TECHNICAL REPORT

ÅPNE OG LENKEDE DATA

EN INFORMASJONSINFRASTRUKTUR FOR ELEKTRONISK SAMHANDLING

SEMICOLON-PROSJEKTET

REPORT NO. 2011-276
REVISION NO. 1

DET NORSKE VERITAS

DET NORSKE VERITAS

TECHNICAL REPORT

Date of first issue: 2011-02-28	Project No.: 913A0304
Approved by: Erling Svendby	Organisational unit: DNV Sustainability and Innovation
Client: Semicolon Research Partners	Client ref.: Semicolon, Terje Grimstad

DET NORSKE VERITAS AS

Veritasveien 1,
1322 HØVIK, Norway
Tel: +47 67 57 99 00
Fax: +47 67 57 99 11
http://www.dnv.com
Org. No: NO 945 748 931 MVA

Summary:

The importance of electronic cooperation between the public sector and citizens & business, as well as internal cooperation within the public sector leads to an increasing demand on new digital solutions for information management and publication. Semantic technology is a new field of solutions allowing for new ways of managing, publishing and sharing digital information. Probably the most important improvement contributed by semantic technology is the use of information models with a formal specification of meaning. Such formal semantics allow for retrieval, automated merging, manipulation and visualisation of information from a variety of sources, independent from many implementation aspects, thereby contributing to the vision of genuine "open data".

Semicolon implemented and executed three pilot-studies in order to test semantic technologies and better understand organisational and juridical aspect of open data. Besides these pilot studies, a literature study has been conducted. Pitfalls and possibilities have been discussed with the consortium partners. Besides forming the basis for the proposed tasks for the follow-up project Semicolon II and in order to show the benefits of semantic technology, the results of the Semicolon I are already re-used in cooperation with other research projects.

This work has received extensive national and international attention and has been presented and promoted at numerous events.

Providing a thorough discussion of possibilities and challenges within the field of open data, the report concludes with a concrete and to-the-point list with recommendations.

Report No.: 2011-276	Subject Group:
Report title: Åpne og lenkede data	
Work carried out by: Robert Engels and Per Myrseth	
Work verified by: Vibeke Dalberg	

Date of this revision:	Rev. No.:	Number of pages:
2011-02-28	1	43

Indexing terms

Key words	Service Area
Semantic technology, Linked Open Data, eGovernment services	Systems & Software
	Market Sector
	General IndustryGeneral

☒ Unrestricted distribution (internal and external)

☐ Unrestricted distribution within DNV

☐ Limited distribution within DNV after 3 years

☐ No distribution (confidential)

Table of Content *Page*

1 SAMMENDRAG

Offentlig sektor deltar i økende grad i elektronisk samhandling. Dette er samhandling mellom offentlige enheter samt mellom offentlig sektor og borgere & næringsliv. Semantisk teknologi er en nyvinning inne IT som gjør det mulig å tilby åpne data i nye former og dette gir andre fordeler enn tidligere teknologier. Det viktigste nye momentet i denne teknologien er bruk av formelle modeller av datas betydning. Dette gjør det mulig å: slå sammen data fra ulike kilder, gjenfinne, manipulere og visualisere på nye måter.

Semicolon-prosjektet har gjennomført flere piloter for å teste ut semantiske teknologier og for å forstå mer av organisatoriske og juridiske forhold rundt åpne data. Konkret har prosjektet laget tre tekniske piloter, gjennomført litteraturstudie og diskutert muligheter og resultater med en rekke aktører. Resultatene har ført til samarbeid med andre forskningsprosjekter og inngått som sentralt tema i Semicolon II prosjektet.

Aktiviteten har fått stor oppmerksomhet og resultatene er blitt presentert i en rekke nasjonale og internasjonale fora.

Rapporten diskuterer muligheter og utfordringer ved åpne data primært fra offentlig sektor og avsluttes med en punktliste med anbefalinger.

2 INNLEDNING

Denne rapporten oppsummerer flere leveranser og litteraturstudie knyttet til Semicolonprosjektets arbeid med åpne data og metadata. Semicolon er et brukerstyrt innovasjonsprosjekt, med delfinansiering fra Norges Forskningsråds VERDIKT-program.

Semicolonprosjektets hovedmål er å utvikle og utprøve IKT-baserte metoder, verktøy og metrikker for hurtigere og billigere å oppnå semantisk og organisatorisk interoperabilitet innen offentlig sektor og mellom offentlig sektor og både næringsliv og borgere.

Semicolon plandokumenter for arbeidet med åpne data og metadata er som følger:

- Aktivitetsplan for 2009 – Open data and metadata, Versjon 1.0, 2009-10-05
- Aktivitetsplan for Semicolon Metadata 2010. SSB Case, økt tilgjengeliggjøring av metadata og makrodata.
- Studentoppgave for IFI, sommer 2010. Versjon 1.0, den 9. juni 2010

Bakgrunn for aktiviteten:

(I) Dagens arkitekturer for elektroniske samhandling benytter ofte informasjons-utveksling som stafettpinne mellom prosesser. Realisering av slike løsninger er ofte kostbare og komplekse å etablere og forvalte. Noen få arkitekturer for elektronisk samhandling har vokst frem, eksempler er tradisjonell EDI, Web-Services, batch duplisering av data og innsynsløsninger.

(II) Offentliglova av 1.1.2009, realiserer EU-direktiv om gjenbruk av den offentlige sektors informasjon (direktiv 2003/98/EF). Det fremheves i direktivet og motivasjonen for loven at god tilgang på offentlige data er viktig for å skape verdiøkende tjenester, og dermed fremme informasjonssamfunnet. Loven skal sikre mulighet for viderebruk av informasjon

(III) Dagens publiserte offentlige informasjon er i vesentlig grad låst til informasjonseiers formater og tekniske tjenester.

Basert på W3C[1] sitt initiativ Semantic Web, også kalt Web 3.0, ønsker Semicolon å vise at (i) Semantic Web kan benyttes som en arkitektur for samhandling og (ii) oppfyllelse av offentlighetsloven kan gjøres mulig ved bruk av Semantic Web. Den delen av Semantic Web vi benytter er omtalt som Linked Data og Linked Open Data. Denne arkitekturen for samhandling skal kunne fungere side om side med andre arkitekturer.

2.1 Målgruppe for leveransen

Alle som er underlagt offentlighetslova eller som har informasjon som er av allmenn interesse, eller som har informasjon som kan skape forbedringer og innovasjon hos andre. Dette må sees i sammenheng med målet for Semicolonprosjektet om hurtigere og billigere samhandling.

2.2 Forskningsutfordringer i aktiviteten

I plandokumentene ble det satt opp følgende forskningsutfordringer knyttet til offentlige åpne data og Linked Open Data generelt:

- Hele spekteret av det som omtales som Semantic Web Sciences.
- Egen overbyggende struktur som gir sammenhenger mellom etaters metadata.
- Distribuert bygging og forvaltning av ontologi og ontologi evolusjon.
- Informasjonsforvalting
  - Systematisere og beskrive livssyklus for datasett
  - Data kvalitet
- Rettigheter og prising av informasjon

2.3 Resultatmål for åpne data og metadata aktiviteten i Semicolon

Plandokumentet for LOD aktiviteten listet en rekke primære resultatmål. Punktene under viser hvordan disse punktene er besvart.

Pilotene som er nevnt videre er alle kort beskrevet i kapittelet *Pilotene* i dette dokumentet. ORIS med mer er beskrevet fyldig i [Giese11]. Sommerstudentpiloten er beskrevet i [Bruce10a].

Resultatmål 1:
Infrastruktur for klassifisering av åpne offentlige data. Med klassifisering menes her typebeskrivelser, utvalg, juridiske med mer. Også andre grupperinger som rollebegrensninger for bruk, prising og så videre vil være ønskelig.

[1] www.w3c.org World Wide Web Consortium

Er besvart ved piloten ORIS. Det ble hentet innspill fra sommerstudentpiloten hos SSB, supplert med presentasjoner og feedback i prosjektmøter, seminarer og konferanser.

Resultatmål 2:
Infrastruktur for publisering og gjenbruk av åpne offentlige data

- Anbefale representasjonsformater for publisering i infrastrukturen.

Er besvart ved piloten ORIS og sommerstudentpiloten hos SSB.

- Supplert med presentasjoner i prosjektmøter, seminarer og konferanser.
- Anbefalinger og erfaringer er presentert i et eget kapittel i dette dokumentet.

Resultatmål 3:
Piloter som viser ulike typer anvendelser av infrastrukturen;

- Pilotere publisering av offentlige data i gjenbrukbare formater, og liste offentlige data i en sentral portal.
- Anvendelse av portalen som viser gevinsten for separate formål.
- Portal skal gi mulighet for innsyn, navigasjon, søking i data

Er besvart ved pilotene, ORIS, Brønnøysund-piloten og Sommerstudentpiloten hos SSB.

Resultatmål 4:
En serie forskningsartikler som beskriver hva vi har gjort og hvilke effektmål vi kan oppnå.

Er besvart ved rapportene listet under samt forskningsartikler:

- IFI sin sluttrapport til Semicolon [Giese11]
- Denne rapporten, se eget appendiks i dette dokumentet som lister presentasjoner, artikler og medieomtale.
- Sommerstudentenes sluttrapport [Bruce10a]

Resultatmål 5:
Presentasjonsmateriell for prosjektet og til annen formidling

Er besvart ved presentasjoner en rekke steder. Se eget appendiks i dette dokumentet som lister presentasjoner, artikler og medieomtale.

Resultatmål 6
En PhD grad (ferdig etter Semicolon prosjektet) og flere mastergrader innen utgangen av 2010.

Er besvart ved IFI sin innsats og studentfremdrift, se [Giese11] og kapittelet arbeidsform i dette dokumentet.

Erfaringer fra arbeidet med disse resultatmålene er diskutert i kapittel *Erfaringer og diskujon* og blir oppsummert med en rekke anbefalinger i kapittelet *Anbefalinger og konklusjon.*

2.4 Leveranser fra Semicolon åpne data og metadata aktiviteten

Følgende piloter er utviklet:

- ORIS, portal som lister ulike datakilder, formatkonverterer, sammenstiller og visualiserer spørreresultater.
- Sommerstudentpiloten, portal med makrodata fra SSB som Linked Open Data. Bygging av ontologier, formatkonverteringer, sammenstiller og visualiserer spørreresultater.
- Foretaksregister som Linked Open Data. Innpakking av web-services som linked open data kilde. Denne tjenesten er brukt i følgende andre piloter:
    - Sesam4 prosjektet og Semicolon har sammen gjort en kobling av NFR prosjektdata og Brønnøysundregisterenes Foretaksdata.
    - Sesam4 prosjektet har koblet et stort nyhetsarkiv (ca 40 millioner nyhetsartikler fra CyberWatcher) og oppdatert foretaksinformasjon fra Foretaksregisteret ved hjelp av semantisk teknologi.

Andre leveranser er

1. Eget kurs i semantisk teknologier på IFI, INF3580 Semantiske Teknologier
2. Mastergradsoppgaver er ferdig eller blir ferdig i 2011 [Giese11]. Dekker følgende tema:
    1. Evaluering av spørringer over enveis datakilder.
    2. Generering av XML fra semantiske datalagre
    3. Åpne offentlige data for mobile håndholdte enheter
    4. Transformering av tabulære data til RDF
    5. Visualisering av offentlige data i Oris/Synapse

Beskrivelse av leveransene gjøres tredelt:

1. Dette dokumentet
2. Oppsummeringsrapport fra IFI [Giese11]
3. Oppsummeringsrapport fra Sommerstudentarbeidet [Bruce10a]

2.5 Sentrale termer brukt i dette dokumentet

I tabellen under lister vi en del sentrale begreper som er fremkommet gjennom arbeidet i Semicolonprosjektet og anvendt i denne rapporten.

Beskrevne data	Vil si at data er forklart/beskrevet på en slik måte at andre enn eier kan forstå betydningen av data og begrensninger i data. Dette gjelder både selve innholdet og forvaltningsbeskrivelse av data, (de kvalitetsregimer data er underlagt). Beskrivelsen bør være gjort i henhold til en metode for ontologi bygging og forvaltning som andre enn etaten selv har kompetanse på og beskrivelsene må betraktes som maskinlesbare data.
Data	Er tall, tekst eller binære objekter som bilder, lyd eller film egnet for lagring, utveksling, viderebruk og bearbeiding (inspirert av OAIS Referanse Modell definisjonen).
Maskinlesbare data	Data som er representert på et åpent og veldefinert format slik at et mangfold av IT-systemer kan lese, manipulere og gjennomføre tapsfrie transformasjoner mellom ulike datarepresentasjoner. Dette uten at en benytter manuell datafangst / re-punching eller annen tapsbasert datafangst som skanning med OCR el.
Offentlige data / Public Sector Information (PSI)	Vil si at offentlig sektor alene eller sammen med andre har finansiert og eller opparbeidet data, er ansvarlig for å forvalte dem, og at offentlig sektor har rettigheter knyttet til egen bruk og viderebruk.
Sterke identifikatorer	Er et datum som entydig skiller et datasett / tuppel fra et annet og som er egnet til å koble datasett sammen. Eksempler er personnummer, biometri eller organisasjonsnummer. Termen primærnøkkel fra relasjonsdatabaseteori er ofte brukt som alias til sterke identifikator. En URN kan være en sterk identifikator.
Åpne data	Vil si data som en borger, etat eller næringsdrivende kan få tilgang til ved innsyn og evt. en kopi. Formålet med bruken skal ikke være avgjørende for innsyn/tilgang.

Overstående termer inngår i termen under. Termene under er listet alfabetisk.

