML tabell med og uten hyperlinker.

Hyperlinker lar deg referere ned i detaljene dersom du trenger å referere til dataene utenfra. Ta for eksempel en side http://eksempel.no/index.html som inneholder en tabell som følger

Person	Organisasjon
Espen Askeladd	Innovasjon AS
Trollmor	Trubadurene DA
Huldra	Trubadurene DA

I denne tabellen kan man ikke referere direkte personene, ei heller til organisasjonene. Man kan heller ikke referere til hva som ligger i definisjonen på en "Person" eller "Organisasjon".

Dersom vi legger på hyperlinker kan vi både refereres til og vi kan referere direkte til mer detaljer om organisasjonene. Merk at dette er hyperlinkene bak teksten - teksten som brukeren ser er den samme som over.

http://vokabular.no#Person	http://vokabular.no#Organisasjon
http://eksempel.no#EspenAskeladd	http://eksempel.no#InnovasjonDA
http://eksempel.no#Trollmor	http://eksempel.no#TrubadureneDA
http://eksempel.no#Huldra	http://eksempel.no#TrubadureneDA

Man kan da ha en tilsvarende tabell som har detaljer om organisasjonene.

http://vokabular.no#Organisasjon
http://eksempel.no#InnovasjonDA
http://eksempel.no#TrubadureneDA

Og definisjonene kan finnes i en annen refererbar tabell

Begrep	Definisjon
http://vokabular.no#Person	En person er…
http://vokabular.no#Organisasjon	En organisasjon er…

Vi ønsker med andre ord å etablere URI-er som angir unikt (med forenklet prefix-syntaks):

skd-begrep:inntekt
nav-begrep:inntekt
vegdirektoratet-bil:DN19690

br-enhet:986352325

Referansene blir stabile, og entydige. De kan refereres til av andre, og hvor man ved opplag kan få informasjon om hva «tingen» eller definisjonen er. Altså en mekanisme som lar oss få mer informasjon om det vi har lenket til.

Det er behov for felles designregler for bruk av URI-er slik at vi får noen trafikkregler i offentlig sektors informasjonsinfrastruktur. Dette dokumentet er en guide som ivaretar viktige hensyn ved etablering av URI-er for dette formålet.

URI-typer

Ovenfor har vi skilt mellom to typer referanser, referanser til konkrete ting og til begreper. Vi kaller gjerne disse ID-URI-er og Begrep-URI-er. Tabellen under beskriver dette mer formelt.

Type ressurs	Type URI som navngir ressursen	Kommentar
«Ting» i den virkelige verden	ID-URI	Dette er både fysiske og abstrakte «ting». Synonym: Instans Eksempel: en konkret bil, en bestemt person eller en spesifik sektor, en hendelse
Begrep	Begrep-URI	Dette er en identifikator for et begrep. F.eks. «bil» eller «person». Synonym: Konsept Merk at det er definisjonen og ikke ordet (termen) som avgjør om noe er et begrep.

En ID-URI er en identifikator for en konkret fysisk objekt (f.eks. bilen, båten, personen eller organisasjonen), men objektet selv kan ikke lastes ned over nettet. Du kan imidlertid få informasjon om objektet, i form av en html side, xml-fil e.l. Adressen til denne informasjonen kalles en informasjonsressurs. En informasjonsressurs er enhver ting som man i prinsippet kan laste ned over nettet, det mest kjente eksemplet er en URL for et HTML-dokuement.

Vi skiller derfor mellom ID-URI og Dokument-URI, der de siste representerer et dokument. En Begrep-URI vil også ha et Dokument-URI som representerer innslaget i ordboka.

Type ressurs	Type URI som navngir ressursen	Kommentar

Type ressurs	Type URI som navngir ressursen	Kommentar
Informasjon på Web-en om «Ting»	Dokument-URI	En informasjonsressurs er i prinsippet en GET-bar ressurs. Det er et dokument (i videste forstand) som er akesserbart på Webenog som omhandler en «Ting» en som er identifiserbar med en ID-URI eller Begreps-URI. Merk at Dokument her kan omfatte både et menneske-lesbart dokument (f.eks. HTML) og et maksin-lesbart dokument (f.eks. RDF).

Forskjellen mellom ID-URI-er og Dokument-URI-er er illustrert i figuren under.

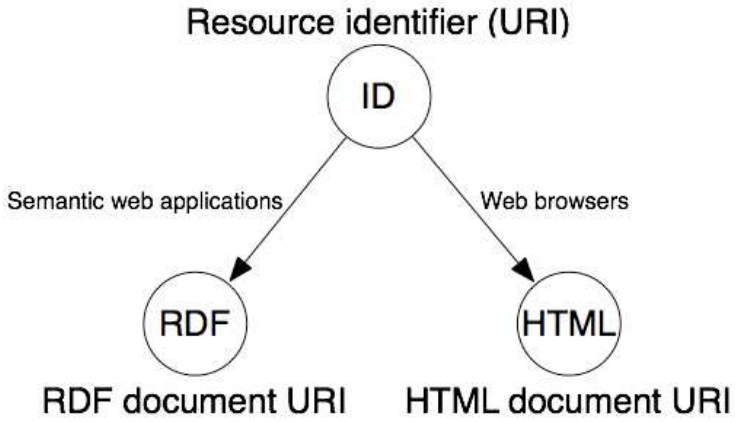


Figure 1 - ID-URI vs. Dokument-URIer

Et URI-sett er en samling referansedata (samling av beskrivelser av «ting») som blir publisert med URI-er.

Type ressurs	Type URI som navngir ressursen	Kommentar
Datasett	Sett-URI	Denne identifikatoren representerer et sett av «Ting».

Vi definerer også relasjons-URI-er.

Type ressurs	Type URI som navngir ressursen	Kommentar

Type ressurs	Type URI som navngir ressursen	Kommentar
Relasjon	Relasjons-URI	Dette er en identifikator for relasjoner mellom begreper, f.eks. del-av (en organisasjon er del-av en annen organisasjon).

Nedenfor finnes flere relevante definisjoner.

Definisjoner

Definisjonene under slik de benyttes i dette dokumentet.

Navn	Beskrivelse
Referansedata	En etat har gjerne en liste over «Ting» de er ansvarlige for eller som inngår i prosesserer de utfører. Disse har en identifiserende referanse som benyttes når man snakker om «Tingen» i dataene. Listen er «referansedata» som gir en felles mening og felles identifikator og refererer til samme «Ting» i etaten. Eksempel: Enheter i Enhetsregisteret, saker i sakssystemet, ansatte i lønns- og personalsystemet,
Ressurs	Ressurser svarer til «Ting» som har en tydelig identitet i en gitt kontekst: steder, mennesker, bøker, hendelser, abstrakte konsepter, forhold mellom ting etc.
URI	Universal Resource Identifier, er en streng av tegn som benyttes som en identifikator på Internett. Typisk vil en URI unikt navngi en «Ting» (‘Ressurs’). En URI kan i Linked Data-sammenheng være både et begrep (gruppe av ‘ting’ som ‘enhet’) eller en ‘ting’ selv, f.eks. en konkret enhet. Eksempel: http://brreg.no/vocab/enhetsregisteret/enhet eller http://brreg.no/id/enhet/123456789
URI-sett	Referansedata som har en felles oppbygging av URI-er som gjør det enkelt å gjenbruke (lenke til). Ofte endrer kun siste ledd i URI-en seg for en konkret «Ting» (‘Ressurs’). Eksempel: Enheter med felles system for URI-er.
Navnerom	Et navnerom er et felles prefiks for URI-er av samme type. Typisk vil prefiks benyttes til å forenkle lesbarheten. Eksempel: http://brreg.no/id/enhetsregisteret/enhet/ prefikses med er: og enheten 123456789 kan da skrives er:123456789