Elektronisk samhandling	Samarbeid mellom to eller flere aktører gjennom (automatisk) digital interaksjon mellom datasystemer med den hensikt å oppnå et felles forretningsmessig mål. Samhandling forutsetter at det er en felles forståelse for innholdet i data som blir utvekslet.
Lenkede data / Linked Data	At innhold i et sett med data kan knyttes til andre sett med data, og at det gjøres med sterke identifikatorer for å sikre felles betydning. Er summen av følgende egenskaper: • Beskrevne data • Maskinlesbare data
Lenkede åpne data / Linked Open Data (LOD)	Lenkede åpne data er summen av følgende egenskaper: • Beskrevne data • Åpne data • Maskinlesbare data

	• Lenkede data
Lenkede åpne offentlige data	Er summen av følgende viktige egenskaper: • Offentlige data • Beskrevne data • Åpne data • Maskinlesbare data • Lenkede data
Mikroontologi	Er en ontologi, som er avgrenset til et fåtall typer konsepter og tilhørende relasjoner, og kan gjerne være en kodeliste eller en systematisk sammenheng. Eksempel kan være (i) en liste over verdens land, (ii) sammenheng mellom år, mnd, dager, uker, (iii) typer sivilstatus, eller (iv) sammenheng mellom postnummer, kommuner, fylker og politidistrikt.
Ontologi	Ontologi er i datateknologien og informasjonsvitenskap en formell representasjon av et sett av begreper innenfor et domene, samt hvordan disse begrepene er relatert til hverandre. En ontologi brukes til å markere dataene sin betydning og lette smart dataprossesering av dataprogrammer i domenet. (Wikipedia).
Rådata	Er *data* i sin mest detaljerte form, dvs ikke bearbeidet, summert eller aggregert. De kan være anonymisert, kvalitetstestet, feilkorrigert etc, men datasettet må da beskrives som sådan. For de fleste bruksformål antar vi at rådata som er kvalitetstestet og feilkorrigert er å anse som mest interessant for viderebruk fordi disse brukes av virksomheter eller etatene i sin virksomhetsutøvelse.
Semantisk teknologi	Er datateknologi som er i stand å sammenstille og utføre regelbaserte operasjoner på rådata ved bruk av strukturerte modeller (ontologi) som beskriver meningen med data. En eksplisitt, entydig og publisert formell semantikk definerer meningsinnholdet. Programvare som kan utføre operasjoner på slike semantiske modeller og tilhørende data, kalles gjerne med en fellesbetegnelse semantisk teknologi.
Åpne offentlige data / Open Government Data	Åpne offentlige data er summen av følgende egenskaper: • Offentlige data • Beskrevne data • Åpne data • Maskinlesbare data

I resten av dette dokumentet bruker vi definisjonene i tabellen over. Definisjonene over er inspirert av definisjonene benyttet i bl.a. [FAD10a] og kilder som Wikipedia (merket).

3 ÅPNE DATA, HISTORIEN FREM TIL I DAG

3.1 Historien

Under har vi beskrevet noen sentrale hendelser som har ledet frem til dagens Linked Open Data trend.

"The key thing about all the world's big problems is that they have to be dealt with collectively. If we don't get collectively smarter, we're doomed." Doug Engelbart - Intelligence in the Internet Age, New York Times, 9/19/05

Tanken om at informasjon må knyttes sammen for å få en betydning og en relevans kan føres så langt tilbake som til klassifikasjonstankene til Aristoteles. Filosofer har bearbeidet temaet ytterlige, mens psykologer har forsøkt å koble det mot modeller av resonering og forståelse. Også bibliotekarer og arkivarer har bidratt med gode tanker og initiativer i mange år. I en god del av disse diskusjoner har det dreid seg om klassifikasjonsskjemaer, emner og kontrollerte vokabularer. Når vi beskriver historien om lenkede data retter vi oss hovedsaklig mot den nyere historie, hvor disse mer filosofiske tanker treffer den mekaniske /digitale verden. Historien til åpne lenkede data (Linked Open Data eller LOD) inneholder bidrag fra flere fagmiljøer og ender opp med et rammeverk (Semantisk Web av W3C) som definerer de byggeklosser og verktøy man trenger for å få lenkede åpne data implementert og brukt i praksis. Vi starter vår historie med Vannevar Bush og Doug Engelbart.

Memex (1945) – Vannevar Bush

Det var i 1945 at Vannevar Bush begynte å fundere på måter å lagre informasjon om alt han hadde lært i sitt liv i en slags maskin. Denne maskinen måtte så senere kunne spørres for å få ut informasjonen igjen. Denne virtuelle maskinen kalte han Memex (fra «MEMory Extension»). Memex-en er et apparat som komprimerer og indekserer alle ens bøker, notater, poster og kommunikasjon som er mekanisert slikt at den kan bli konsultert med økende effektivitet og fleksibilitet [Wikipedia]. Hovedtanken bak Vannevar Bush' teoretiske datamaskin er at man kan lagre alt av informasjon man kommer over mens man legger til assosiative lenker som senere kan følges for å lenke sammen informasjon man har bearbeidet tidligere. Bush' tanker og beskrivelser følger stort sett tilgjengelig teknologi etter andre verdenskrig, hvor mikrofilm og kameraer spiller en hovedrolle. Grunnen til at Bush ofte blir referert til i sammenheng med dagens Internett er at hans tanker er veldig likt grunnlaget til Internett. Han beskriver informasjonsenheter som er linket sammen med assoiative linker som selv har typer/roller. En av hovedtankene bak Semantisk Web er nettopp denne type linking som i tillegg tillater individualiserte linker som kan deles med andre brukere. Denne tanken som ligger bak Memex systemet fra 1945 er nå i ferd å bli virkelighet i initiativer som Linked Open Data (LOD).

Augmented human intellect – Doug Engelbart

Doug Engelbart er kanskje mest kjent for sitt patent på "musen", pekeinstrumentet han fikk patent på i 1970, og som senere er blitt brukt av millioner brukere av datamaskiner verden over. Han er minst like kjent for sin forskning på Stanford Universitet. Han hadde et sterkt ønske å gjøre verden bedre og mente at nøkkelen til forbedring er å lage et kollektivt menneskelig intellekt av alle mennesker som bidrar med effektive løsninger. Har man fått det til, så burde det være innen rekkevidde å løse store og vanskelige oppgaver i verden ved å ta i bruk mange, desentraliserte datamaskiner i en "Augmented Knowledge Workshop" [ENGEL73]. Mange viktige momenter fra arbeidet til Engelbart kommer tilbake i dagens "semantisk web" filosofi og implementasjon. Engelbart's forskning førte til en tidlig versjon av Internett (kjent som ARPANET), som var ett nettverk med fire noder som ble ferdigstilt i 1969 og HyperText, som representerte en måte å lenke informasjon som var tilgjengelig fra forskjellige noder i ett nettverk av datamaskiner. Den tekniske implementasjonen av disse ideer var en viktig forutsetning for en senere implementering av Engelbarts ideer om "augmented" knowledge i datanettverk.

Illustrasjon 1: Doug Engelbart's Workstation for testing screen selection and knowledge augmenting (1965)

Knowledge representation on the Internet - SHOE

Som en av de første, koordinerte tilnærminger kom Simple HTML Ontology Extensions (SHOE) både med en filosofi og et rammeverk for kunnskapsrepresentasjon på digitale, distribuerte systemer. SHOE tok utgangspunkt i utvidelser av den eksisterende HTML standarden og la til en mulighet for å inkludere maskin-lesbare, semantiske annotasjoner i web-baserte dokumenter [SHOE00].

SHOE standarden tillater at man deler opp eksisterende web-dokumenter i deler. Disse deler kan så annoteres hver for seg, relasjoner mellom informasjonselementer og et felles begrepssystem (tilsvarende en ontologi) kan defineres. Ontologiene kan inneholde klassifikasjoner, relasjoner og avledet kunnskap ("inferred knowledge").

Fra Internett til Semantisk Web – Tim Berners Lee

The first step is putting data on the Web in a form that machines can naturally understand, or converting it to that form. This creates what I call a Semantic Web – a web of data that can be processed directly or indirectly by machines. Tim Berners Lee – Weaving the Semantic web [TBL+99, page 191]

I 1999 la Tim Berners-Lee grunnlaget for semantisk web i sitt paper "Weaving the web" [TBL+99]. Utviklingen av semantisk web gikk over flere år og resulterte i et sett med teknologier som støtter hverandre. Sammensatt danner de på mange måter en implementasjon av tankene fremlagt av Bush og Engelbart 40 år tidligere.

Teknologien bak semantisk web-filosofien gjør at individer og maskiner kan annotere ressurser (som i *digitale representasjoner av eksisterende, fysiske objekter*), men også tanker, ideer og konsepter. Alle slike annotasjoner for en identifikator gjør dem refererbar. Indikatorene og annotasjonene legges ut på internett for gjenbruk av andre. Alle har lov og mulighet til å uttrykke seg om ressurser og sammenhenger mellom ressurser. Enighet om uttalelser kreves ikke, rammeverket tillater at det eksisterer motsigelser. Bakgrunnen i F-Logic [KLW95], sett-teori og beskrivelses-logikk (description logics) gjør at datamaskiner får muligheten til resonering over informasjonen som legges ut.

Porteføljen av tekniske spesifikasjoner som utgjør Semantisk web-teknologien forvaltes av World Wide Web Consortium (W3C) gjennom en global dugnad i åpne diskusjoner mellom forskjellige fagområder, akademia og industri. Mange implementasjoner baserer seg på åpen og fritt tilgjengelig programvare, noe som har hjulpet på utbredelse.

Standarder som RDF/S og OWL baserer seg på filosofi og logiske rammeverk som semantiske nettverk, Frame-Logic og Description Logics, og har dermed ikke tatt hensyn til temporale aspekter (som «temporal logics»). OWL kommer i tre varianter. *OWL Lite* som inneholder lite semantikk og er gjerne brukt i applikasjoner med en liten datamodell som inneholder mange instanser. I andre enden er det *OWL Full*, som tillater at man definerer komplekse datamodeller for komplekse domener, men resonering i OWL Full-domenet er ikke-deterministisk. For mange applikasjoner rekker det med «mellom-varianten» *OWL DL,* som også er det mest brukte alternativet, hvor modellene er litt forenklet, men som i prinsipp tillater resonering over modellen og instanser.

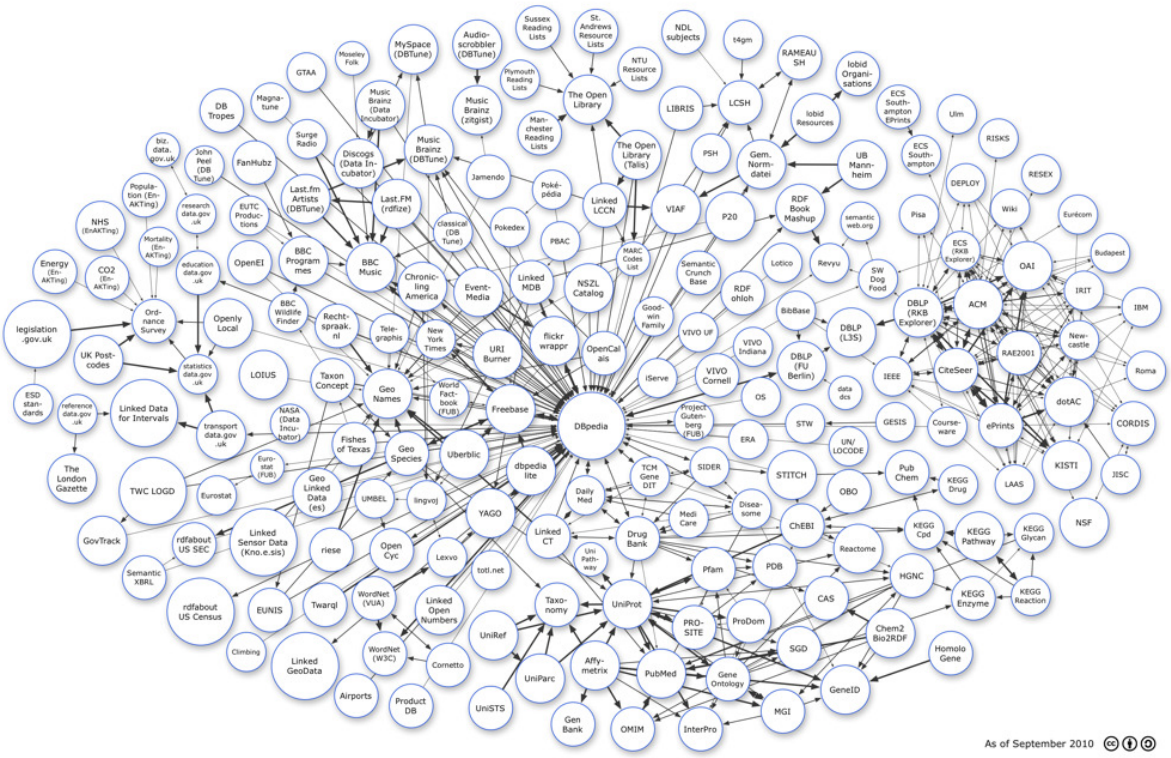
Den nyeste OWL standarden (OWL2) inneholder noen flere matematiske operatorer som «linear equations», i tillegg til en forbedret logikk basert på OWL. OWL2 har også tre sub-språk (EL, QL og RL), som har sine egne, definerte bruksområder. Det er også initiativer som OWL Time, men status er at ontologi-evolusjon er under utvikling. Ved å ta i bruk for komplekse teknologier som krever mye kompetanse vil vi også heve terskelen for å legge ut data og begrense viderebruk. Dette er en problemstilling som vil bli vesentlig å finne en god balanse på i de kommende år, hvor ulike domener trolig vil lande på ulike ambisjonsnivåer.

3.2 Dagens situasjon

Semantisk Web i praksis: Linked Open Data (LOD)
Selv om teorien i semantisk web begynner å bli moden, har praktisk utbredelse tatt lengre tid enn det optimistene hadde forespeilet seg. Et initiativ som viser en mulig vei til implementasjon av en visjon som innebærer "semantikk på web" er nettopp Linked Open Data. Alle byggesteiner for linking av eksisterende datasett er levert i de spesifikasjoner som utgjør "semantisk web" tanken. Man finner byggeklosser for å beskrive data på det laveste nivå (instanser med egenskaper), man finner rammeverk for å definere, publisere og dele ontologier, og man har

tilgang til et spørrespråk som bruker ontologien (i form av en RDF/OWL "grafstruktur") for å spørre på forskjellige kilder.

Kort forklart er SPARQL laget for å utføre forskjellige oppgaver uavhengig av den aktuelle implementasjonen av en SPARQL-endpoint. Både protokollen og returnerte svar er standardisert. SPARQL definisjonen kommer med fem "hoved"-konstrukter: *SELECT, ASK, DESCRIBE, CONSTRUCT* og *UPDATE.* Avhengig av hvilken oppgave som er tenkt utført bruker man *SELECT* for å definere en utvalg av en (distribuert) graf-strukur, *ASK* for å få sjekket sannhet i et uttrykk, *DESCRIBE* for å få informasjon om en (ukjent) del av grafen rund en kjent "punkt" i grafen, *CONSTRUCT* for å lage en ny graf basert på en template og *UPDATE* for å endre eller tilføye data/uttrykk i en eksisterende graf.
SPARQL protokollen kan også brukes for å slå sammen deler av forskjellige grafer i én og samme spørring. Mekanismene bak SPARQL er såpass kraftig at det har blitt grunnlag for Linked (Open) Data.



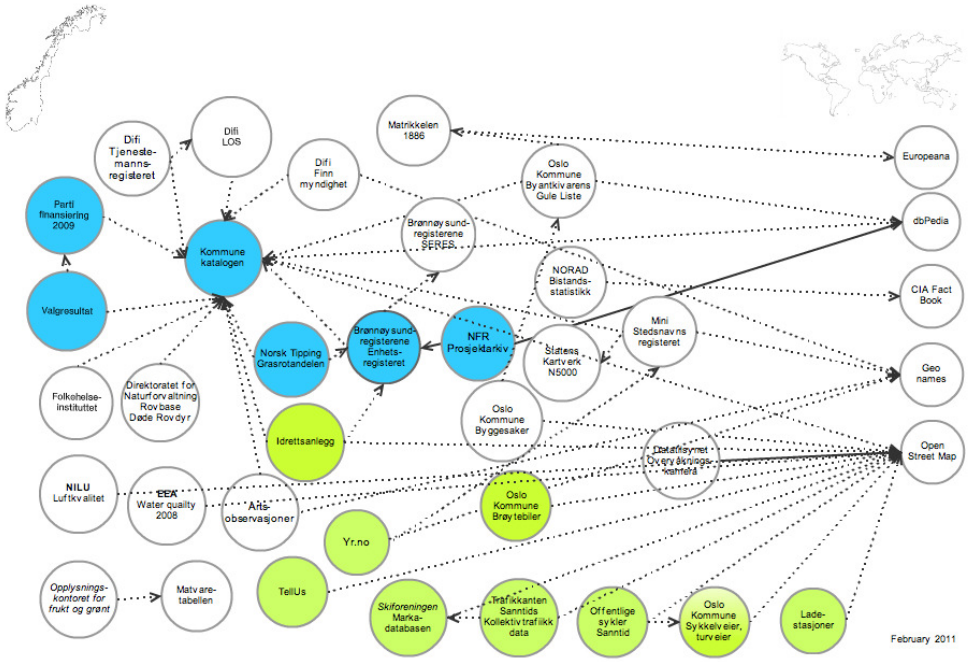
Illustrasjon 2: Linked Open Data Cloud[2]

Noe som er interessant med Linked Open Data som initiativ, er at det illustrerer godt hvordan man vha semantisk web-teknologi, publiserer ontologi for gjenbruk, der vidt forskjellige data-eiere kan semantisere (aka "løfte") sine datakilder. Gjennom ontologiene får man tilgang til

[2] Figuren i full størrelse se: http://richard.cyganiak.de/2007/10/lod/

tilhørende instanser (entitieter) og ontologiene tillater andre å lenke tredje-parts data til ontologien.

Flere og flere datakilder blir nå konvertert og lagt åpne på nett. En av de første artiklene om Linked (Open) Data [TBL06] inneholder en konkret "beste praksis" på hvordan man går frem for å legge ut sin egen informasjon som "linked (open) data".



Illustrasjon 3, Norske åpne data per februar 2011, laget av David Norheim, Computas.

Også den norske delen av skyen vokser, og illustrasjon 3 viser lenker til en del internasjonale data. Gruppering av datasett skaper ofte debatt, men det er gjort et forsøk i illustrasjonen over hvor (i) blå viser eksempel for regional/kommunal utvikling, (ii) grønn er eksempel som viser transport og miljø. Heltrukne linjer viser referanser til andre datasett, mens stiplede viser en mulig kobling.

En trend er at flere og flere organisasjoner beskriver sine datakilder ved hjelp av RDF/RDFS/OWL og tilgjengeliggjør dem for nye typer aksess og spørrespråk. Gjennom kryssreferanser mellom slike åpne, standardiserte informasjonskilder kobles innhold sammen til et globalt tilgjengelig kunnskapsnettverk. Linked Open Data er initiativet for å få så mange forskjellige datakilder som mulig koplet sammen. Illustrasjon 2 viser status på LOD skyen i 2010.

Ser man litt nærmere på illustrasjon 2, finner man relativt mange velkjente informasjonskilder. Det er ingen som har lagt føringer på bruken av denne skyen. Størrelsen på kildene angir hvor

mye data som ligger i dem, tykkelse og retning på lenkene angir hvem som referer til informasjon i hvilke andre baser, og hvor mange relasjoner som er blitt definert.

Sentralt i illustrasjon 2 finner man den semantiske versjonen av Wikipedia, kalt DBPedia. Her ligger alt av informasjon som Wikipedia også publiserer, men tilgjengeliggjort via et semantisk format gjennom et standardisert webgrensesnitt. Ellers finner man en del kilder til musikk og film (bla. BBC data), offentlige kilder (US sensus data mm), bio-medisinske data, og mange biblioteks- og forlagskilder.

3.3 Noen typer åpne data og trender relevant for disse

Selv om fremtiden er ukjent, finner man noen trender som gir oss en pekepinne for hvilken retning ting utvikler seg i.

Nasjonale masterdata
Norge har en forvaltingstradisjon for å etablere registre for nasjonale masterdata. Eksempler er Foretaksregisteret, Det Sentrale Personregistert, Helseforetakregistert, Arbeidsgiver – Arbeidstakerregisteret, Regnskapsregisteret, Biobankregisteret, Partiregisteret, DNA-registeret, Ektepaktregistert osv[3]. Et raskt overslag tilsier at vi har langt over hundre ulike nasjonale registre i det offentliges regi. Dette er datakilder som utgjør en sentral basis i informasjons-samfunnet og som bør inngå i en offentlig strategi for åpne data.

Mikroontologier
Mikroontologier brukes som (a) basis for regelanvendelse og regelforvaltning, (b) ved utveksling av data og kobling av datasett og ved (c) rapportering og statistikkproduksjon. Ved økt samhandling og utveksling av data blir enighet og felles forvaltning av disse mikroontologiene viktigere.

Offisiell statistikk
I Norge har vi store mengder makrodata og statistikk som er samlet gjennom flere hundre år, men første statistiske årbok for Norge som ligger åpen hos SSB er fra 1879[4]. Siden den gang har mengden data om landet vokst voldsomt. SSB har store mengder data som er egnet som åpne data og har allerede lagt ut store mengder. I tillegg har de lagt ut informasjon om hva variabeldefinisjoner betyr i sin metadata-portal, samt at informasjon om datasett som er lagt ut under tittelen ”om statistikken”.

Felles referanser til definisjoner og felles definisjoner
Mange datakilder er nå åpnet og lagt på nett. En viktig tanke er at ontologiene for disse datakildene må referere til hverandre, det vil si bruke hverandres definisjoner av allerede modellerte deler av verden. Jo mer slike ontologier blir gjenbrukt, jo enklere blir sammenslåing av data fra forskjellige kilder etterpå, og jo mer kan man finne ut om hva slags informasjon som er tilgjengelig gjennom «surfing» av linked dataskyer. Der det finnes gode, strukturerte og ofte brukte vokabularer kan man «semantisere» disse og legge dem ut på nett for felles gjenbruk. Eksempler på slike vokabularer er Dublin Core (purl.org), vCard (ofte brukt for

3 http://www.offentlighet.no/Registeroffentlighet/Alle_registre/
4 http://www.ssb.no/histstat/

informasjonsutveksling mellom Personal Information Management systemer), men også ontologier som Friend-of-a-Friend (FOAF). I norsk offentlig sektor benyttes allerede Dublin Core i en rekke anvendelser både innen arkivområdet og som del av ulike fagsystemer.

Økende bruk av lenkede data prinsipper i interne systemer
Flere bedrifter har begynt å ta i bruk prinsippene til lenkede data internt i sine informasjons-systemer. Selv om slike data er konfidensielle og ikke skal publiseres åpent på nett, er det sett på som en fordel å publisere også disse ontologiene. Bruk av andres ontologier tillater nemlig at man kan hente inn åpne data (f.eks kan man knytte tilleggsinformasjon fra steder som wikipedia eller World Fact Book direkte til egne objekter) for så å berike «intern» informasjon.

Nyere datakilder mer åpne enn datakilder som har vært forvaltet over lang tid
Noe å merke seg er at store og viktige initiativer som Europeana [EUR11] (som knytter sammen museumsdatabaser) ikke har kommet så langt at de publiserer innholdet med samme kunnskapsrepresentasjon, selv om det er tiltenkt. En interessant observasjon er også at de datakilder som ble lagt ut tidlig var relativt "nye" datakilder. Det viser seg at konvertering av eksisterende datakilder, dog teknisk mulig, er vesentlig mer ressurskrevende. Som viktigste årsak blir det ofte nevnt behovet for kvalitetssikring, rettighetsavklaringer og redefinering av målsetning med en kilde («er Linked Open Data virkelig hva vi ønsker oss?»). Norske aktører fra kulturarvsektoren leverer innhold til Europeana, og en oversikt og fremtidsperspektiv er tilgjengelig i [ABM10].

Strukturerte data, lagring og utveksling
På midten av 70-tallet ble relasjonsdatabase-teorien etablert og dannet grunnlaget for flere ti-års spennende videreutvikling av databaseteori og teknologi. Nettverksdatabasene fra 70-tallet fikk en del utfordringer med hardware-skalering. Objektorientert databaseteknologi tok opp deler av tankegodset fra nettverksdatabasene, men fikk kun moderat utbredelse. Fremdeles er det relasjonsdatabasene som har størst utbredelse for å lagre og bearbeide strukturerte data.

Fagmiljøer innen databaseteori har utviklet modeller på flere ulike abstraksjons-nivåer for å beskrive innholdet i en database. Først siste tiåret har det blitt vanlig å fokusere på betydningen av data som egne modeller. Etater som driver elektronisk samhandling trenger tre typer modeller i sin informasjonsforvaltning. Disse er: (i) modell av betydning, (ii) modell for lagring og (iii) modell for utveksling. Disse tre modellene er gjensidig avhengig av hverandre og nødvendige for at en moderne og informasjonsintens virksomhet skal kunne delta i elektronisk samhandling.

Teknologi og modeller anvendt ved elektronisk samhandling i form av elektronisk datautveksling (EDI) har på tilsvarende måte som databaseteknologien gått i retning av å skille modellen fra hva informasjon betyr, og fra hvordan informasjon utveksles.

Linked Open Data initiativet anbefaler å beskrive betydningen av data i henhold til RDFS eller OWL. Disse modellene kan anvendes på mange bruksområder for informasjon som lagring, utveksling, filtrering, søking og sammenstilling av data.

3.4 Myndighetstiltak for utbredelse av åpne data

Fornyings-, administrasjons- og kirkedepartementet (FAD) har personell og samarbeid med bl.a. DIFI for å sikre fremdrift på utbredelse av åpne offentlige data samt bidra til nytteeffekter av disse data. Et viktig grep er gjort ved å benytte tildelingsbrevene til etater for å få dem til å bidra til mer og bedre åpne data.

Pressemelding, 17.11.2010 [FAD10b]

Rådata skal kunne brukes av flere
Flere offentlige rådata skal nå bli tilgjengelige for de som måtte ønske å bygge videre på dem. I dag finnes det både nett-tjenester og mobile applikasjoner som viser alt fra værmeldinger og flytider til bussruter og hvor nærmeste idrettsanlegg er. Disse tjenestene er ofte utviklet av private virksomheter selv om de er basert på informasjon som forvaltes av det offentlige. Fra 1. januar 2011 skal alle egnede rådata gjøres tilgjengelig i maskinlesbare formater.

- *Dette gjelder informasjon som har samfunnsmessig verdi, som kan viderebrukes, som ikke er taushetsbelagt og der kostnadene ved tilgjengeliggjøring antas å være beskjedne, sier fornyingsminister Rigmor Aasrud.*

- *Kravet skal inngå i alle tildelingsbrev som departementene skal sende til sine direktorater og etater for 2011 for å beskrive hva de skal gjøre kommende år.*

Grepet med å stille krav til åpning av data som del av tildelingsbrevet er et av de sterkeste virkemidlene offentlig sektor kan benytte seg av, så dette er et viktig og gledelig steg fremover.

Det pågår for tiden en utredning i regi av FAD om samfunnsøkonomiske betraktninger av offentlig sektors Linked Open Data initiativ. Utlysningen fra FAD hadde tittelen: "Utredning av markedspotensialet knyttet til viderebruk av offentlige data". Rapporten var ikke publisert da denne rapporten ble ferdigstilt.

4 VISJONEN OM ÅPNE DATA

Ofte betraktes potensialet i Open Government Data» (OGD) fra sluttbrukeren sin side. Sluttbrukeren er da et individ i samfunnet med spesifiserte ønsker og gjøremål. I denne konteksten blir potensialet gjerne målt i fire dimensjoner [Kaplan10]: Transparens & ansvar, nye & forbedrede tjenester, ny kunnskap & innsikt, samt økt deltagelse av borgere i den demokratiske prosessen.

> *«The idea that we have to choose between government providing services to citizens and leaving everything to the private sector is a false dichotomy. Tim Berners-Lee didn't develop hundreds of millions of websites; Google didn't develop thousands of Google Maps mashups; Apple developed only a few of the tens of thousands of applications for the iPhone.*
>
> *Being a platform provider means government stripped down to the essentials. A platform provider builds essential infrastructure, creates core applications that demonstrate the*

power of the platform and inspire outside developers to push the platform even further, and enforces "rules of the road" that ensure that applications work well together.
---- Tim O'Reilly. Government as a Platform»

Semicolon sitt mål er å utvikle og utprøve IKT-baserte metoder, verktøy og metrikker for hurtigere og billigere å oppnå semantisk og organisatorisk interoperabilitet innen og med offentlig sektor.

I lys av dette målet passer åpne data som tematikk inn på flere måter. Vi skiller mellom ulike typer åpne, lenkede og offentlige data. (Se terminologi kapittelet for ytterligere forklaringer, vi velger å avgrense oss til kun maskinlesbare data). Disse er alle kombinasjoner av følgende punktliste:

- Åpne Data vs ikke Åpne Data
- Offentlige Data vs ikke Offentlige Data
- Lenkede Data vs ikke Lenkede Data

Semicolon-prosjektet har ikke bare fokusert på sluttbrukere som individ, men også tatt utgangspunkt i statens egne organer som «sluttbrukere» og sett på potensial og mulige gevinster for disse. Med organisasjoner som SSB, Brønnøysundregistrene, Skattedirektoratet, KITH, Kommunenes Sentralforbund og Helsedirektoratet i konsortiet er det like naturlig å vurdere hva trenden for åpne data kan gi etatene, som hva trenden kan gi næringsliv og publikum.