Navn	Beskrivelse
Begrep	En bestemmelse (egenskap, trekk, kjennetegn) eller et kompleks av bestemmelser som karakteriserer eller avgrenser, altså definerer, en klasse av ting. (Store Norske Leksikon). Et begrep i vår sammenheng er også en Ressurs i seg selv. Eksempel: Person
Dokument	Et dokument i denne sammenheng er en beskrivelse av en (eller flere) Ressurser i et gitt format. Se figur Figure 1 - ID-URI vs. Dokument-URIer. Synonym: Informasjonsressurs
Begrepsapparat	En samling av begreper innenfor samme domene, samt forhold mellom disse. Synonym: Begrepssystem, Begrepsmodell, Terminologi
Vokabular	Samlinger av begreper med en klart definert betydning. Ressursene er representert av URI-er som består blant annet av vokabulareierens/-skaperens domenenavn. Synonym: Ontologi (mer formell bruk). Synonym: Begrepsapparat (mer uformell bruk) Merk at det er grader av formalitet som skiller et vokabular fra en ontologi, men begrepene benyttes ofte synonymt. Vokabular benyttes ofte om det er en spesifik bruk av et begrepsapparat der begrepene har klare URI-er. Ontologier benyttes ofte ut over dette der begrepene har formelle logiske beskrankninger, f.eks. en Mann er ikke en Kvinne.
Ontologi	En sammenstilling av et begrepsapparat med logisk formelle sammenhenger. En slik logisk sammenheng mellom to begreper er f.eks. Kvinne «is-A» Person. Med dette menes at noe som er en Kvinne også er en Person. Synonym: Vokabular (mer uformell bruk) Eksempel: Ontologien for Persondomenet
ID-URI	URI-en som unikt identifiserer en ting (abstrakt eller konkret) Eksempel: en bestemt bil, en organisasjon, en person.
Dokument-URI	URI-en til et dokument som beskriver en ting. Merk at dokument-URI-er generelt sett også kan være dokumenter som ikke omhandler en ID-URI eller Begreps-URI. Denne tolkningen er ikke relevant her.
Data	Det laveste abstraksjonsnivået informasjon kan baseres på; som en variabel eller et sett av variablers kvalitative eller kvantitative egenskaper.

Navn	Beskrivelse
Referanse-identifikatorer	En lokal referanse som ikke er oppslagsbar over HTTP (altså ikke en URI). Merk: Referanseidentifikatorer er unik innenfor et domene f.eks. personnummer og organisasjonsnummer de kan også være unike på tvers av domener (f.eks. organisasjonsnummer og personnummer), men ikke oppslagbar over HTTP.
Informasjonsbærende URI	En URI som inneholder klartekstelementer som antyder mening. Eksempel: http://dbpedia.org/resource/Grotten har et vist meningsinnhold gjennom at man kan lese at den er om en ressurs kalt Grotten.
Informasjonsløs URI	En URI som ikke inneholder deler som antyder mening. Det er ofte grader av informasjonsløshet, hvor f.eks. domene-delen fortsatt har en viss antydning av mening, mens leddene etter er infromasjonsløse. Eksempel: Man kan ikke lese av URIen at http://sws.geonames.org/2988507/ refererer til Paris.
Lenkede data	Lenkede Data (engelsk: Linked Data) refererer til et sett av beste praks prinsipper som for å publisere og lenke mellom strukturerte data på Web.
REST	Representational State Transfer (REST) er en modell for web-tjenester basert kun på HTTP. Ved REST-tjenester, i motsetning til tradisjonell web-services, opererer man på URI-en som er det faktiske objektet som leses eller endres.
HTTP	HyperText Transfer Protocol (HTTP) er den grunnleggende protokollen som brukes av World Wide Web.
HTTP innholdsforhandling	En teknikk hvor man i HTTP header angir hvilke formater han forstår, og serveren responderer i henhold til denne forespørselen. Engelsk: Content negotiation Synonym: format-forhandling

Prinsipper og designvalg

I det følgende adresseres prinsipper for oppbygging av URI-er med en drøfting som begrunner prinsippene. Prinsippene er:

Nr.	Navn
1	Eierskap og opphav
2	Sti-struktur i URI-er
3	Levetidsutfordringer
4	Oppslag på URI-er
5	Dokumenter for maskin- og menneskelesbarhet
6	Kvalitetskarakteristikker

Generelt kan man si at et URI-mønster for begreper og data identifikatorer er bygget opp av

http://{domene}/{type}/{term}/{referanse}.{format} for data og
http://{domene}/{type}/{vokabular}/{term}.{format} for begreper

Se tabell side 24 for nærmere beskrivelse. Om en sammenlikner med ISO 11179-6 [5] for en slik struktur, vil den kunne se ut som følger:

http://{domene}/{begrep}/{versjon}

Der {begrep} inkluderer feltene {type}, {term} og {referanse} eventuelt {type}, {vokabular} og {term}. Domene må her være en organisasjon som har autorisasjon til å definere og publisere begreper innen domentet, eventuelt delegerere.

Prinsippene er knyttet til denne strukturen som følger:

- Prinsipp 1 knytter seg direkte til domene-delen av et URI-mønster. 2 dreier seg om oppbyggingen av den resterende URI-delen: sti-strukturen, mens 3 tar for seg levetidsutfordringer i tilknytning til denne.

- Prinsipp 4 omhandler hva URI-oppslag skal referere til: dokumentrepresentasjoner av ID-ressurser og hva som skjer i bakkant av slike oppslag, mens prinsipp 5 beskriver hvilke representasjons-formater oppslag på dokumenter skal støtte.

- Prinsipp 6 beskriver kvalitetskarakteristikker som angår et konkret datasett: kvaliteten av det URI-mønsteret omhandler.

Prinsipp 1: Eierskap og opphav

Formulering	Stabile/permanente domener er nødvendig for å sikre at oppslag på en URI viser en beskrivelse av en ressurs uansett om oppslaget gjøres i dag, i morgen eller om 10 år.

Begrunnelse	Du kan ikke garantere levedyktigheten til domener utenfor din egen etats kontroll. Data forvaltet av en etat kan i framtiden flyttes til en annen etat. Navn på etater er dessuten ikke stabile. Eks: Moderniseringsdepartementet → Fornyings- og administrasjonsdepartementet → Fornyings-, administrasjons- og kirkedepartementet
Konsekvenser	• Ikke benytt navnerom/domener utenfor din kontroll • Unngå å vise eierskap i domenet i så stor grad som mulig

Hovedprinsippet her er å sikre stabilitet i URI-ene når man velger hovedkomponenten for URI-en – domenenavnet.

Permanente URI-er

Permanente ID-URI-er er viktig for å sikre at informasjon om ressurser som viderebrukes av andre alltid vil være tilgjengelig. Med «alltid» menes å sikre at den har mest mulig stabilitet f.eks. uavhengig av omorganisering av offentlig sektor. F.eks. vil domenenavnet «norge.no» sannsynlig være mer stabilt enn «difi.no».

Når et datasett kobles til et annet datasett der ID-URI-ene ikke er permanente, vil man potensielt (videre-) bruke URI-er som det en gang i framtiden ikke lenger vil være mulig å gjøre oppslag på. Et slikt oppslag vil da gi en HTTP-404-respons fra serveren – en beskjed om at ressursen ikke lenger er tilgjengelig. For å tydeliggjøre hva en ID-URI identifiserer og slik skape forståelse for hvordan den bør brukes i sammenstilling med andre data er det viktig at HTTP-oppslag på URI-en i det hele tatt returnerer data.

Merk at dette står i motsetning til den umiddelbare troverdigheten/tilliten som ligger i domenenavnet til en etat.

Hold deg unna navnerom du ikke kontrollerer

Tom Heath og Christian Bizer beskriver i [6].noen retningslinjer for utforming av gode URI-er En av disse retningslinjene dreier seg om viktigheten av å unngå navnerom en ikke selv kontrollerer ved utforming av URI-er.

Ved å bruke et domene man ikke har kontroll over som del av en URI har man ingen mulighet for å selv gjøre URI-en oppslagsbar. Derfor bør man ikke definere URI-er i domener man ikke kontrollerer. Man kan derimot lenke til dem.

Eierskap vist i URI-er

De fleste datasett har en tydelig eier, enten det er en offentlig etat, et firma eller en privatperson. F.eks. har Enhetsregisteret en eier i Brønnøysundregistrene. Dataeieren forvalter datainstansene som inngår i et datasett og modellerer begrepene som inngår i et begrepsapparat. Det kan imidlertid være politisk-, og troverdighetsmotivert å beholde et element av eierskap i URI-ene.

URI-en kan gjenspeile eierskap dersom det er viktig å tydeliggjøre hvem dataene har opphav i. Synlig eierskap kan også bidra til at URI-en blir mer meningsbærende ved at eierens navn er knyttet til dataenes emnetilhørighet.