De spesifikasjoner som inngår i realisering av åpne data som infrastruktur, danner en ny infrastruktur for elektronisk samhandling. Denne infrastrukturen kan bringe oss fra forvaltning av etatsvise og enkeltstående it-systemer til etablering av en kontrollert og velfungerende infrastruktur for informasjonsformidling og gjenbruk. Tradisjonelle informasjons-forvaltningsbehov som konfidensialitet, tilgangskontroll, versjonering og integritet kan benyttes etter behov.

Visjonen for det offentlige er at «åpne data» kan føre til effektivitetsgevinster, bedre og enklere kommunikasjonsprosesser mellom etater, og en kvalitetsgevinst. Det er interessant at det viser seg å ligge et demokratisk forbedringspotensial i å åpne opp data i et demokrati. Som Tim O'Reilly skriver er «crowdsourcing[5]» en måte å få tilbakemeldinger fra «massene» på. I noen sammenhenger vil denne tilnærmingen kunne supplere eller erstatte tradisjonelle ekspertutvalg eller masseundersøkelser.. En slik effekt kan også ses i Semicolon sammenheng, hvor det å gjøre offentlige prosesser, resultater og data åpne og synlige innenfor og mellom etater allerede har som effekt at man bygger ned siloene og gjør at man blir bevisst behovet for flere dimensjoner av datakvalitet.

En forutsetning for en velfungerende rettsstat er åpenhet rundt hva offentlig sektor gjør og gjennom det hvilke data de opparbeider og benytter i sitt tjenestetilbud, styring og samfunnsregulering. Innsyn i arbeidsprosesser, regler og data er fundamentale for at presse, borgere og næringsliv har mulighet til å etterprøve offentlig sektors gjøren og laden.

[5] **Crowdsourcing** is the act of outsourcing tasks, traditionally performed by an employee or contractor, to an undefined, large group of people or community (a crowd), through an open call. Wikipedia

5 PILOTENE

Semicolon-prosjektet har gjennomført flere piloter for å teste ut teknologi og for å forstå mer av organisatoriske og juridiske forhold rundt åpne data. Kapitlene under gir en kort beskrivelse av de enkelte pilotene.

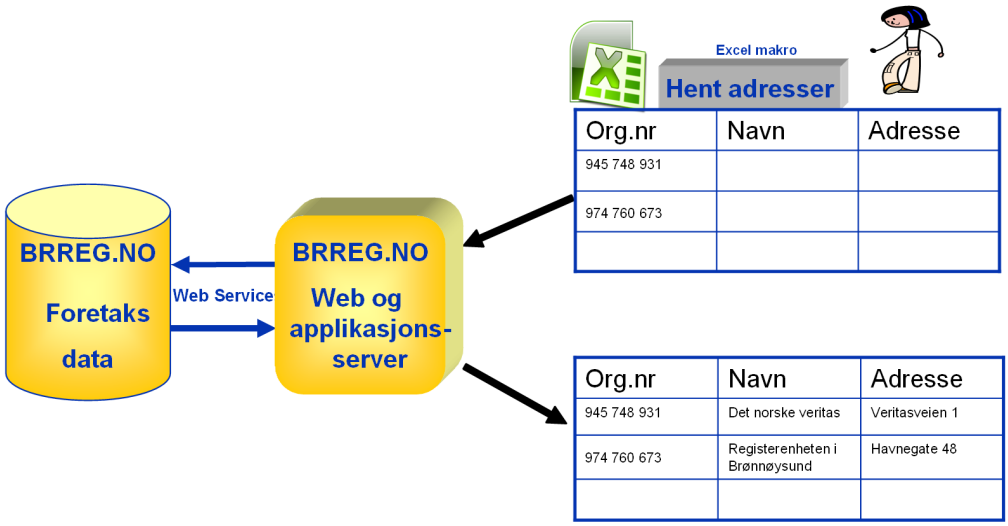
5.1 BR piloten, enklere tilgang til offentlige data

Brønnøysundregistrene (BR) har i mange år distribuert innholdet fra Foretaksregisteret i digital form, men ønsket sammen med Semicolon-prosjektet å prøve ut nye metoder for å gjøre disse dataene lettere tilgjengelig. Initiativet er delvis egeninitiert og delvis basert på det engasjement Fornyings-, administrasjons- og kirkedepartementet (FAD) viser i sin satsning på åpne offentlige data. I tråd med dette prøver nå Brønnøysundregistrene ut semantiske teknologi for å åpne opp og distribuere data.

Brønnøysundregistrene publiserte et nyhetsoppslag på sine web-sider i mai 2010 om at Semicolonprosjektet hadde laget en løsning for utprøving av semantiske teknologier på foretaksdata. Brønnøysundregistrene ønsket tilbakemeldinger på hvordan semantiske teknologier best kan benyttes for å dekke flest mulige behov hos brukerne. Se nyhetsoppslaget på: http://www.brreg.no/nyheter/2010/05/semicolon_tilgang_data.html

Den nye tjenesten gir mulighet for direkte oppslag i foretaksregisteret ved å spørre med organisasjonsnummeret som nøkkel. Du sender et spørsmål, men sender samtidig med beskjed om hvordan du vil ha svaret levert. Dermed får du mulighet for å få svaret vist tekstsmessig som HTML eller får svaret i RDF/XML struktur. Dette gjør at brukere som ønsker svaret vist i et brukergrensesnitt eller ønsker svaret direkte integrert i sin applikasjon kan bruke samme tjenesten.

I praksis vil dette bety at et oppslag på internettadressen https://ws.brreg.no/lod/enhet/974760673 gir et utvalg data om foretaket med organisasjonsnummer 974760673. I dette eksempelet er Brønnøysundregisterenes eget organisasjonsnummer brukt, men tilsvarende adresser er satt opp for alle norske foretak.



Illustrasjon 4, konseptuel skisse over BR-piloten, Foretaksdata som LOD

Som et trivielt eksempel på hvordan denne tjenesten kan benyttes med minimale krav til brukers IT-utstyr og kompetanse om semantiske teknologier, har Semicolon laget et excel regneark som ut fra en kolonne med organisasjonsnummer gjør oppslag mot Foretaksregisteret og fyller ut adresseinformasjon i celler på regnearket. På denne måten kan foretak, foreninger og andre holde adresselistene sine oppdatert med foretaksregisteret som kilde.

Løsningen er demonstrert en rekke ganger nasjonalt, men også på internasjonale konferanser. BR-piloten er blitt omtalt i SemanticWeb.com [Zaino10] og ut fra det vi vet er BR-piloten det første eksempelet på at et nasjonalt Foretaksregisteret er lagt åpent ut ved bruk av semantisk teknologi.

Fordelen med tjenester basert på semantisk-teknologi som denne er demonstrert sammen med SESAM4 prosjektet [SESAM11]. SESAM4 piloten består av en semantisk modellert kommersiell base med rating-informasjon om foretak. Denne er koblet med et stort nyhetsarkiv (ca 40 millioner nyhetsartikler fra CyberWatcher) og oppdatert foretaksinformasjon fra BR-piloten ved hjelp av semantisk teknologi.

I SESAM4-prosjektet er det foretatt en ekstraksjon av NFR prosjektarkivet, som inneholder mer en 29.000 prosjekter som har fått støtte av Norges Forskningsråd. Informasjon om prosjektene inneholder prosjektets tittel, beskrivelse, start- og stoppdato pluss prosjektets ansvarlige partner. I samarbeid med Semicolon (arbeidspakke 3/DNV) har denne konverteringen blitt koplet opp mot Brønnøysundregistrenes Foretaksregister som Linked Open Data. Dermed blir NFRs prosjektarkiv utvidet med ny funksjonalitet som har en dynamisk kobling mot Foretaksregister, noe som innebærer at endringer av informasjon om hovedpartnere slik den er registrert i Foretaksregisteret er direkte tilgjengelig.

Brønnøysundregistrene jobber aktivt med formidling av data og det pågår diskusjoner om å utvide piloten til flere av data-attributtene som ligger i foretaksregisteret. Videre vil det vært naturlig å vurdere om og hvordan de andre registrene Brønnøysundregistrene forvalter bør tilgjengeliggjøres ved bruk av semantisk teknologi. Brønnøysundregistrene sine nasjonale masterdata-kilder er ansett som kritiske informasjonsressurser i det moderne norske samfunnet.

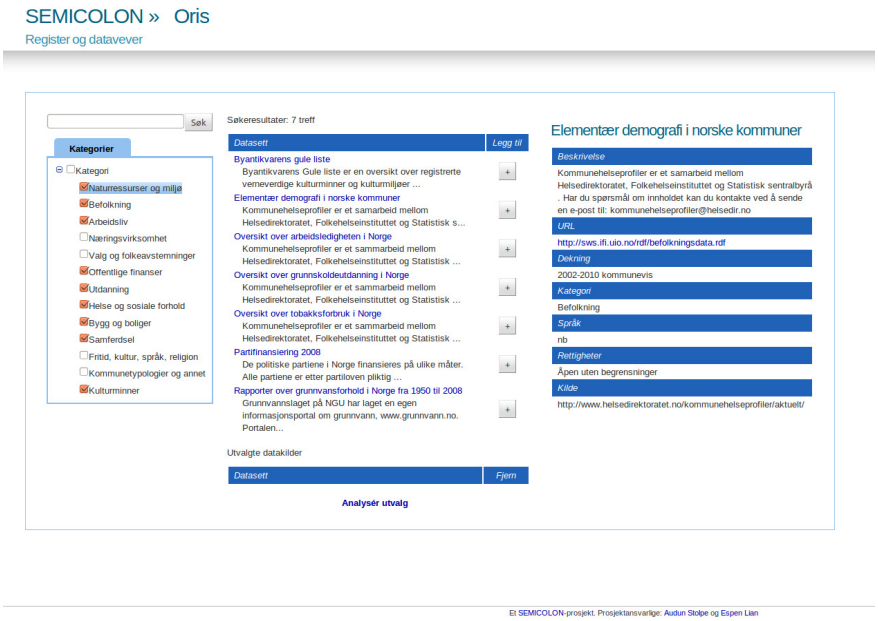
5.2 ORIS / SYNAPSE

I CASE-arbeidet og litteraturstudier i Semicolon erfarte vi at mengden åpne offentlig data vil komme til å bestå av mange tusen datakilder, og forståelsen av disse data vil ofte være vanskelig tilgjengelig for andre enn eier av data. Videre så vi at data publiseres i et mangfold av formater som gjør det vanskelig å sette sammen datakilder til større datasett evt. at kobling av distribuerte datakilder og søking på disse er vanskelig.

Denne bakgrunnen ledet til arbeid med å

1. etablere en infrastruktur hvor etater kan publisere metadata om sine datasett
2. etablere et verktøy hvor brukere av åpne data kan (i) velge datasett, (ii) gjøre spørringer og filtreringer og (iii) lage rapporter og visualiseringer basert på flere datakilder.

Løsningen ble en portal utviklet av Universitetet i Oslo, Institutt for Informatikk, som demonstrerer både infrastrukturen, hvor etater kan publisere metadata om sine datasett, samt en portal hvor brukere kan sette sammen, søke, manipulere og visualisere de datasett som er publisert. En vesentlig del av løsningen er formatkonvertering til formater som støtter semantiske spørrespråk som SPARQL.



Illustrasjon 5, ORIS oversikt over datakilder lokalisert på nettet.

I brukergrensesnittet lar brukeren velge sin kombinasjon av datakilder for viderebruk.

En god oversikt over bakgrunn og funksjonalitet i ORIS / SYNAPSE er beskrevet i [Giese10].

5.3 SSB case: Cambridge Semantics demo

Bakgrunn:
Basert på dialog mellom SSB og Semicolon ønsker SSB å gi Semicolon anledning til å demonstrere hvordan W3C standarder kan øke bruk og forbedre anvendelse av makrodata. De W3C standarder som er relevante er godt egnet for å sikre enkel viderebruk av data. Benyttet software er Anzo-suiten, levert av Cambridge Semantics.

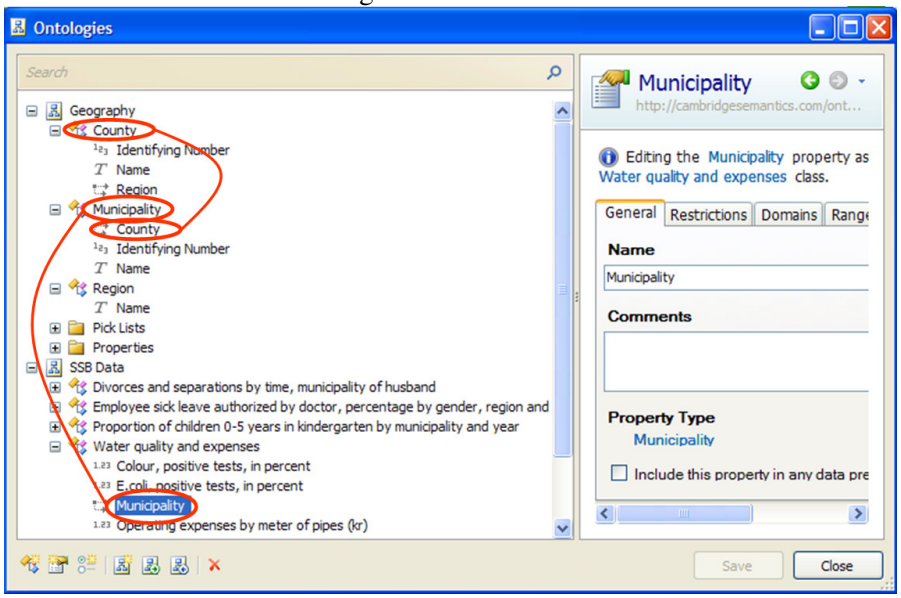
Formål:

- Teste ut nye teknologier og fremgangsmåter for å distribuere og visualisere innholdet i statistikkbanken.
- Resultatet trenger ikke inneholde spennende samfunnsmessige forskningsresultater ut fra dataene, men skal gi eksempel på verktøystøtte for å sette sammen datasett, filtrere datasett og visualisere resultater. Eksemplene trenger ikke ha avanserte statistiske beregninger som del av løsningen.

SSBs interne prosjekt for ”nye SSB” arbeider med overlappende initiativer, se neste avsnitt.

Resultatet av sommerprosjektet er beskrevet i mer detalj i rapporten Semantiske Teknologier for Distribusjon og Visualisering av Statistikk [Bruce10a]. Resultatene av et sommerprosjekt 2010 for Semicolon-prosjektet ved Universitetet i Oslo, Institutt for Informatikk.