Dataenes troverdighet henger også sammen med opphav i, f.eks. vil en ontologi modellert av en faglig tungtveiende aktør som World Wide Web Consortium i de fleste tilfeller ha større gjennomslagskraft enn en tilsvarende ontologi modellert av en privatperson. Samtidig finnes det ontologier modellert av "små" aktører som har blitt ledende innen enkelte områder. Et eksempel på dette er FOAF (Friend-of-a-Friend), et vokabular som brukes for å modellere personer og forhold mellom dem. Dette vokabularet framstår som nøytral og emneorientert og vil kanskje derfor være mer nærliggende å bruke enn en tilsvarende vokabular som var modellert av en kommersiell aktør som f.eks. Facebook.

I rapporten "Designing URI Sets for the UK Public Sector" [3] vektlegges det at ID-URI-er i UK bør dannes med basis i domenet data.gov.uk med sektortilhørighet som subdomene, for eksempel skal utdanningsdata identifiseres under domenet http://education.data.gov.uk. URI-ene skal ikke inneholde navnet på etaten som definerer dataene da dette kan endres. Oppslag på URI-ene kan gi respons enten på data.gov.uk selv eller ved at DNS brukes for å redirecte oppslaget til etaten som forvalter dataene sin server [3, s. 6]. Videre defineres termen "URI Set" som en samling av data publisert med URI-er av samme aktør innen samme domene, f.eks. skoler eller veier. Metadata om slike URI-sett skal foreligge og inneholde blant annet informasjon om presisjon på dataene, hvor lenge deres levetid garanteres og provenance (dataenes opphav/eierskap) og mål [3, s. 5]. Slik metadata bør være endel av RDF-beskrivelsen av settet. (se mer under prinsipp 6 kvalitetsbeskrivelser).

I Storbritannia anbefales det altså at eierskap til offentlig sektors data ikke reflekteres i identifikatorene (ID-URI-ene), men at rot-domenet data.gov.uk benyttes for å tydeliggjøre at de har tilhørighet til det offentlige og at metadata skal brukes for å beskrive blant annet eierskap.

En liknende måte bør velges i Norge der det foreligger en sektorinndeling, med f.eks. basis i data.norge.no som rot-domene. En slik URI vil bygges opp ved at {domene} blir erstattet av {sektor}.{permanent-domene}, f.eks. helse.data.norge.no. Det ser viktig at noen påtar seg en forvaltningsoppgave for fellesdomener som helse.data.norge.no

Untakksmessig bør det være akseptert å reflektere eierskap direkte i ID-URI-ene som brukes slik som i Enhetsregisteret og Kartverket i UK (Ordanance Survey) http://data.ordnancesurvey.co.uk/id/50kGazetteer/218015. En slik URI vil bygges opp ved at {domenet} er etatens eget domenenavn, f.eks. brreg.no. Man bør imidlertid her påkrevve at man overfor sentral forvaltningsaktør (f.eks. Difi) argumenter for hvorfor et slik unntak velges.

Prinsipp 2: Sti-struktur i URI-er

Formulering	Definere en struktur som ivaretar: • Balanse mellom informasjons-bærende og informasjons-løse URI-er • Identifisering av en gruppe av instanser (knyttet sammen via et «begrep») • Identifisering av en konkret instans («ting») med lokal referanse • Identifisering av type-dokument «om» konseptet (menneskelesbart og maskinlesbart)

Begrunnelse	En utbredt sti-struktur for URI-er vil bidra til å forenkle prosessen med å få oversikt over- og arbeide med URI-er på tvers av etater. Man vil også kunne si noe om sett av URI-er f.eks. knyttet til datakvalitet og eierskap.
Konsekvenser	Mal for utforming av (domene +) sti-struktur i URI-er • ID-URI (instans): http://{domene}/id/{term}/{referanse} • Dokument-URI for instanser (menneskelesbar): http://{domene}/page/{term}/{referanse} • Dokument-URI for instanser (maskinlesbar): http://{domene}/data/{term}/{referanse} • ID-URI (begrep): http://{domene}/vocab/{vokabular}/{term} • Dokument-URI for begreper (menneskelesbar): http://{domene}/page/{vokabular}/{term} • Dokument-URI for begreper (maskinlesbar): http://{domene}/data/{vokabular}/{term}

Hovedprinsippet her er å etablere en struktur som sikrer en gjenkjennbar struktur på URI-er.

Type

Vi har tidligere nevnt behovet for å skille mellom ID-URI-er og Dokument-URI-er. Disse ulike typene fremgår av det første elementet i en sti-struktur. Som minimum bør dette dekke:

- ID-URI-er for begreper
- ID-URI-er for konkrete «ting» (instanser)
- Dokument-URI for menneskelesbare dokumenter
- Dokument-URI for maskinlesbare dokumenter
- Sett av URI-er

De tilsvarende verdiene for {type} som anbefales er derfor *id*, *vocab*, *page*, *data*, *set.*

Gruppering av instanser

Det er nyttig å angi typen til en gruppe av instanser. Dette øker lesbarhet. Dette bør som angitt i [3] være et element fra virkeligheten. Dvs. f.eks. «enhet»

http://domene.no/id/enhet/123456789

Merk at man bør trå varsomt her, og ikke forsøke å gjenoppbygge ontologien (nedbrytingen av begepet) på denne måten der identifikatoren er ment å representere den mer generelle tolkningen.

For eksempel kan en enhet brytes ned i ulike typer enheter, for eksempel aksjeselskap, frivillig organisasjon osv. Man må unngå situasjoner som

http://domene.no/id/aksjeselskap/123456789

For det første ender man opp med unødvendige mange identifikatorer på samme ting. Dessuten står man i fare for å lage en URI som har mindre stabilitet ved at reglene for om denne enheten er et aksjeselskap kan endres, og URI-en dermed ikke lenger er gyldig.

Informasjons-bærende og informasjons-løse URI-er (referanse-element)

At en ID-URI gir informasjon i seg selv er viktig for at den skal være enkel å gjenbruke for mennesker, spesielt gjelder dette for URI-er som navngir et begrep. Hvis det kommer tydelig frem fra URI-en hva den identifiserer, reduseres sannsynligheten for at den brukes i "feil" sammenheng. Den mest brukte navnet på begrepet (såkalt preferert term) brukes derfor i URI-en som informasjons-bærende element.

For instansdata vil det normalt være kombinasjonen av {term} og {referanse} som utgjør informasjons-bærende element. I endel tilfeller kan det være det vanskelig å lage en informasjons-bærende ID-URI for {referanse}-leddet. Det vil for eksempel være problematisk å lage informasjons-bærende ID-URI-er for endel abstrakte ting, slik som en instans av bistand mellom en giver og mottaker. I slike tilfeller er det dermed mest naturlig å konstruere en egen numerisk identifikator for hver instans innen et datasett.

Basert på erfaringer fra IT-industrien generelt og database design spesielt så bør en være varsom med å legge betydning inn i deler eller hele verdien til en referanse-identifikator. Det er flere eksempler på at det kan koste dyrt når en på et senere tidspunkt må endre. Fødselsnummer er et eksempel på en referanse-identifikator som bærer mening i ulike ledd av identifikatoren, denne skaper problemer ved personers endring av kjønn, siden et siffer indikerer nettopp kjønn. Tilsvarende “Numerisk adresse”, en identifikator for adresser, som svikter fordi den inneholder kommunenummer og vegnummer, og det skaper problemer ved sammenslåing av kommuner. Hvis samme vegnummer er i bruk i begge kommunene blir det kollisjon.
Det kan derfor argumenteres for at referanse-identifikatorer ikke bør inneholde noen form for betydning. Dette gir noen utfordringer for mennesker som skal lese identifikatorene, men betyr lite for maskiner som allikevel må slå opp for å finne betydningen. Ved å legge noe meningsbærende inn i en referanse-identifikator påvirker URI-regimet strukturen for begreper og eventuell kategorisering og en slik kobling bør en unngå.

Merk at {term} alltid skal angis i entallsform.

Identifisering av en konkret instans

Den konkrete instansen identifiseres med dens referanse delen av en URI (som for eksempel kan være en primærnøkkel i databasesystemet). Det er viktig at denne gir best mulig mening. I de fleste tilfeller der man har unike navn bør disse benyttes. F.eks. anbefaler [3] at der navnet blir langt og omfattende, bør dette unngås.

http://domene.no/id/fylke/Sogn-%20og%20Fjordande

Der man ikke har unike navn, men har unike referanse-felter, som f.eks. kommuner der de første sifferene angir hvilke fylke kommunen ligger bør dette benyttes.

http://domene.no/id/kommune/0926

for Lillesand. Merk forøvrig at dette siste eksempelet også viser at man kan komme til å falle for fristelsen å bryte dette generelle prinsippet og bygge opp dette som en part-of (hierarkisk) struktur som følger

http://domene.no/id/geo-inndeling/norge/09/26

Selv om dette er god skikk innen REST-arkitekturer, kan det skape en uheldig situasjon dersom kommunen skulle endre fylke – noe som ikke er særlig hypotetisk jfr. folkeavstemninger ved siste valg (Mer om versjonering under prinsipp 3). En annen utfordring oppstår dersom inndelingen kan gjøres på ulike vis, f.eks. gjennom landsdel.

Skille mellom ressurser og dokumenter

Det er en vesentlig forskjell mellom en ressurs som kan angis med en ID-URI og et dokument som beskriver ressursen, en såkalt dokument-URI. Hvis vi gjør oppslag på en ressurs kan vi ikke returnere selve ressursen, men kun en beskrivelse av den. Dette bør gjenspeiles i URI-strukturene som navngir henholdsvis ressursen (ID-URI) og dokumentene som beskriver den (dokument-URI-ene).

Vanlig praksis, blant annet i DBPedia, er å introdusere et skille mellom URI-er som tilbyr et dokument egnet for lesing av mennesker (HTML) og maskinlesbare dokumenter (RDF) i tillegg til skillet mellom ID-URI-en og dokument-URI-ene. Slik dannes et URI-mønster bestående av tre deler.

- ID-URI
- Dokument-URI (RDF)
- Dokument-URI (HTML)

I DBpedias URI-er gjenspeiles dette skillet på følgende måte:

- ID-URI: http://dbpedia.org/resource/Grotten
- Dokument-URI (HTML): http://dbpedia.org/page/Grotten
- Dokument-URI (RDF): http://dbpedia.org/data/Grotten

Med en standardisert måte å gjøre disse skillene på, vil det være lettere å huske hvordan en skal identifisere en ting kontra å se en dokumentbeskrivelse av tingen. For eksempel kan en tenke seg at en ser på en HTML-dokumentbeskrivelse av en ressurs, http://brreg.no/page/enhet/974760673, og ønsker å kopiere denne URI-en fra nettleseren inn i en editor hvor man ønsker å bruke ressursen som beskrives sin ID-URI i et RDF-trippel. I et slikt tilfelle vil man måtte endre URI-en man får ved innliming slik at det ikke lenger står "page" i den, men det ordet som er med på å skape identifikatoren. Det kan være vanskelig å huske hvordan URI-en skal endres fra være en dokument-URI til å være en ID-URI igjen når det ikke brukes en standard struktur, derfor er det fornuftig å ha bestemte konvensjoner her.

Dersom "id" alltid brukes som del av en ID-URI, mens "page" eller "data" brukes i dokument-URI-er (og dette er det eneste som skiller ID- og dokument-URI-ene fra hverandre), minsker risikoen for at URI-ene brukes "feil". *Id* er anbefalt valgt framfor DBpedias' *resource* for å tydeliggjøre at det er identifikatoren det er snakk om. *Resource* kan sies å være et mer generelt ord enn *id* og dermed misvisende da dokumenter også er ressurser.

I [6, paragraf 4.1.3] belyses en svakhet med den ovennevnte måten å skille mellom ID-URI-er og dokument-URI-er på. Det er vanskelig å se – spesielt for de som ikke er vant med denne strukturen – at et oppslag på en ID-URI i en nettleser leder til en annen URI (dokument-URI) da de ulike URI-ene er såpass like. Dermed er det fort gjort å bruke en dokument-URI som navn (ID-URI) f.eks. når en kopierer URI-en fra adresselinjen. En alternativ, mer visuelt distinkt struktur gitt av subdomenbruk lanseres derfor i [6]:

• ID-URI: http://id.{domene}/{id}
• Dokument-URI (HTML): http://pages.{domene}/{id}
• Dokument-URI (RDF): http://data.{domene}/{id}

I Data.gov.uks dokument "Creating URIs" ([3]) lages et eget skille for hvordan begrepers URI-er bør skapes kontra instansdata. Vi anbefaler at slike ID-URI-er skapes analogt med ovennevnte måter for instansdata, men med *vocab* istedenfor *id* som type for å illustrere at det er snakk om et begrepsapparat:

- ID-URI: http://{domene}/vocab/{begrep}/{referanse}

Prinsipp 3: Levetidsutfordringer

Formulering	Det er et poeng at URI-er er permanente i størst mulig grad. Dette for å sikre at informasjon om ressurser som viderebrukes av andre i dag også skal finnes tilgjengelige i framtiden. I prinsipp 1 ble utfordringer i forbindelse med dette knyttet til domenetilhørighet. I *dette* prinsippet knyttes levetidsutfordringer til URI-enes sti-struktur og historiske utvikling: versjonering.
Begrunnelse	Implementasjonsdetaljer, slik som servers portnummer, filformatangivelser er ustabile deler av en sti-struktur som kan komme til å endres i takt med (teknologisk) utvikling.
Konsekvenser	• Unngå implementasjonsdetaljer i URI-ene • Versjonering, dato, bør derfor kun inkluderes i en ID-URI dersom det anses som sannsynlig at betydningen til det URI-en navngir vil endres med tiden.

Unngå implementasjonsdetaljer i URI-ene

Detaljer om den tekniske implementasjonen bak serveren som støtter URI-oppslag bør utelates fra URI-ene.

Informasjon som eksempelvis portnummer og detaljer om implementasjonsspråk slik vist i http://example.com:8080/education/schools.php?id=3211 må derfor ikke forekomme. Dette fordi en URI der slik informasjon inngår vil endres i takt med utbedring og utskiftning av teknisk løsning og dermed ikke være permanent. Dessuten gjør teknisk informasjon URI-ene unødvendig lange og vanskelige å lese. Prinsippet er også diskutert i [6].

Dette gjelder også for protokoller. Man kan falle for fristelsen til å lage en URI som benytter https fordi trafikken skal krypteres. Det må likevel benytte http til identifikatoren som eventuelt kan redirecte til https ved oppslag.

Et URI-mønster der skillet mellom ID-URI-er og dokument-URI kun er gitt av et dokumentformattillegg til slutt i en URI går igjen i flere datasett.

- ID-URI: http://brreg.no/vocab#orgnr
- Dokument-URI (HTML): http://brreg.no/vocab.html
- Dokument-URI (RDF): https://brreg.no/vocab.rdf

Denne typen skille er kanskje mest vanlig i hash-URI-er (se prinsipp 4); hvor dokument-URI-ene tilsvare ID-URI-en før hashen (#) strippes vekk før sending til serveren. Man kan da ikke eksplisitt kan referere til en dokument-URI uten å angi format. Format kan her sies å være en implementasjonsdetalj, og dermed noe en i framtiden kan ønske å endre. Formatangivelse på slutten av URI-ene bør derfor ikke være eneste skille mellom ID-URI-er og dokument-URI-er.

Versjonering

Normalen er at et begrepsapparat er stort sett stabilt over tid – nye begreper kan komme til, men dette er ikke problematisk da ID-URI-ene til begrepene som finnes fra før forblir uendret. På tross av dette kan betydning til og anvendelse av et begrep vil være i drift over tid, f.eks. «Rusmiddel», «Media». Informasjonsbehandling og saksbehandling i offentlig sektor må også håndtere endring i lover og regler som angår offentlig sektor, f.eks. «Ekteskap». God design av Begreps-URI'er bør ha som formål å begrense utfordringen ved semantisk drift.