Studentene demonstrerte med sitt arbeid hvordan man kan lage ontologier som beskriver forholdet mellom ulike statistikk-variabler, merke statistikken med vokabular fra ontologien, lagre den som RDF-tripler og visualisere statistikken i et web-grensesnitt, hvor sluttbrukeren selv kan bestemme hvilke data som skal sammenlignes.



Illustrasjon 6, Anzo sin ontologieditor som viser hvordan en mikro-ontologi for kommune-fylkes inndelingen i Norge kobles til et makro-datasett fra SSB.

Anvendelse av semantiske teknologier slik de er pakket inn i Anzo-suiten fra Cambridge Semantics tillater brukeren å generere grafer som inneholder data fra flere kilder så lenge disse data har en felles variabel/ dimensjon som for eksempel kommune eller tidsperiode. Brukeren kan selv velge hvilket aggregeringsnivå han ønsker å visualisere dataene på, for eksempel per kommune eller per fylke.

Vi ser at Anzo har stort potensial for å gjøre tilgjengelig offentlig statistikk for sluttbrukere. Anzo vil gjøre det lettere å samarbeide om statistikk og metadata. Muligheten for å laste opp egen statistikk vil gjøre det enklere for blant annet forskere og bedrifter å bedrive forskning. Piloten er godt beskrevet av studentene i deres rapport [Bruce10a]. De laget også en video som demonstrer løsningen brukt på SSB sine makrodata, videoen er lagt ut på www.youtube.com under tittelen *Semicolon and IFI with Anzo by Cambridge Semantics* [Bruce10b].

5.4 Nye SSB

Dette er ikke en pilot i Semicolon regi, men et internt prosjekt i SSB for å etablere ny informasjonsportal som skal være klar til høsten 2011. Prosjektet har følgende visjon og mål:

Visjon:

Ssb.no er stedet du vender tilbake til når du skal vite mer om det norske samfunnet.

Delmål:

- Ssb.no skal gi brukerne rask og enkel tilgang til fakta om det norske samfunnet, ved både overordnet og detaljert informasjon.
- Ssb.no skal bidra til at brukerne forstår hvordan tall og statistikk framkommer og henger sammen, slik at de kan omdanne statistiske fakta til kunnskap om Norge.
- Ssb.no skal drive folkeopplysning ved å gi brukerne informasjon de ikke visste at de trengte, og ved hjelp av ulike verktøy hjelpe brukerne til å anvende og å spre kunnskapen.

SSB har ca 5000 datasett som ligger tilgjengelig for visualisering og nedlastning i flere ulike formater. Således har ikke den nye offentlighetsloven påvirket mengden data SSB publiserer. Disse datasettene er knyttet til klassifikasjoner og variabeldefinisjoner. SSB har også tradisjon for å legge ut beskrivelser kalt *om statistikken*, som beskriver hvordan datasett er innhentet og bearbeidet, endringer i tidsserier etc. Klassifikasjonene til SSB er en form for små selvstendige ontologier som benyttes for å klassifisere og dermed også er egnet for å koble datasett via disse klassifikasjonene. I andre sammenhenger kalles slike klassifikasjoner gjerne mikroontologier.

Eksempel på slik mikroontologi er listen over land. Denne har SSB lagt ut på sine websider[6]. Semic.eu har forespurt SSB om tilgang til klassifikasjonene, så samme klassifikasjonen er tilgjengelig via Semic.eu sine distribusjonskanaler[7].

Det foreligger et skjema for unike identifikatorer (URN) for variabeldefinisjonene hos SSB, men dette er foreløpig ikke åpent utad. Med enkle grep vil disse identifikatorene kunne bli tilgjengelige som tjenester som svarer med sine beskrivelser når de blir kalt/benyttet.

Avsnittet er basert på samtaler og møter med SSB, samt presentasjoner og dokumenter fra eksempelvis [Løvbak10] og [Bjørnskau08].

6 ERFARINGER OG DISKUSJON

Dette kapittelet er strukturert etter forskningsutfordringene som ble satt opp initielt i plandokumentene. Gjennom prosjektets gang har vi gjort noen kursjusteringer og supplert med tema som definering av regime for URN, journalistenes ønsker knyttet til åpne data med mer.

Siste tiår har den digitale verden rund oss forandret seg. Det har vært et skifte til små mobile plattformer med prosesseringskapasitet og lagringskapasitet som tidligere var forbeholdt laptops. Det har blitt signifikant mer fokus på data versus software, og det har kommet føringer fra politisk hold om veien videre når det gjelder gjenbruk av offentlig data. Semicolon prosjektets målsetning er å bidra til enklere og billigere former for elektronisk samhandling der offentlig sektor deltar.

[6] Oversikt over land brukt i SSB sin social statistics hos SSB: http://www3.ssb.no/stabas/ItemsFrames.asp?ID=5588005&Language=en&VersionLevel=ClassLevel

[7] Oversikt over land brukt i SSB sin social statistics hos Semic.eu: http://www.semic.eu/semic/view/Asset/Asset.SingleView.xhtml?id=59513

Fordelen med åpne data blir ofte illustrert ved eksempler som er knyttet til privatpersoners fritidsbehov eller et ønske om å bedre oversikt over politikere og politiske partiers forslag, avstemming, finansiering, mm. Vi har sett mer på hva offentlig sektor trenger for å få etablert et mer effektivt informasjonssamfunn, og våre anbefalinger må sees i lys av dette utgangspunktet.

Det man observerer er at:

- Dagens offentlig og åpen informasjon er i vesentlig grad låst til informasjonseiers formater og tekniske tjenester.
- Offentlig sektor forvalter store mengder masterdata og registerdata for Norge som nasjon. Næringsliv og offentlig sektor er avhengig av tilgang og tilstrekkelig kvalitet på disse informasjonsressursene.
- Dagens elektroniske samhandlingsformer benytter ofte informasjonsutveksling som stafettpinne mellom organisasjoner i verdikjeder og verdinettverk. Realisering av slike løsninger er ofte kostbare og komplekse både å etablere og forvalte.
- Offentliglova av 1.1.2009 realiserer EU-direktiv om gjenbruk av den offentlige sektors informasjon (direktiv 2003/98/EF). Det fremheves i direktivet og motivasjonen for loven at god tilgang på offentlige data er viktig for å skape verdiøkende tjenester, og dermed for å fremme informasjonssamfunnet. Loven skal sikre mulighet for viderebruk av informasjon.
- Norske myndigheter (i likhet med myndigheter i andre land) har lagt føringer for å åpne for viderebruk data som er «samfunnseid» og som ikke bryter med personvernloven og opphavsrettigheter.
- Ved siden av initiativer som er igangsatt gjennom politiske vedtak, finnes det en gruppe med datasett som er blitt laget, forvaltet og publisert av grupperinger av entusiaster. Disse datasett inneholder ofte data som er laget av individer under en slags felles «dugnad». Noen av disse kildene er meget sentrale i LOD satsingen (for eksempel Wikipedia, Musicbrainz, mm).
- Det etableres konkurrerende datasett til de offentlige datasett. Disse har tilstrekkelig kvalitet for mange formål, og er egnet for viderebruk. Eksempel er kartinformasjon i OpenStreetMap, leksikon som Wikipedia, læremidler, historie.
- Motvilje mot å publisere er begrunnet med forklaringer av ulik art, eksempelvis kulturelle årsaker, merarbeid, personvernhensyn, usikker eller lav datakvalitet, rettigheter etc.

En av bærebjelkene i tanken rund semantisk web er at all data kobles sammen gjennom distribuerte og publiserte ontologier. Slike ontologier kan være åpne eller bare tilgjengelig innenfor en virksomhet. Ved å bruke applikasjoner som «forstår» semantikken i disse ontologiene, evt. jobber med samme modell, kan man lett integrere «fremmed» data i egne applikasjoner. Det vil si at trenden med koblede åpne data baserer seg på at distribuerte datakilder skapt og forvaltet under ulike rammebetingelser skal kunne bidra positivt til utvikling av informasjonssamfunnet.

6.1 Etablering, forvaltning og tilgjengeliggjøring av åpne data

Eierskap og dugnad

For å få etablert gode data og ontologier er det viktig at mange bidrar. Det er ressurskrevende å lære seg andres metadata og datasett, så en vil fort kunne velge vekk gjenbruk til fordel for egen

kontroll, egne data og kompetanse. Derfor trenger man gode incentiver for å bidra til å videreforedle felles datasett. Norge har forvaltningstradisjon for å etablere nasjonale registre for en mengde typer data, og dette er en fordel for nasjonen når det skal bygges en informasjons-infrastruktur.

Eierskap til resultatet, et datasett, og opplevelsen av å delta i innovasjon ser ut til å være sentrale nok drivere til at entusiaster bidrar med å bygge opp interessant innhold som legges åpent ut på nettet. Dette forutsetter at man er bevisst personvern og IPR[8], og ikke legger ut data som man ikke har rettigheter eller lov til å legge ut. Danskene har laget en juridisk veileder for dette formålet [IT-Tele10b]. Det eksisterer mye data som kan være av relevans for andre, selv om denne relevansen ikke er åpenbar i utgangspunktet. Et viktig punkt er at man aldri skal si at en datakilde «ikke er interessant». Hvis datakilden ikke blir brukt, så skader det heller ikke.

Basert på en enkelt litteraturstudie og diskusjoner i prosjektet har vi laget et eget appendiks kalt *Legal aspects of mashups.* I Dette appendikset er sjekklister som er relevant for dem som legger ut åpne data og for dem som tar i bruk åpne data.

6.1.1 Langtidsforvaltning av informasjon

Langtidsforvaltning av datakilder som er skapt av entusiaster er en utfordring som må håndteres. Dagens forslag til Linked Open Data har ikke mye fokus på tillit, datakvalitet, konseptdrift [Gulla10], evolusjon av ontologier, samt temporale aspekter ved data som tidsserier, kronologi i datasett og kontekst data må tolkes i. Dersom ontologier skal være i bruk og refererbare over lengre tid, så må eventuelle versjoner av en ontologi bli gjort eksplisitt slik at konsekvensene av evolusjonen kan tas i betraktning i applikasjoner. Og datasett må forvaltes slik at endringer i ontologi og datasett er dokumentert, konsistente og håndterbare over tid. Denne utfordringen drøftes grundigere i *Semantic technologies in Information governance* [Had+10].

6.1.2 Informasjonsforvaltning

Når flere kilder blir tilgjengelige og flere applikasjoner gjenbruker større menger av data, da blir det viktig at datakilder blir behandlet forutsigbart og metodisk, og at det publiseres metadata om datasettene som informerer om hva som er blitt gjort med dataene. Koblede åpne data er en infrastruktur for distribuerte datasett hvor datasett blir noder i en informasjonsgraf. Ulike subsett av disse nodene kobles sammen til ulike formål og brukere må kunne stille krav til forutsigbarhet på mange områder. Dette inkluderer: Stabilitet i tekniske grensesnitt, service level agreements, versjonering av ontologi for å kunne håndtere semantiske kaskadeeffekter mellom ulike informasjonsressurser, eventuelt oppdatering av datasett i forholdt til endret ontologi, konverteringer i formater, med mer.

Både de opprinnelige anbefalinger og det nyere «star-skjema» foreslått av W3C tilrettelegger for metoder som adresserer: definisjon av nye datakilder, analyse, rensing og vasking av data, og deretter konvertering/mapping og publikasjon av data.

8Intellectual Property Rights

6.1.3 Rådata

Når man skal publisere data som Linked Open Data er det viktig at man avveier om det er mulig å legge ut «rå-dataene» i stedet for behandlede data. Fra datafangst til bearbeidede data har ulike virksomheter ofte en lengre bearbeidingsprosess. Prosessene hos både SSB og SKD baserer seg på datafangst, kopiering av registre, kvalitetstesting og feilkorrigering. Hos SSB benyttes også en rekke avanserte statistiske metoder for å håndtere kjente svakheter eller skjevheter i rådata-tallmaterialet når analyser av data gjøres og statistikk produseres. Det er derfor flere nivåer av rådata-tilstander som må vurderes. For de fleste bruksformål antar vi at rådata som er kvalitetstestet og feilkorrigert er å anse som mest interessant fordi det er disse data som inngår i etatenes saksbehandling og tjenesteproduksjon.

Det er viktig at rådata beskrives så godt som mulig slik at det blir lettere for eksterne å forstå dataene som legges ut og bruke dem på riktig måte. Typiske eksempler er betydning av numerisk data. Vær tydelig på hva et bestemt tall betyr. Er det et relativ tall (som en prosentandel), absolutt tall (antall forekomster), er tallet målt ved en metode og henhold til en kjent skala (lengde, tyngde, tetthet, periode, tilstand, tidspunkt, bevegelse osv), i et definert populasjonsutvalg, målt under gitte forutsetninger, avgitt i en kontekst til et formål, osv?

Behandlet data, om det nå er parametere som slås sammen til nye parametere, eller om man filtrerer bort ekstremer («outliers»), er nesten alltid mindre «rik» enn ubehandlete data (på godt og vondt). Dermed legger man føringer som setter en stopper for bestemte bruksformål.

> For eksempel i et datasett om svømmebassenger blir lengde x bredde x høyde sammenslått til volum, siden det er den vesentlige parameter når man skal fylle de med vann. For en applikasjon som beregner vannbruken er det ikke noen vesentlig forbedring, men kan ses som en effektiv måte å forenkle på. Hvis man, noen år senere, trenger å finne alle bassenger til en konkurranse som er 25 meter eller lengre, så kan man ikke benytte tallmaterialet til dette formålet.

Derfor kan aggregering og bearbeiding av data på forhånd være problematisk. Man bør helst overlate denne jobben til sluttbrukeren (så vidt det er mulig og realistisk). Eksempelet viser hvorfor det ikke er anbefalt å bearbeide data slik at en oppgir aggregeringer og lignende uten å kjenne til senere bruk. Derfor er det et viktig prinsipp i LOD som sier at man ikke skal legge føringer for fremtidig bruk, og helst skal legge ut veldokumentert rådata.

6.1.4 Datakvalitet

I motsetning til uønskede aggregeringer på rådata, kan det derimot være riktig og hensiktsmessig å rense/vaske data. Typiske datakilder inneholder støy (data-feil) som kan forårsakes av upresis måling, språkproblemer, kjente skjevheter i populasjon, men også sensor-feil mm. Slike feil gjør at datagrunnlaget blir dårligere enn nødvendig, uten at man i etterkant har tilgang til nok kunnskap for å kunne behandle feilkildene i datagrunnlaget. Et typisk eksempel er at en spesiell temperaturmåler viser 2 grader for mye. Dataeksperten og data-eier vet normalt om slike feil, men hvis det ikke blir dokumentert, så er disse dataene mindre verdt for andre. I slike tilfeller

kan det være hensiktsmessig å gjøre en kvalitetshevningsjobb før man legger ut data. Andre eksempler på kvalitetsløfting av data er håndtering av duplikater, null og nil verdier, feil i kodelister eller fremmednøkler etc.