På instansnivå har vi også stabilitets-utfordringer. I Norge er en kommune et godt eksempel på en type instans som kan være ustabil. Kommuner slås sammen og noen endrer navn. Selv om kommunenavn trolig vil fungere som den mest informasjons-bærende delen av en URI som navngir en kommune, betyr ikke dette at det er den beste identifikatoren som kan brukes. Anta at vi hadde ID-URI-en http://domene.no/id/kommune/Barbu for å identifisere kommunen Barbu og at Barbu slo seg sammen med Arendal kommune (noe som skjedde i 1902). Skulle denne URI-en da refererer til den gamle kommunen Barbu som ikke fantes lenger, eller den nye som var en del av Arendal? Hva med identifikatoren til Arendal? Dette bringer diskusjonen frem et annet designvalg, versjonering: hva skal gjøres, hvis noe, for å modellere versjonsendringer i en URI?

FOAF-vokabularet inkorporerte versjons-informasjon i URI-ene sine. Versjonsnavnet, 0.1, står derfor fortsatt som en del av URI-ene for hvert begrep i FOAF: http://xmlns.com/foaf/0.1/. Dette til tross for at vokabularet har vært i utvikling siden denne versjonen. Som FOAF beskriver selv, står dette til advarsel for andre som vurderer å blande metadata inn i identifikatorene sine [http://xmlns.com/foaf/spec/#sec-evolution]:

> «We are left with the digits "0.1" in our URI. This stands as a warning to all those who might embed metadata in their vocabulary identifiers.»

Dette støtter opp under Data.gov.uks måte å skape URI-er på hvor metadata skilles fra URI-ene ut til et såkalt URI-sett (mer om dette i avsnittet om domener), som igjen henger sammen med ønsket om permanente URI-er: en URI bør forbli uendret etter den er laget [3, s.8].

Tennison argumenterer i [8] for at ID-URI-er ikke må inneholde informasjon det er sannsynlig at kan komme til å endres. Derfor består en rekke ID-URI-er av tall istedenfor tekststrenger til tross for at dette gjør dem mindre meningsbærende. Tim Berners-Lee tar dette enda lenger. I notatet fra 1998, "Cool URIs don't change" [9] presiserer Berners-Lee at all informasjon som inngår i en URI er med å skape trøbbel. Til og med type-informasjon (altså termen som angir typen til instansen, f.eks. «person»). Dette begrunner han med at ved å bruke et type-navn i en URI binder du deg til en klassifikasjon du i framtiden vil kunne ønske å endre. Et unntak gjøres imidlertid. Datoen en URI ble laget på er en type informasjon som ikke vil endres i følge Berners-Lee. «I Cool URIs» foreslår han derfor at dato bør inngå som del av en URI. ID-URI-ene definert av W3C, f.eks. i RDF-vokabularet, inneholder derfor dato: http://www.w3.org/1999/02/22-rdf-syntax-ns#type.

Ved å bruke dato (årstall) kunne man enkelt skille mellom Arendal kommune før og etter den ble slått sammen med Barbu: http://domene.no/id/kommune/1610/Arendal og http://domene.no/id/kommune/1902/Arendal. I følge Tennison [8] virker det dog som at sannsynligheten for at et term blir brukt med en annen mening i framtiden er så liten at det ikke er verdt å inkludere dato da det vil svekke URI-ens lesbarhet og føre til økt behov for vedlikehold. Versjonering, dato, bør derfor kanskje kun inkluderes i en ID-URI dersom det anses som sannsynlig at betydningen til det URI-en navngir vil endres med tiden?

Til tross for Berners Lees notat, kan man med fordel inkludere type-informasjon, i URI-er der det ikke anses som sannsynlig at type-navnet skal bli foreldet, f.eks. «person», «enhet». Dette fordi denne typen er sentralt for å gjøre en URI meningsbærende og fordi det muliggjør oppslag på samlinger gjennom URI-sett.

Dersom man har sterkt behov for å referere til konkrete versjoner av en ID-URI som endrer seg hyppig, bør man som også foreslått av ISO 11179-6 ha versjon som siste ledd i sine identifikatorer. Da kan man operere både med en versjonsløs og en versjonert variant. Dette bør imidlertid unngås om mulig.

Prinsipp 4: Oppslag på URI-er

Formulering	URI-er benyttes ikke kun for å generere globale primærnøkler, men også for å kunne lage RESTfulle tjenester som kan slå opp på disse. Oppslag er også kjent som de-referering av URI. URI-oppslag bidrar til å bre kjennskapen og tilgangen til ressursen og deres definisjon (begrepet) og at andre relaterte ressurser kan oppdages.
Begrunnelse	303-URI-er muliggjør oppslag på en identifikator, mens beskrivelsene i ulike serialiseringer (dokument-URI-er) skilles ut.
Konsekvenser	ID-URI-er derefereres til dokument-URI-er ved hjelp av 303-URI-er. Man unngår bl.a. situasjonen at et selskap identifiseres med sin hjemmeside.

Dereferbare ressurser – dokumentoppslag

Ressurser navngis med ID-URI-er og bør være dereferbare. At de er dereferbare betyr at HTTP GET oppslag på ID-URI-ene skal resultere i dokumenter – med egne dokument-URI-er – som tilbyr mer informasjon om ressursene og eventuelle andre tilknyttede ressurser. Dette i henhold til Berners-Lees retningslinjer for lenkede data. Slike dokumenter bør finnes både i formater som er lettleselige for mennesker: vanligvis HTML, og som maskinlesbare dokumenter, RDF. "Designing URI Sets for the UK Public Sector" anbefaler at ID-URI-oppslag bør resultere i dokumentrepresentasjoner av ressursen i forskjellige formater og at disse formatene bør lenke til de andre formatene som eventuelt foreligger som representasjon av ressursene.

Slik spesifisert i World Wide Web Consortiums notat "cool URIs for the semantic web" [7], er det to ulike måter å gjøre URI-er dereferbare på. Enten ved hjelp av 303-URI-er, eller som Hash-URI-er.

Redirection - 303-URI-er

Med denne teknikken vil et HTTP-oppslag returnere HTTP responskoden "303 - see other". 303-URI-er kalles også for Slash-URI-er da slash ("/") brukes for å unikt identifisere en ressurs.

Etter 303-responsen følger en URL til et dokument hvis format avhenger av hvilken innholdstype (content type) som er forespurt, gjennom HTTP-innholdsforhandling. En innholdsforhandling baseres på hva slags MIME-type(r) som spesifiseres i en aksepthode til en HTTP-forespørsel og hvilken klienttype som står for forespørselen. En ressurs returneres i det høyest prioriterte dokumentformatet, MIME-typen, som serveren støtter [10, s. 42].

Et oppslag på ID-URI-en: http://domene.no/id/enhet/123456789 vil med content type *text/html* returnere en 303-respons til en dokument-URI http://domene.no/page/enhet/123456789 som gir en html representasjon.

Et tilsvarende oppslag med content type application/rdf+xml resulterer i en 303-respons til en dokument-URI http://domene.no/data/enhet/123456789 og returnere et maskinlesbart dokument i RDF/XML-format.

I følge [7] bør nevnte variant å lage en 303-URI på velges når de forskjellige datatypene – RDF og HTML (+ evt. andre) – har forskjellig innhold, f.eks når HTML-dokumentet inneholder mer informasjon enn RDF-representasjonen. Dersom de ulike formatene har det samme innholdet – samme dokument, forskjellig format – bør en annen variant velges der URL-en som følger 303-responsen peker til en felles (abstrakt) generisk URL hvor resultatet av et oppslag igjen avhenger av innholdsforhandling. Da to dokumentrepresentasjoner av samme ressurs i forskjellige formater i de fleste tilfeller bør være like innholdsmessig er det denne varianten å gjøre 303-redirects på som vanligvis bør velges og som anbefales i [8]. I praksis ser det imidlertid ut til at den første varianten er vanligst å bruke, selv når det ikke er vesentlige forskjeller i innholdet mellom de ulike dokumenttypene som kan leveres ved innholdsforhandling.

Hash-URI-er

Generlet kan en URI kan inneholde et hash-symbol ("#") med innhold etter dette. For eksempel: http://vocab.lenka.no/hvor#adresse. Ved oppslag på en slik ID-URI krever HTTP-protokollen, i følge [7], at delen etter hash-symbolet fjernes før URI-forespørselen sendes serveren, noe som vil gjøre dem til dokument-URI-er. Dette skjer automatisk av klienten og betyr at en Hash-URI ikke kan slås opp på direkte og dermed heller ikke identifisere et web-dokument (HTML eller RDF f.eks.) De kan derimot brukes for å identifisere ressurser som ikke er web-dokumenter. Hash-URI-er brukes altså utelukkende som ID-URI-er mens HTTP-oppslag på disse URI-ene gir dokumenter som beskriver dem. Flere dokumentformater (i det minste HTTP og RDF) bør støttes ved oppslag, og hvilket av dem som returneres avgjøres ved innholdsforhandling.