En dataeier kan pga sin domenekompetanse klare å benytte seg av data med kjente datakvalitets-utfordringer, fordi han kjenner hvilke problemer datakvaliteten medfører og kan kompensere for disse. Andre brukere har ikke denne kompetansen og vil ha utfordringer med å benytte de samme dataene selv til det samme formålet. Derfor er tydeliggjøring av hvilken kvalitet en datakilde har vesentlig for hvor egnet datakilden vil være for viderebruk.

Ulike bruksområder vil ha ulike krav til datakvalitet. Med tilgang til mange åpne datakilder vil det være mulig å vurdere to tiltak opp mot hverandre når det gjelder bruk av koblede åpne data:

- kvalitetsmessig løfting av åpne data fra primærkilde og sikre at eier tilgjengeliggjør data
- koble flere åpne datakilder og benytte nødvendige kvalitetsmetrikker og kvalitetskorrigeringer som gjør at den nye samlede kilden er tilstrekkelig egnet til formålet

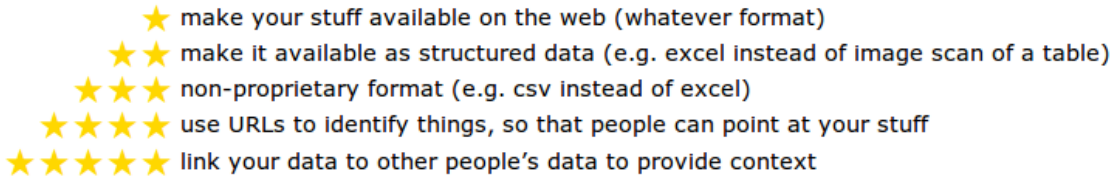
Det er viktig at det foretas en undersøkelse av hvilke datakilder som er tilgjengelige før man setter i gang arbeid med å koble sin virksomhet til en eller flere åpne datakilder. Både for tilbyder av åpne data og bruker av åpne data er det nyttig å sette opp noen prinsipper for hvordan man skal gå frem når nye datasett skal bygges opp. W3C har startet dette arbeidet gjennom sine Linked Open Data prinsipper.

6.1.5 Prinsipper for distribuering og tilgjengeliggjøring av Linked Open Data

For å hjelpe til med publikasjon av Linked Open Data har W3C jobbet med prosessen for å konvertere eller løfte eksisterende datakilder til strukturerte formater som forholder seg til standardene. Allerede for noen år tilbake ble det beskrevet et sett med anbefalinger for publikasjon av Linked Open Data (ref [TBL06]). Anbefalingen inneholder fire punkter:

1. Bruk URI som navn for ting du skal beskrive
2. Bruk HTTP/URI slik at man kan slå opp beskrivelsene
3. Hvis noen slår opp en beskrivelse, tilby verdifull informasjon ved å respondere på forespørselen ved hjelp av SPARQL/RDF
4. Inkluder lenker til andre datakilder og ontologier, slik at man kan slå opp og finne beskrivelse av data og finne flere interessante datasett.

Et viktig prinsipp er at man ikke nødvendigvis trenger å publisere data på en «perfekt» måte. Alle data som blir publisert på nett i en mer eller mindre strukturert form kan være av nytte for noen. Et par år etter anbefalingene til W3C ble det foreslått «Linked Open Data star scheme» som klassifiserer publiserte datasett på fem nivåer ([DERI10]):



Illustrasjon 7: Linked Open Data Star Scheme ([DERI10])

Dette forslaget gir en tydelig tilbakemelding til eventuelle brukere av en datakilde om hva man kan forvente av datakvalitet (hvor langt man har kommet i publikasjonen av en bestemt datakilde) og brukbarhet av den.

Ved å ta i bruk for komplekse teknologier som krever mye kompetanse vil vi også heve terskelen for både å legge ut data og begrense viderebruk. Dette er en problemstilling som vil bli vesentlig å finne en god balanse på i de kommende år, hvor ulike domener trolig vil lande på ulike ambisjonsnivåer.

Danskene har laget veilederen *Sådan udstiller du dine data* [IT-Tele10a]. Dette er en pragmatisk veileder som er egnet for å senke terskelen for å få ut åpne data som ikke støtter alle fasetter av W3C sine anbefalinger.

6.2 Etablering, forvaltning og tilgjengeliggjøring av metadata / ontologier

Det er primært to ting man kan legge ut på nett relatert til LOD: *ontologier* og *instans-data.* Det første punktet refererer til beskrivelsen av et domene og inneholder ikke noe faktisk data om individer i domenet. Disse ontologiene burde man publisere, selv om man ikke tenker å publisere data i første omgang. Kun publiserte ontologier blir gjenbrukt, noe som senere kan medføre at dine data blir enklere å berike eller dele. Eksempler på slike vokabularer er Dublin Core, vCard, Friend-of-a-Friend. I dagens norske offentlig sektor benyttes allerede Dublin Core i en rekke anvendelser både innen arkivområdet og som del av ulike fagsystemer.

Mikroontologier
Vi velger å se på en systematisering av begreper innen *tid* som en mikroontologi, det vil si nedbrytningen år – måneder – uker, samt kopling til kvartaler, sesonger, halvår mm er et eksempel. Et annet mikroontologi eksempel er land – fylker – kommuner, samt kopling til postdistrikt, politidistrikt, kirkesokn mm. Slike mikroontologier brukes som basis i mange datasett og i mange IT-systemer. Offentlig sektor er storforbruker av slike mikroontologier i fagsystemer, status og produksjonsrapporter, samt i sin statistikkproduksjon. SSBs klassifikasjonsdatabase består av et stort sett av mikroontologier [Stabas11]. Denne kilden er per i dag åpen, men kunne med fordel vært publisert i henhold til Linked Open Data prinsipper og markedsført sterkere som gjenbrukbare mikroontologier.

Offentlig sektor bør bruke sin spisskompetanse til å hjelpe brukere av åpne data med å

- gi eksempler på hvordan etatens data kan kobles sammen
- gi eksempler på hvordan etatens data kan kobles med andre etaters data
- vise hva som er kjente begrensninger for bruk av etatens data

Slike koblinger ses på som fletting av ontologier.

Kvalitet på ontologier
For å oppfylle sin funksjon må en ontologi være av en viss kvalitet. Noen av disse kvalitetsegenskapene vil kunne måles maskinelt, andre vil være vanskeligere å måle maskinelt. Dette er et område hvor en er kommet kort og mye arbeid gjenstår, men noen regimer for ontologikvalitet eller metadatakvalitet foreligger. Semicolon bidrag innen dette feltet er artikkelene *A data quality framework applied to e-government metadata* [MYR+10] og *Visualization of Complex Relations in E-Government Knowledge Taxonomies*. [MYR+11].

6.2.1 Regime for å identifisere ting, personer, informasjonsobjekter med mer.

Adressering og informasjon som gis tilbake
Det er en god, anbefalt praksis at man publiserer identifikatorer som leverer data tilbake når den blir forespurt (eksempelvis REST grensesnitt). Med andre ord: Sender man en forespørsel (request) til en URI, så skal man få tilbake en minimal beskrivelse av hva det er som beskrives (en ressurs, en relasjon, et begrep eller en data-instans). I tillegg bør nok metadata medfølge til at man kan identifisere objektet på en god måte. Spør man et system som er et repository for en ontologi eller metadata-repository som Semantikkregisteret for elektronisk samhandling [SERES] skal man få tilbake data om konseptet/begrepet.

Man kan også ha et ønske om å publisere komplette datakilder som inneholder både ontologier og instanser. Et eksempel på en slik kilde kan være DBPedia (semantisk versjon av Wikipedia). I så fall må man tenke rettigheter, eierskap og personvern, noe som gjør saken vanskeligere, mens det blir viktigere at man er ryddig.

6.2.2 Hvem definerer URIer / definisjonsmakten.

Det finnes ikke noen prinsipielle eller tekniske begrensninger på internett for hvem som kan legge ut begrepsdefinisjoner eller ontologier. Det er i prinsippet tillat at alle legger ut sine egne definisjoner og sine indentifikatorer til disse definisjonene. Om alle definerer betydning av sine data på ulike måter vil sammenstilling av flere åpne datasett være vanskelig.

I noen domener kan man imidlertid ønske seg litt mer styring, ikke minst for å kunne tilby data med forutsigbar og høy kvalitet. Et eksempel på et slikt domene er data fra det offentlige. Det er flere aktører i det offentlige som kan bidra med interessante data, og det er hensiktsmessig (og i god LOD ånd) å la aktører gjenbruke og publisere definisjoner i størst mulig grad. Lovgivende myndighet og offentlig sektors hierarkisk forvaltningsorganisering har ført til mange begreper som er overlappende, men ikke tilstrekkelig like. Dette gjør at det vil forekomme mange definisjoner med overlappende funksjonsområder. Dette gjør det vanskelig å gjenbruke data på tvers av offentlig sektor, og skaper usikkerhet ved viderebruk utenfor offentlig sektor.

DIFI har et pågående arbeid om nasjonal metadatastrategi som beskriver en prosess for hvordan harmonisering av felles begreper skal gjennomføres [DIFI10c]. Semicolonprosjektet har laget en tilsvarende prosessbeskrivelse, men med fokus på den enkelte etats behov for å etablere og forvalte sine begreper kalt *Livssyklusprosess for utarbeidelse og forvaltning av begrepssystemer* [Dalb+11]

Et godt eksempel er SSB definisjoner som er relatert til statistikk som legges ut i Norge. Statistikkene til SSB er mange og svære, og det er sannsynlig at det er mange mindre aktører som forholder seg til de samme dataene som SSB leverer. Dermed er det også hensiktsmessig for disse aktørene å gjenbruke definisjonene slik SSB benytter dem. SSB henter data fra mange offentlige kilder. Mange av variabeldefinisjonene til SSB er styrt av hva datakildene har som sine definisjoner. SSB har dermed allerede gjort en del harmonisering av sine begreper opp mot andre kilder.

Distribuerte kunnskapsnettverk og distribuerte data krever at alle objekter som blir omtalt er unikt identifiserbare. Til formålet har det blitt fremmet forskjellige forslag, fra Object Identifiers (OID) [OID10] til Published Subject Identifiers (PSI) [PSI10] og Uniform Resource Identifier (URI)/Unified Resource Name (URN) [URI10]. Eksempelvis så bruker Linked Open Data URI prinsippene til persistent publikasjon av objekter på nett [SAU07].

Ikke bare objektene burde være unikt identifiserbare, men også definisjonene av klasser og relasjoner mellom klassene i en ontologi. I mange tilfeller er ikke all tilgjengelig informasjon for et objekt like interessant eller relevant. I slike tilfeller garanterer denne fremgangsmåten at man kan kvalitetssikre og filtrere bort informasjon under bruk.

I sitt pågående arbeid med *nasjonal metadatastrategi* må DIFI sikre at arbeidet leder frem til anvendbare begreps- og definisjonskataloger som gjør metadata lett tilgjengelig og dermed også gjør data egnet for viderebruk. Det er ressurskrevende å etablere helhetlige tilnærminger for kompetanse, metoder og verktøy for å etablere begrepsapparat. SERES signaliserer ambisjoner om å bli en slik felleskomponent. DIFI har foreløpig ikke definert et felles ontologi / metadata-repository som en felleskomponent [DIFI10b].

Mye offentlig tjenesteyting og regulering er strengt styrt av lover og forskrifter. Dette fører til at en vesentlig del av begrepene som benyttes i datafangst og tolkning av data må gjøres i lys av gjeldende lovverk. Skatteetatens *Ligning ABC for 2010* lister 64 ulike lover for Skatteetatens forvaltningsområde [Lign10].

6.3 Indeks over åpne data kilder

Det finnes flere måter å finne igjen kilder med bestemt informasjon og en bestemt relevans/ autoritet/ kvalitet. Man kan ta utgangspunkt i en respektert ontologi og følge linkene fra denne modellen til ontologier for andre deler av «dataskyen». Man finner direkte frem til relevante tredjepartskilder og kan gjenbruke slik informasjon nesten helt automatisk. På den andre siden kan man bevisst foreta et informert søk (educated guess) gjennom indeks-portaler, som rapporterer om dataeiere og datakilder som er åpent tilgjengelige. En slik fremgangsmåte er mer arbeidsintensiv, siden det betyr at man må søke gjennom mange datakilder manuelt, men på den

annen side kan det bidra til en vesentlig bedre tilknytning av gode, hittil mindre kjente datakilder til LOD. Spørsmålet blir i så fall endret fra «hvor finner jeg en egnete datakilder?» til «hvor finner jeg en god dataindeks?».

LOD hadde ikke vært en «åpen» bevegelse hvis ikke det å lage en indeks også var en åpen, demokratisk prosess. Alle som vil kan starte en egen indeksside og begynne å beskrive og henvise til åpne datakilder på nett. Det finnes forskjellige initiativer som beskriver og/eller indekserer datakilder, med sider som Swoogle[9], Sindice[10], Sig.ma[11] mm. En mye brukt plattform blir utviklet av *Open Knowledge Foundation (OKF)* er en non-profit organisasjon som har som mål å bidra til åpne data med et ressursnettverk og verktøy. Det er OKF som står bak *Comprehensiv Knowledge Archive Network* (CKAN) som igjen er rammeverket brukt i offentlige nettportaler som data.gov.uk, data.norge.no, overheid.nl, offenedaten.de, fr.ckan.net, m.m., men også prosjekter som LOD/LOD2 som åpner opp og publiserer data fra EU kommisjonen. Status er at en ser ut til å gå fra listeoversikter til en søkemotortilnærming med kategoriseringer for å få oversikt over mulige datasett. Dette er en naturlig utvikling når antall datasett er høyt.

På verdensbasis finnes det noen initiativer som jobber med definisjon av begrepssystemer som beskriver «verdenskunnskap» i CYC and openCYC [LEN90] eller mer spesifiserte områder som næringsvirksomheter og produkter i UDDI/XML [UDDI10]. Felles for disse er at begrepsdefinisjonene er ment for å forstå og kople sammen begrep som brukes til resonering og/eller informasjonsutveksling mellom systemer og applikasjoner.