303 versus hash – hva bør velges?

Hash har fordelen av færre nødvendige HTTP-oppslag og dermed mindre overhead enn 303. Et oppslag på en ressurs med en hash-URI vil returnere et dokument som representerer alle ressurser som deler navn med URI-en det gjøres oppslag på fram til hash-symbolet. Dette skjer fordi serveren ikke behandler delen etter hashen. I enkelte tilfeller kan dette være ønskelig, f.eks. i en ontologi der man ønsker å raskt få oversikt over et begreps relasjon til tilknyttede begreper som også inngår i ontologien. I andre tilfeller, der man kun ønsker å få informasjon om ressursen man slår opp på, gir hash-URI-oppslag mer informasjon en ønskelig. For store datasett vil hash-oppslag dessuten returnere veldig mye data, noe som både kan være uheldig med tanke på datamengden som returneres klienten ved oppslag og hvor enkelt det vil være for en person å skaffe seg oversikt over den tilgjengelige informasjonen om termen han slo opp på.

303 er mer fleksibelt med tanke på hva et oppslag kan returnere. Et oppslag på ID-URI-en http://linkedgeodata.org/ontology/hasCity returnerer f.eks. kun en dokumentrepresentasjon av akkurat denne ressursen, mens et oppslag på http://linkedgeodata.org/ontology returnerer informasjon om alle andre ressurser som inngår i denne ontologien. Dersom linkedgeodata hadde navngitt ressursene sine med hash-URI-er, ville kun et oppslag med sistnevnte funksjonalitet være mulig.

Jeni Tenisson nevner noen praktiske vanskeligheter med 303-URI-er i [11]. Blant annet at det kreves tilgang til web server-konfigurasjonen for å legge til 303 redirects, noe som hever terskelen for hvem som kan konstruere dem.

Det er med tiden blitt mer etablert praksis å benytte 303-URI-er, noe vi også anbefaler.

REST, URL-navigasjon som generell API

For instanser er det typen (dvs. {term}) og eventuell instans-referanse som brukes for å identifisere en bestemt ressurs innenfor et navnerom. http://domene.no/id/{term}/{referanse}. Ved hjelp av slike skjemaer identifiseres ulike instanser av data med samme type, f.eks. ulike enheter: http://brreg.no/id/enhet/123456789 og http://brreg.no/id/enhet/987654321. Ved å fjerne instans-ID-en fra URI-en kan man i RESTfulle systemer referere til en samling instanser av en bestemt type: http://brreg.no/id/enhet [10, s.42]. Et oppslag på instansnivå bør resultere i en representasjon av denne instansen, mens et oppslag på samlingsnivå bør returnere en representasjon av alle tilhørende instanser. Tilsvarende for begreper.

Prinsipp 5: Dokumenter for maskin-lesbarhet og menneskelesbarhet

Formulering	Det er under *URI-oppbygging* nevnt hvordan man kan skille mellom menneske- og maskinlesbare formater. Dette prinsippet omhandler hvilke slike formater som bør støttes.
Begrunnelse	Det er sentralt for mennesker å raskt kunne få oversikt over hva en ID-URI identifiserer. Dette kan de få gjennom en pent formatert HTML-side. For maskiner er det sentralt å få tolket semantikken i dataene som beskrives i et oppslag mot en ID-URI og å ha mulighet til å sammenstille dataene med- og å følge lenker til andre datakilder. RDF tilbyr dette.
Konsekvenser	En anbefaling til hvilke formater som bør kunne leveres ved ID-URI-oppslag

Det følgende uttrykker hvilke formater (content-types) som bør støttes som respons på oppslag på ID-URI-er. Blant annet [3] nevner en rekke vurderinger på dette feltet. Basert på disse, og norske forhold er det viktig å støtte som et minimum formater som er i utstrakt bruk.

Dokumenter for menneskelesbarhet

- URI- oppslag (med content negotiation) må støtte HTML
- URI- oppslag (med content negotiation) bør støtte HTML+RDFa

Dokumenter for maskinlesbarhet

- URI- oppslag (med content negotiation) må støtte RDF/XML og Turtle (Turtle er i ferd med å overta for RDF/XML som primær-formatet for RDF)
- URI- oppslag bør videre kunne støtte JSON
- URI- oppslag på begreper bør støtte OWL og XMI (UML eller SERES metamodell)

Prinsipp 6: Kvalitetskarakteristikker

Formulering	Et datasett (URI-sett) bør ha sin egen URI som beskriver kvalitetskarakteristikker for datasettet. Oppslag på en slik URI bør resultere i et eget dokument som beskriver datasettet.
Begrunnelse	Ved å unngå eierskap i URI-er trengs andre mekanismer for å beskrive eierskap, kvalitet osv.
Konsekvenser	URI-sett gir mulighet for å si noe om hele datasettet.

Identifisere datasett og egenskaper ved datasettet

Kvalitetskarakteristikker legges på datasettet som sådan, denne identifiseres ved følgende mønster

Sett-URI: http://{domene}/set/{term}

URI'en peker på en ressurs som beskriver egenskaper ved datasettet. Disse egenskapene er typisk datakvalitet- og eierskapsinformasjon av ulik art.

Den følgende informasjonen bør fremkomme ved oppslag på sett-URI.:

Innhold	Kommentar
Begrepsdefinisjon	En OWL representasjon (ved RDF oppslag) eller en HTML representasjon av Begrepet. F.eks. http://brreg.no/set/enhetsregisteret/enhet vil medføre en OWL-klasse som representerer en enhet. f.eks. http://seres.no/Brønnøysundregisterene/Enhet
Eierskap	Kilden til referansedataene f.eks. Enhetsregisteret
Status	Status, dvs. ulike forvaltingsmessige statuser URI-ene i settet kan ha
Nøyaktighet	Hvor god er nøyaktigheten i datasettet
Kompletthet	Hvor komplett er settet av URIer i forhold til virkeligheten som ligger bak definisjonen av begrepet. f.eks. inneholder det alle Enheter?
Oppdaterthet	Hvor tidsmessig korrekt er dataene i forhold til virkeligheten?
Lisensbetingelser	Hvilke lisenser er dataene i settet underlagt?
Påtenkt levetid	Hvor lenge antas det at datasettet er tilgjengelig?
Påtenke brukere	Hvem kan antas å ha nytte av datasettet?
Tilgjengelige representasjoner	Beskrive mengden av fil-formater URI-ene i settet finnes i

En rekke vokabularer finnes for å beskrive denne type informasjon. Dublin Core-vokabularet [16] inneholder endel predikater som er velegnede til å angi eierskap i metadata-beskrivelser. Et annet mer ekspressivt vokabular, The Open Provenance Model [17] kan også brukes. Det foreligger også et vokabular kalt VoID som beskriver denne type informasjon kalt VoID, Vocabulary for Interlinked Datasets [13] som inneholder aspekter av denne type beskrivelser.

Hvordan disse vokabularene kan benyttes for å beskriv denne type informasjon, vil bli behandlet i et eget dokument som oppfølging til dette dokumentet.

Anbefaling

Anbefalingen for URIer for «ting», «dokumenter» (dokumentressurser) og definisjoner (begrep) er gitt i tabellen under. Merk at det her er anbefalt en sentralt URI-regime/domene som i UK, men også en lokalt alternativ hvor etatens domenenavn benyttes.