6.3.1 Gjenbruk av data og metadata – Hvordan finner man frem til interessante kilder?

Et av de viktigste prinsippene i LOD er at datakilder som blir lagt ut kobles sammen. Dette innebærer at man bør beskrive sitt eget domene så mye som mulig med definisjoner og begrep som allerede er publisert.

Markedsfør dine datasett, og sikre deg stor utbredelse av dine data. Din etats rolle i informasjonssamfunnet øker i betydning ved økt bruk. Det er kun informasjon som kan bli funnet som blir gjenbrukt. Fra «data-provider» perspektiv betyr dette at man ikke bare publiserer sin ontologi og sine data, men at man også må gjøre den synlig. Det kan gjøres på forskjellige måter. Det enkleste er at man lager et interessant datasett som er lenket mot noen store kilder. Når man så begynner å bruke datasettet (som inneholder relasjoner til velkjente og ofte brukte eksterne ontologier) er det sannsynlig at ditt datasett vil bli lagt merke til etter hvert. Denne vekststrategien av «ontologi-linking» er organisk, i og med at vekst i bruk avhenger av ytre faktorer og man kan håpe at semantiske søkemotorer finner publikasjonen av ontologien av seg selv.

En annen effektiv metode for å synliggjøre en ontologi og promotere den er «Ontologi pushing», en metode som innebærer at man går aktivt ut mot forskjellige kataloger for ontologier og

[9] http://**swoogle**.umbc.edu
[10] http://sindice.com/
[11] http://sig.ma/

semantisk annotert data for eksplisitt å anmelde en semantisk annotert datakilde. Det finnes mange initiativer med portaler som Sindice, SWOOGLE, Semantisk Web Search Engine (SWSE), HAKIA, FALCON, Watson, Evri, Nepomuk osv, alle opptatt av å indeksere semantiske annotasjoner og gjøre disse tilgjengelige til allmennheten. Ytterligere eksempler på indeksing tjenester og referanser se [HLW10]. De fleste av disse tilbyr tjenester for innmelding av nye datakilder.

6.4 Bruk og forvaltning av flere åpne datakilder som endres over tid

Det endelige målet med å lenke sammen data fra mange forskjellige kilder er en forbedring i tilgang til (mer) informasjon. Det kan tenkes at man har mulighet til å ta bedre beslutninger hvis flere forskjellige meninger/datasett tas i betraktning, eller informasjon rett og slett er mer detaljert.

En faktor som spiller en avgjørende rolle i et slikt scenario er om man kan stole på data-settene man har tenkt å (gjen-)bruke [MOR08]. Flere forskjellige dimensjoner avgjør om man skal eller ikke skal se ytterlige på en bestemt datakilde:

- Hvem som har laget eller vedlikeholdt dataene. Denne informasjonen kan være direkte avgjørende om man kan/tør/bør gjenbruke data. Ikke alle kilder er troverdige, noen kilder er uønsket, mens andre kilder kan være en «autoritet» på området og kan brukes direkte.
- Er datakilden beskrevet i nok detalj? En data-eier kan være akseptert som autoritet, men hvis datasettene ikke er beskrevet kan det være vanskelig å gjenbruke disse. Spesielt data som er kontekst- og tidsavhengig (aktualiteter, helseinformasjon, m.m.) kan være ubrukelig hvis ikke det fremkommer:
    - hvor gamle dataene er, eller hvilken versjon det jobbes med,
    - hva årsaken er til at datasettet ble etablert, og hovedoppgaven det var ment å løse,
    - hvilken kvalitet de forskjellige attributtene i dataene har,
    - hvilke konsepter/egenskaper og relasjoner som beskrives i dataene (ontologi data)

Det er altså viktig å forstå dataene før man begynner å bruke dem. Heldigvis tar LOD delvis hensyn til det gjennom et klart skille mellom ontologi og selve instansdataene. Dette innebærer at man kan utforske en ontologi for å bestemme nytteverdi før man tar i bruk hele datakilden. Ontologien kan i så fall brukes til både kvalitetssikring, innholdsdokumentasjon og i en senere fase som hjelpemiddel ved integrasjon, utveksling eller sammenslåing av informasjon.

6.5 Tilgjengeliggjøring av dynamiske data

Data som blir gjort offentlig tilgjengelig kan i hovedsak være *statisk* eller *dynamisk*. I tilfellet statisk data, kan man for eksempel konvertere datasettet til en RDF/OWL modell som legges ut og gjøres tilgjengelige gjennom en SPARQL-endpoint og RESTfull grensesnitt. Statisk data kan konverteres som en «snap-shot», tidsstemples, og trenger deretter ofte ingen oppdateringer.

Det motsatte scenariet innebærer at data er ekstremt dynamisk, for eksempel børsinformasjon, trafikkdata, informasjon som blir vedlikeholdt i wikipedia, offentlige baser /registre (Brønnøysund) eller statistikk og data generert av sensorer.

Fagsystemer som inneholder potensielle dynamiske åpne data er gjerne basert på kjent Relasjonsdatabase-teknologi (RDBMS). Det koster tid, penger og kompetanse å konvertere en slik base til en RDF-graf samt gjennomføre de nødvendige oppdateringer. En vanlig tilnærming i slike tilfeller er å etablere en mapping mellom entity-relationship modellen i RDBMSen og en RDF-graf/ontologi modell. Denne mappingen må gjøres eksplisitt og tilgjengeliggjøres for de involverte systemene. Bruksmønsteret blir da at en klient sender et SPARQL-spørsmål til et SPARQL endpoint som ligger på toppen av/ pakker inn RDBMS-implementeringen. SPARQL endpoint mottar innkommende spørsmål og oversetter disse til ER modellen. En PL-SQL/SQL spørring blir sendt til RDBMS systemet. Svaret fra RDBMS blir konvertert av SPARQL endpoint og resultatet leveres tilbake til klienten iht LOD design prinsipper. Dermed blir en slik base komplett transparent for sluttbrukerne og kan brukes i LOD sammenheng. Semicolon-piloten hos BR er et eksempel på en slik løsning, men vi benyttet oss der av web-services som allerede hadde pakket inn RDBMS løsningen. Det vil si at transformasjonene gikk fra LOD forespørsel => Web-services => RDBMS og deretter retur med svaret tilbake via to lag med transformasjoner. I mange tilfeller kan en slik løsning være fornuftig, men det er rapportert dårligere ytelse enn løsninger med en ”native” RDF-base[12].

6.6 Journalistenes behov

I arbeidet med LOD har Semicolon vært i kontakt med journalister i flere ulike roller. I disse diskusjonene har vi sammen satt opp følgende liste over ønskede kilder og noen egenskaper ved disse:

- Personrelatert informasjon, supplement til skattelister etc.
- Info om objekter/referenter (enkelt bedrifter, personer, geografisk sted, objekter av ulik art.)
- Repeterende saker relatert til eksempelvis politiske valg, årstider etc.
- Kart med lag av supplerende data
- Tidsserier / utvikling / endring
- Legg til rette for tema som undervisningssektoren årlig er opptatt av
- Asylsøking, innvandring og integrering
- Svineinfluensa, hvordan finne god bakgrunnsinfo?
- Levekårs-bakgrunnsdata og undersøkelser
- Helse, sykehusvalg, norsk pasient register (Norsk Pasient Register), etc.
- Helseinformasjon, koble hendelser til tid, sted og helsetilbud, hvor brekker flest lårhalsen og hvor ble behandlingen foretatt?
- Faktaopplysninger tilgjengelig til sammenligning med politisk uttalelser: Eksempelvis sykehuskøer
- Innsikt/ dybdesaker, trenger gode bakgrunnsfakta tar lang tid å finne frem om en i det hele tatt finner tilstrekkelig åpne data.

[12] http://www.w3.org/RDF/FAQ

I denne dialogen kom det frem at journalister sjelden har verktøy og kompetanse til å sette sammen enkle datasett. Bruk av regneark ble ansett som avansert databehandling for gjennomsnittsjournalisten.

7 ANBEFALINGER OG KONKLUSJON

Utgangspunktet for Semicolons arbeid med åpne data er bedre elektronisk samhandling i offentlig sektor og mellom offentlig sektor og publikum. Sett i et slikt lys er vi opptatt av hva som kan bidra til en bedre offentlig forvaltning og samfunnsøkonomiske gevinster. Mange av eksemplene som brukes for å illustrere fordelen med åpne data er rettet mot privatpersoners (fritids)behov og et ønske om å bedre innsyn i politisk aktivitet. Vi har brukt målbildet for semicolonprosjektet og sett på hva offentlig sektor trenger for å få etablert et mer effektivt informasjonssamfunn, og våre anbefalinger må sees i lys av dette utgangspunktet. Forslagene under må også sees som innspill til videre arbeid i Semicolon II prosjektet (2011-2013).

7.1 Oversikt over andres anbefalinger for å få økt nytte av LOD

Vanlige anbefalinger fra andre kilder som [FAD10a], [IT-Tele10a] og [ABM10]

- Orienter deg om prosessen med å åpne opp data steg for steg, se f.eks: [LIFE09]
- Legg ut dine rådata uten unødvendige endringer.
    - Dataeier vet ikke på forhånd hvem som kan ha hvilken nytte og evt. hvilken innovasjon etatens data kan inngå i.
- Legg ut metadata om ditt datasett
- Publiser dine data på for eksempel data.norge.no
- Bruk åpne og velprøvde lisenser og vurder juridiske sider, inkl. rettigheter og personvern før du publiserer
- Følg stjerneprinsippene for publisering, prøv å få så mange stjerner som mulig!
- Motiver for og lag piloter og produksjonsløsninger basert på åpne data

7.2 Semicolons supplerende anbefalinger

I tilegg anbefaler Semicolons Linked Open Data ressurser følgende i uprioritert rekkefølge:

1. Publiser mikroontologier eksempelvis SSB åpne klassifikasjonsdatabase [Stabas11] og etatenes anbefalte forslag til koblinger mellom datasett
    a. Mikroontologier er en type felleskomponent og bør forvaltes deretter.
    b. Etatene må hjelpe brukere med å lage gode koblinger mellom datasett basert på etatens data over mot andres datakilder.
2. Næringslivet, journalister, studenter og forskere / "Folket" trenger et operativt miljø hvor en kan sette sammen, filtrere og visualisere data uten at de trenger å investere i mye utstyr og kompetanse.
    a. Ikke bare store mediehus og næringsdrivende med egen IT-stab skal kunne bruke åpne data. Menigmann må få mulighet og verktøy som krever minimal teknisk kunnskap til å gjenbruke data/ lage mashups uten for stor kostnad.
3. Motiver en etat til å bli vert for og tilby viderekobling andre etaters åpne data kilder, fordi det krever tid og kompetanse å tilby gode tjenester for åpne data.

   a. Det blir dyrt om hver offentlige enheter skal etablere kompetanse og driftsløsninger for sine egne åpne data kilder. Vi ønsker mange offentlige åpne datasett, men de kan gjerne samles på noen få serverparker.
   b. Koble inn nasjonal metadata-strategi og fellesløsning for metadata direkte i LOD arbeidet og i en slik host-tankegang.
4. Etabler en arena for både data-jegere på linje med danskenes tiltak og brukere av ulike typer data. Et eksempel på en slik norsk arena som er etablert er Origo. Denne har deler av dette operativt allerede, men det er også eksempler på at de enkelte datakildeeierne ønsker å tilby og kanskje styre disse arenaene. Eksempel på et godt initiativ er http://tillegg.yr.no/
5. Etabler klare kriterier for hvordan åpne data skal beskrives og måles innen
   a. Datakvalitet
   b. Ontologi / metadata-kvalitet om datasettet / kilden
   c. Beskrivende informasjon om datasettets historie og forvaltning. Dette er informasjon som skal følge et datasett.
   d. Kobling til andre kjente datasett.
6. Visualiseringshjelpemidler for å finne frem i oversikter over både data og metadata.
   a. Enkeltetater (SSB) har allerede i dag flere tusen datasett som er åpent lagt ut. Dette fører til at terskelen for å velge optimalt datasett for en oppgave fort blir en spesialistoppgave.
   b. Etabler en oversikt over hvor metadata for offentlig sektor finnes, ikke bare data.
7. Etabler prosessteg for hvordan etater og virksomheter kan dra nytte av LOD for å bedre sin interne virksomhet og supplere interne datakilder. Entreprise linked data kombinert med LOD.
8. Få ut kommunenøkkelen og statskalenderen som LOD. Dette er enkle virksomhets-modeller for offentlig sektor og er nyttig kobling mot et mangfold av andre datasett.
9. Lag sjekkliste/ beste praksis/ kompetansemiljø for å få nasjonale masterdata ut som LOD (inklusive eksempler på hvordan etater har foretatt LOD innpakning av rdbms og web-services)
10. Legg ut rådata som er kvalitetstestet, feilkorrigert, men ikke aggregert, summert el. Denne tilstanden på data er å anse som mest interessant for flest formål, fordi det er disse data som inngår i etatenes saksbehandling og tjenesteproduksjon.
11. Markedsfør dine datasett, og sikre deg stor utbredelse av dine data. Stor utbredelse er nyttig argument i budsjettdiskusjonen både internt i organisasjonen og overfor eget departement. Din rolle i informasjonssamfunnet øker i betydning ved økt bruk.
12. Test ut lisenser for åpne data for å etablere erfaring med åpen distribusjon av data med ulike rettigheter knyttet til seg. Disse lisensene skal ivareta tilbyder, verdiøker og sluttbruker av åpne data, se appendiks *Legal Aspects Of Mashups*.

8 ARBEIDSFORM

I hovedtrekk har vi fulgt Semicolon I sin metode for casegjennomføring. Informasjonsinnhenting om mulighet og behov for åpne offentlige data er gjort i form av møter, intervjuer, telefonsamtaler og presentasjoner for Semicolon deltakere. Dette inkluderer partnere fra offentlig sektor, forskningspartnere og akademia. I tillegg har vi hatt møter, presentasjoner, mail-dialoger og eller samtaler med:

- NRK: Katrin Hellesnes, nyhetsredaksjonen
- Bergens Tidende: Are Sørenstuen, IT-dir.
- Aftenposten: Trond Myklebust
- VG: Jon Bones
- Stiftelsen Institutt for Journalistikk, Tord Selmer-Nedrelid og André Verløy
- Nettavisen, Gunnar Stavrum, redaktør
- NRK, Espen Andersen, researcher og journalist
- IT-Telestyrelsen i Danmark,
    - Christian Lanng, kontorsjef
    - Cathrine Lippert, kommunikasjonsrådgiver
    - Adam Arndt, Specialkonsulent/ datajeger
- Universitet i Bergen ved Andreas Oppdahl og Johan G. Erikson
- FAD: Sverre Andreas Lunde Danbolt
- DIFI: Kristian Bergem og Endre Grøtnes
- AFIN: Dag Wiese Schartum m. fl.
- Brukere av LOD-tjenesten fra BR.
- Styret i Den Norske Dataforening, faggruppe for Informasjonsarkitekturer og Semantic Web.