URI-er for Data

URI-mønster	Beskrivelse
http://{sektor}.data.norge.no/id/{term}/{referanse}[/{versjon}] (anbefalt) eller http://{sektor}.{autoritet}.no/id/{term}/{referanse}[/{versjon}] (alternativ)	URI for „Ting"
http://{sektor}.data.norge.no/doc/{term}/{referanse}[/{versjon}] (anbefalt) eller http://{sektor}.{autoritet}.no/doc/{term}/{referanse}[/{versjon}] (alternativ)	URI for dokumenter (Informasjons ressurser - herunder HTML og RDF-dokumenter)
http://{sektor}.data.norge.no/set/{term} (anbefalt) eller http://{sektor}.{autoritet}.no/set/{term} (alternativ)	URI for data-sett

URI-er for Begreper

URI-mønster	Beskrivelse
http://{sektor}.data.norge.no/vocab/{vokabular}/{term} (anbefalt) eller http://{sektor}.{autoritet}.no/vocab/{vokabular}/{term} (alternativ)	URI for definisjoner (Klasser, Egenskaper, Begrep)

URI-mønster	Beskrivelse
http://{sektor}.data.norge.no/doc/{term}/{referanse}[/{versjon}] (anbefalt) eller http://{sektor}.{autoritet}.no/doc/{vokabular}/{term}[/{versjon}] (alternativ)	URI for dokumenter (Informasjons ressurser - herunder HTML og RDF-dokumenter)
http://{sektor}.data.norge.no/set/{vokabular} (anbefalt) eller http://{sektor}.{autoritet}.no/set/{vokabular} (alternativ)	URI for data-sett

Beskrivelse av URI-elementer

URI element	Beskrivelse
{sektor}	Navn på en sektor som et begrep og relaterte referansedata forvaltes f.eks. transport, utdanning, miljø
{term}	Et sektor-spesifikt begrepsnavn for typen entiteter («Ting») som referansen angir, f.eks. bil, enhet, skole. For begrepsdefinisjoner er det nyttig at termen uttrykkes i URIen. Dette gjør det lettere å se, og å slå opp på en term
{referanse}	En referanseverdi som benyttes for å skille mellom individuelle instanser av et begrep. Referanseverdier er typisk avledet fra kode-verdier eller unike navn.
{version}	Et valgfritt felt (bør i de fleste tilfeller utelates) for å skille mellom ulike versjoner av objekter eller begreper. Merk at versjoneringen av «tingen» og dens dokument-represenasjon er uanhengig av hverandre. Referanser uten versjon vil i alle tilfeller referere til den siste versjonen som foreligger når referansen ble dereferert (slått opp).
{autoritet}	Det kan være tilfeller hvor en etat ønsker å vise sitt eget eierskap til begeper og data som ekponeres. I slike tilfeller vil man istedet for {sektor}.data.norge.no, benytte egen autoritet som domenenavn.
{vokabular}	En gruppering av begreper som avgrenser domenet. Et vokabular kan være vertikalt definert, f.eks. enhetsregisteret, eller horisontale vokabularer f.eks. et person vokabular.

Bidragsytere

Jens Kilde Mjelva, Computas; Audun Stople, UiO; Per Myrseth, DNV; David Norheim, Computas; Stig Dørmæenen, Computas; Dimitru Roman, Sintef. Steinar Skagemo, Difi; Rune Smistad, Mediarena

Referanser

[1] Livssyklusprosess for utarbeidelse og forvaltning av begrepssystemer, februar 2011. Vibeke Dalberg, Per Myrseth og Jim Yang.
http://www.semicolon.no/Rapport_prosess_begrepssystem_Final.pdf

[2] Åpne og lenkede data. En informasjonsinfrastruktur for elektronisk samhandling. Rapport nr. 2011-276, revisjon nr. 1. Det Norske Veritas, februar 2011. Robert Engels og Per Myrseth. http://www.semicolon.no/Aapne_lenkede_data_Tech_Report_2011-276.pdf

[3] Chief Technology Officer Council Designing URI Sets for the UK Public Sector 2009 Data.gov.uk Creating URIs

[4] Chief Technoogy Officer Council, Designing URI Sets for Location, May 2011. Data.gov uk. http://location.defra.gov.uk/wp-content/uploads/2011/09/Designing_URI_Sets_for_Location-V1.0.pdf

[5] ISO 11179 Information Technology -- Metadata registries (MDR), Part 6 Registration. http://metadata-stds.org/11179/

[6] Tom Heath og Christian Bizer. Linked Data: Evolving the Web into a Global Data Space, bind 1, kapittel 2 og 4. Morgan & Claypool, 2011.

[7] Leo Sauermann og Richard Cyganiak. Cool URIs for the Semantic Web - W3C Interest Group Note, desember 2008.

[8] Robert Battle og Edward Benson. Bridging the semantic web and web 2.0 with representational state transfer (REST). Web Semant., 6:61–69, februar 2008. ISSN 1570-8268.

[9] Jens Kilde Mjelva. Mobile, håndholdte enheter som plattform for semantisk vev-applikasjoner basert på åpne, offentlige data. Mastergradsoppgave, Universitetet i Oslo, 2011.

[10] Jeni Tennison. What Do URIs Mean Anyway? juli, 2011.

[11] Jeni Tennison. Versioning URIs, juli, 2009.

[12] Tim Berners-Lee. Hypertext Style: Cool URIs don't change. 1998

[13] The Vocabulary of Interlinked Datasets) http://www.w3.org/TR/void/

[14] URI, offisielle beskrivelse av syntaks: http://tools.ietf.org/html/rfc3986

[15] SERES, semantikkregisteret for offentlig samhandling.
http://www.brreg.no/samordning/semantikk/

[16] Dublin Core Metadata Terms, http://dublincore.org/documents/dcmi-terms

[17] Open Provenance Model, http://open-biomed.sourceforge.net/opmv/ns.html

Appendix: Eksempler på implementering av URI-regime for Enhetsregisteret

I det følgende vises hvilke konsekvenser en gjennomføring av det skisserte URI-regimet vil ha for en aktør i det offentlige Norge. Merk at implementasjonen er under arbeid.

I forbindelse med Semicolon-prosjektet lages en URI-struktur for identifikasjon av- og mulighet for oppslag på enheter i Brønnøysundregistrenes Enhetsregister (ER). Denne URI-strukturen er laget på en måte som svarer til anbefalingene skissert i dette dokumentet. 303-URI-er brukes på riktig måte og det finnes dokumentrepresentasjoner både i HTML og RDF. Ved oppslag på en ID-URI redirectes brukeren til en RDF-representasjon (Turtle eller RDF/XML) eller en HTML-representasjon av ressursen på bakgrunn av HTTP-innholdsforhandling.

Enhetsregisteret på begrepsnivå

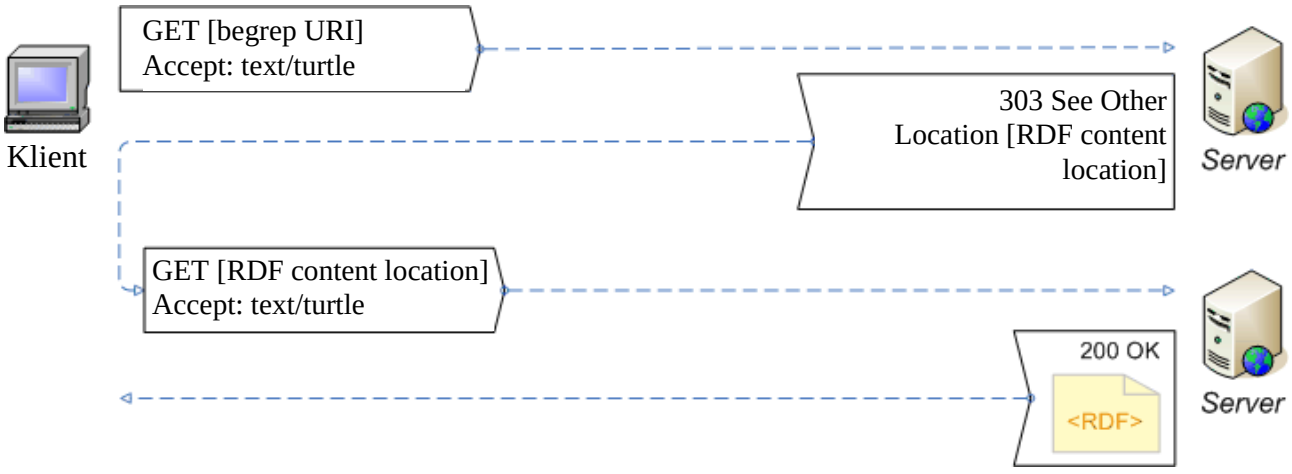
Strukturen for URI-ene på begrepsnivå i Enhetsregisteret under implementering er:

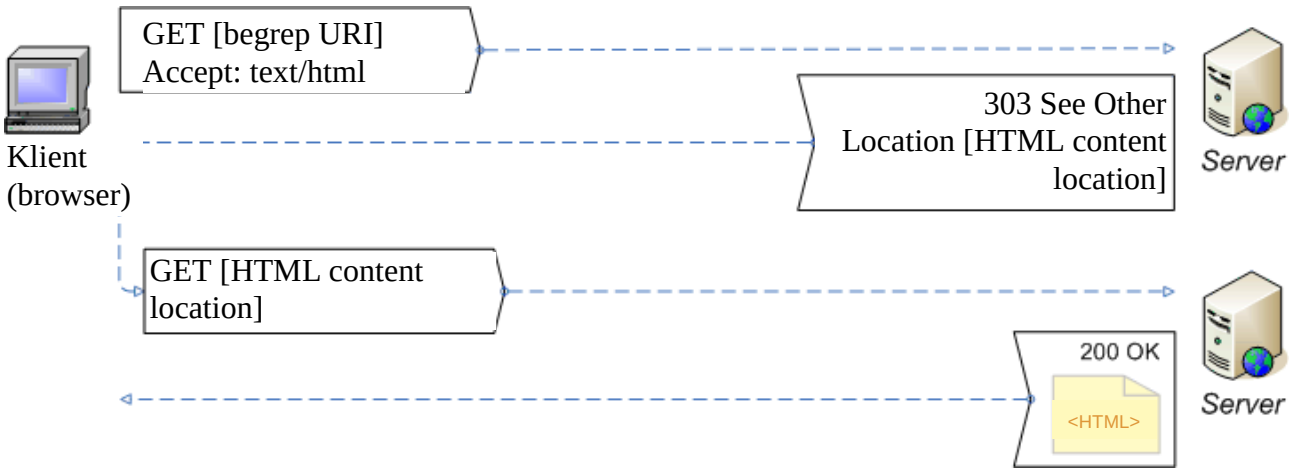
- ID-URI: http://brreg.no/vocab/enhetsregisteret/<begrep>
- Data: http://brreg.no/data/enhetsregisteret/<begrep>
- HTML: http://brreg.no/page/enhetsregisteret/<begrep>

Denne strukturen beskriver hva slags emne begrepene tilhører, nemlig Enhetsregisteret. Emnetilhørighet er inkludert i URI-ene for å tydeliggjøre hva vokabularet omhandler.

Oppslag på et begrep i den nåværende strukturen returnerer kun data som omhandler det begrepet oppslaget gjøres på (/<begrep>). Dermed slipper man å bla gjennom masse data som ikke er relevant for å forstå betydningen av *det* begrepet oppslaget gjøres på. Oppslag som returnerte hele ontologien muliggjøres ved å fjerne <begrep>-delen av URI-en.

Dererferering av begreper





Enhetsregisteret på instansnivå

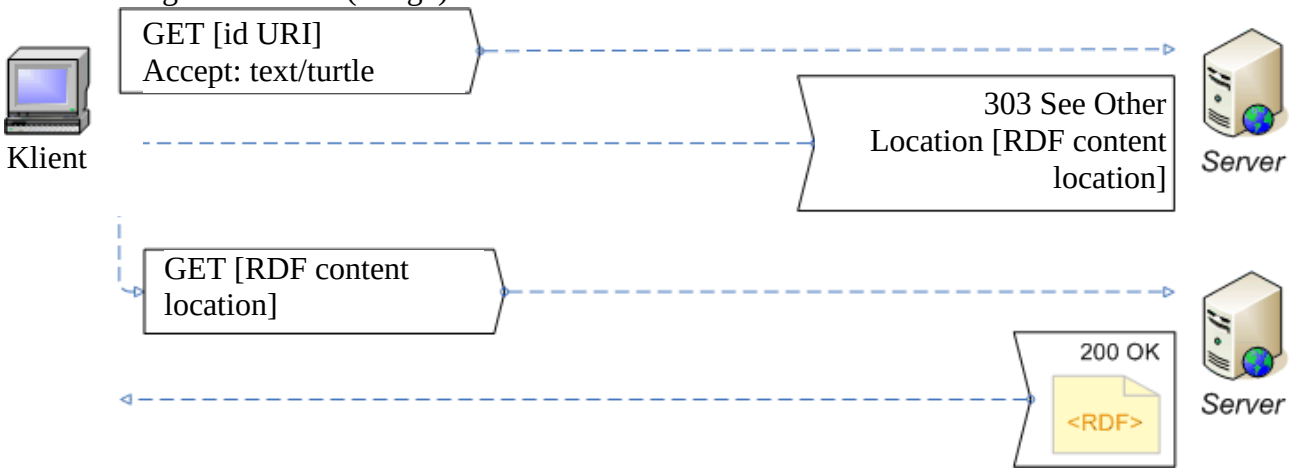
Foreslått skille mellom ID-ressurser og dokumenter (data og HTML) for ER:

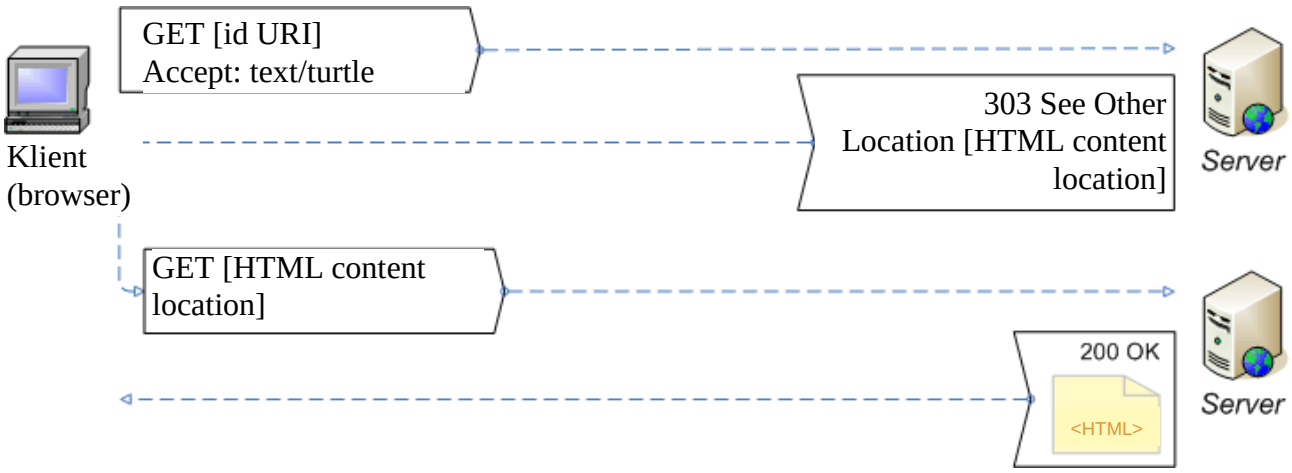
- ID-URI: http://brreg.no/id/enhet/<orgnr>
- Data (RDF): http://brreg.no/data/enhet/<orgnr>
- HTML: http://brreg.no//page/enhet/<orgnr>

En tredeling i URI-strukturen gjennom skille også mellom data og HTML og tillegg til skillet mellom ID-URI og dokument-URI, er et designvalg som muliggjør at man snakke om den menneskelesbare representasjonen og den maskinlesbare representasjonen hver for seg. Dette er naturlig når innholdet i de to dokumentene ikke er helt det samme. Dataene som beskrives i en HTML-dokumentbeskrivelse av en ressurs i Enhetsregisteret er nemlig en delmengde av dataene i en RDF-dokumentbeskrivelse av samme ressurs. Dette for å gjøre HTML-siden mindre omfattende og dermed mer lettleselig.

I den nåværende strukturen finnes altså data-instansenes type, "enhet", angitt i ID-URI-ene. Dermed kan en i henhold til REST og oppslag på samlingsnivå tenke seg å gjøre et oppslag på "enhet" uten angitt organisasjonsnummer for slik å få metadata om enhet som type og kanskje noen eksempelinstanser av denne.

Dererferering av instanser («ting»)




GET [id URI]
Accept: text/turtle
303 See Other
Location [HTML content location]
Server
Klient
(browser)
GET [HTML content location]
200 OK
<HTML>
Server