Semicolon har også deltatt på nasjonale og internasjonale workshops og konferanser hvor Linked Open Data som temaet har vært sentralt. Utøvende ressurser som har deltatt i arbeidet er:

Det Norske Veritas	Per Myrseth, Vibeke Dalberg, Rolf Petter Hancke og Henrik Smith-Meyer
EKor	John Helmut Rygg
ESIS Norge	Robert Engels
Universitetet i Oslo, Institutt for Informatikk	Audun Stolpe (post doc), Martin Giese (post doc), Arild Waaler (professor), Espen Lian (Phd stipendiat), Robert Riise (Phd-stipendiat), Lars Erik Bruce (student), Håvard Otterstad (student)

Ressurser utenfor Semicolon I prosjektet som har kommentert sluttrapporten på den prosjektintern høringen:

Computas	Pia Jøsendal, Stig Dormanen, Roar Fjellheim og David Norheim

9 REFERANSER

[ABM10] Åpen og samordnet tilgang til kulturarven. Anbefalinger for en vellykket tilstedeværelse i den digitale kulturelle verden. Robert Engels. 2. utgave. ABM utvikling 2010

[BER08] F. Berman. Got Data? A guide to Data preservation in the information age, Communication of the ACM December 2008.

[Bjørnskau08] Styringsdokument for nye ssb.no. Thomas Bjørnskau, SSB. 2008. www.ssb.no/Fomssb/Fanbud/FLEVERANSE_S1_1_1_Styringsdokument_nye_ssb_no.doc

[Bruce10a] Semantiske Teknologier for Distribusjon og Visualisering av Statistikk. Lars-Erik Bruce og Håvard Mikkelsen Ottestad. 16. september 2010. http://www.semicolon.no/Ottestad_Bruce_Rapport.pdf

[Bruce10b] Semicolon and IFI with Anzo by Cambridge Semantics. Av Lars-Erik Bruce og Håvard Mikkelsen Ottestad. http://www.youtube.com/watch?v=G913TGTXhK8

[Dalb+11] Livssyklusprosess for utarbeidelse og forvaltning av begrepssystemer. Vibeke Dalberg (DNV) ,Per Myrseth (DNV), Jim Yang (KITH). Semicolon leveranse februar 2011.

[DERI10] Linked Open Data Star Scheme. Proposed by Tim Berners-Lee, with examples by DERI. http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/

[DERI10] http://ld2sd.deri.org/lod-ng-tutorial/#history

[DIFI10a] DIFI-kvalitet. Quality measure regime for 700 Norwegian public web sites, metrics on universal usability and open data are part of the regime. http://kvalitet.difi.no/

[DIFI10b] Nasjonale IT-byggeklossar for det offentlege – Difi-rapport 2010:17. ISSN 1890-6583.
http://www.difi.no/artikkel/2011/01/nasjonale-it-byggeklossar-for-det-offentlege-difi-rapport-2010-17

[DIFI10c] ”Hvis det offentlige visste hva det offentlige vet …” Forslag til en nasjonal strategi for metadata. Utkast, 10. desember 2010.

[DNV+08] Div. authors. State of the art in interoperability for eGovernment. Semicolon deliverables. Report No. 2008-0996, Det Norske Veritas, December 2008. http://www.semicolon.no/Semicolon_SOTA_v1%200.pdf

[ENGEL73] The Augmented Knowledge Workshop. Douglas C. Engelbart, Richard W. Watson, and James C. Norton Stanford Research Institute • Menlo Park, California Presented at the National Computer Conference June 4-8, 1973 (AUGMENT,14724,)

[EUR11] EUROPEANA Consortium webpages. A European initiativ for preservationand publication of digital cultural heritage. 2011. http://www.europeana.eu/portal/aboutus.html

[FAD10a] Veileder om viderebruk av offentlige data skrevet i regi av Fornyings-, administrasjons- og kirkedepartementet. Anne Aaby, Anders Waage Nilsen, Anders Brenna og Pia Virmalainen Jøsendal. http://no.wikibooks.org/wiki/Viderebruksveileder

[FAD10b] Rådata skal kunne brukes av flere. Pressemelding , 17.11.2010. http://www.regjeringen.no/nb/dep/fad/pressesenter/pressemeldinger/2010/radata-skal-kunne-brukes-av-flere.html?id=624786

[Giese11] Sluttrapport til Semicolon, Institutt for Informatikk, Universitetet i Oslo. Giese, Martin. Stolpe, Audun. Waaler, Arild.

[Gulla10] Semantic Drift in Ontologies. WEBIST 2010 - International Conference on Web Information Systems. J.A. Gulla, Geir Solskinnsbakk, Veronika Haderlein, Per Myrseth, and Olga Cerrato.

[Had+10] Semantic technologies in Information governance. V. Haderlein, P. Myrseth, O. Cerrato. DNV Technical Report NO. 2009-1950, REVISION NO. 2

[HEFL+99] SHOE: A Knowledge Representation Language for Internet Applications, by Jeff Heflin, James Hendler, and Sean Luke. *Technical Report CS-TR-4078 (UMIACS TR-99-71).* 1999.

[HEPP07] Possible ontologies. How reality constrains the development of relevant ontologies. Martin Hepp, DERI. IEEE internet computing 2007.

[HLW10] Health Librarian Wiki Canada. Overview over semantic search engines. URL inspection date 12.12.2010. http://hlwiki.slais.ubc.ca/index.php/Semantic_search

[HORR08] I. Horrocks. Ontologies and the Semantic Web. Communication of the ACM December 2008.

[IT-Tele10a] Sådan udstiller du dine data. Teknisk vejledning til at sætte Offentlige Data I Spil. IT- og Telestyrelsen. Versjon 1.0, august 2010. http://digitaliser.dk/resource/559456

[IT-Tele10b] Når du udstiller dine data, Juridisk vejledning i at sætte Offentlige Data I Spil IT- og Telestyrelsen. Version 1.0, november 2010.

[KLW95] M.Kifer, G. Lausen and J.Wu. Logical foundations of object-oriented and frame-based languages. *Journal of the ACM (42:4).* ACM New York, 1995.

[LEN90] Douglas Lenat and R. V. Guha. (1990). *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley. ISBN 0-201-51752-3.

[LIFE09] Lifesized. Open Government Data Poster. 2009. http://www.vrijedata.nl/images/PosterScreen.pdf (original, Nederlandsk), http://www.vrijedata.nl/images/opendataPoster_english01.pdf (oversettelse, Engelsk)

[Lign10] Ligning ABC 25. juni 2010. 31. utgave. Skatteetaten. http://www.skatteetaten.no/no/Bibliotek/Publikasjoner/Handboker/Lignings-ABC/

[LOD10] Linking Open Data, a W3C community project. Started as an activity under the Semantic Web Education and Outreach (SWEO) Interest Group. http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[Løvbak10] Nye ssb.no. Per Olav Løvbak. Nordisk statistikermøde, København, august 2010. www.dst.dk/upload/4.6_nye_ssb.no.ppt

[MOR08] Luc Morea ET AL. The Provenance Of Electronic Data, *Communications of the ACM*. april 2008

[MYR+08] P. Myrseth, T. R. Christiansen and H. Gayorfar. Preservation of semantic value, - a study of the state of the art. Det Norske Veritas report 2008-0123. http://www.LongRec.com/Intranet/ResearchResults/StateOfTheArt/LongRec_Understand-SOTA_v1.01.pdf

[MYR+10] A data quality framework applied to e-government metadata. Per Myrseth, Jørgen Stang and Vibeke Dalberg. The International Conference on E-Business and E-Government (ICEE2011). http://www.icee-meeting.org/2011/

[MYR+11] Visualization of Complex Relations in E-Government Knowledge Taxonomies. Semicolon internal paper, to be submitted. By Per Myrseth, Jørgen Stang and David Skogan.

[OID10] Object Identifier: http://www.oid-info.com/#oid

[OREI10a] Tim O'Reilly. Governemnt as a Platform. Internet published. 2010.

[OREI10b] Tim O'Reilly. Open Government, or Government 2.0, as "Government as a Platform" on which citizens and build things for each other and participate in their government (rather than treating it like a vending machine).

[OWL04] OWL, http://www.w3.org/TR/owl-features/

[PSI10] Published Subject Identifiers (OASIS): http://www.oasis-open.org/committees/tm-pubsubj/docs/recommendations/general.htm

[RASI10] WeGovernment. Andrew Rasiej. Co-founder of Personal Democracy Forum and techPresident . 2010

[SEMI08] Semicolon Semicolon prosjektet. http://www.semicolon.no/

[SERES] Semantikkregisteret for elektronisk samhandling. http://www.brreg.no/samordning/semantikk/

[SHA+06] N. Shadbolt, W. Hall and T. Berners-Lee. The Semantic Web Revisited. IEEE Intelligent Systems 2006.

[SHOE00] The SHOE specification: http://www.cs.umd.edu/projects/plus/SHOE/spec.html

[Stabas11] SSBs klassifikasjonsdatabase, Stabas 2011. www.ssb.no\stabas

[SAU07] Cool Uris For The Semantic Web. http://www.w3.org/tr/2007/wd-cooluris-20071217/

[SESAM11] SESAM4 project (NFR funded). Semi-Semantic Models for Cross-Sector Portals. 2008-2011. http://www.sesam4.net.

[SKOS09] SKOS Simple Knowledge Organisation System Reference http://www.w3.org/2009/07/skos-pr

[TBL+99] *Weaving the Web* Berners-Lee, Tim, with Fischetti, Mark (Harper Collins Publishers,1999) ISBN 0-06-251586-1(cloth) ISBN 0-06-251587-X

[TBL+01] T. Berners-lee, J. Hendler and O. Lassila. The semantic web. May 2001 scientific american magazine http://www.scientificamerican.com/article.cfm?id=the-semantic-web

[TBL06] Design Issues For Linked Open Data. Tim Berners Lee, W3C. 2006 http://www.w3.org/designissues/linkeddata.html .

[UDDI10] Universal Description, Discovery And Integration. The uddi organisation. Url inspection date 12.12.2010. http://uddi.xml.org/

[UNCEFACT10] UN/CEFACT http://www.unece.org/cefact/

[URI10] Offisiele Beskrivelse Av Syntaks: http://tools.ietf.org/html/rfc3986

[URN10] Offisiel Workgroup Beskrivelse: http://datatracker.ietf.org/wg/urn/charter/

[WHIT09] Memorandum on Transparency and Open Government, White House, USA. issued on January 21, 2009. http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment/

[WIKIPEDIA2010a] http://en.wikipedia.org/wiki/Vannevar_Bush

[WIKIPEDIA2010a] http://en.wikiquote.org/wiki/Douglas_Engelbart

[Zaino10] Jennifer Zaino. Norwegian Semantic Web Project Latches On To Linked Open Data's Possibilities. October 27. http://semanticweb.com/norwegian-semantic-web-project-latches-on-to-linked-open-datas-possibilities_b829

APPENDIKS: PRESENTASJONER, PAPERS OG MEDIOMTALE

Følgende er en oversikt over åpen data relaterte presentasjoner om pågående arbeid og resultater.

2009

- Edok, DND konferanse. Offentlige data - hvem, hva, hvor? Per Myrseth. 13. nov 2009.
- eGovMonNet,workshop Ghent Belgia. "Some thoughts on measuring Semantic Interoperability on Public Sector Information". Per Myrseth. 30. nov 2009. http://epractice.eu/en/workshops/egovmonet4
- Per Myrseth og Audun Stolpe: Semantiske teknologier og informasjonsforvaltning hos Statistisk Sentralbyrå (SSB)". Invitert presentasjon ved Software 2009, 10.2.2009. (*)

2010

- Martin Giese. "Strategier for semantisk åpning av DBMS-datalager". Invitert presentasjon ved Software 2010, Oslo, 10. februar 2010. (*)
- Audun Stolpe: "Åpn opp dine data", foredrag ved 'Ut av siloene inn i nettskyen', seminar i regi av Dataforeningen 17, november 2010 Oslo/Kolbotn. (*)
- Audun Stolpe: "ORIS; med handlekurven full av åpne data", foredrag ved 'åpne offentlige data – piloter, erfaringer og anbefalinger' morgenmøte i Den Norske Dataforeningen 14. Desember 2010 Oslo. (*)
- Håvard Ottestad og Lars Erik Bruce: "Semantiske teknologier anvendt på SSB sine makrodata", foredrag ved 'åpne offentlige data – piloter, erfaringer og anbefalinger' morgenmøte i Den Norske Dataforeningen 14. Desember 2010 Oslo. (*)
- Linked open data, Per Myrseth. Track Informasjonshåndtering, DND, 10. februar. 2010. Software 2010
- Trender lokalt: offentlig data som LOD, status, utfordringer og muligheter. Per Myrseth og David Norheim. Track Semantisk Web i utvikling, 10. februar. 2010. DND, Software 2010
- A nation's presence on the Semantic Web. At the Semantic Days 2010. 31.mai 2010 Per Myrseth (Semicolon) and Robert H.P. Engels (Sesam4).
- Frislepp av nasjonale masterdata! Demonstrasjon av masterdata frå Foretaksregisteret som Linked Open Data (LOD). Per Myrseth. 10. juni 2010. Seminar i regi av Ressursnettverk for eForvaltning og NorStella
- Examples on how ontologies can ease aggregations, drilldowns and visualizations on macro data. Per Myrseth. Open Data Foundation Europe 2010 meeting, 7-8 of July 2010, Tilburg University, NL.
- Bedre kvalitet i begrepssystemer gir bedre semantisk interoperabilitet. Av Vibeke Dalberg og Per Myrseth. Seminar**:** Felles begrepsbruk i elektronisk forvaltning på lovstyrte områder. Del av undervisningen på masterstudiet i Forvaltningsinformatikk. (omhandler bl.a. åpne data og metadata for disse åpne data)
- Linked open data, Prosess for å etablere, forvalte og kvalitetssikre metadata/begrepsmodeller**.** DND Frokostmøte 14. desember 2010. Vibeke Dalberg og Per Myrseth, Det Norske Veritas AS

- Linked open data. Erfaringer og anbefalinger. DND frokostmøte 14. desember 2010. Robert Engels, ESIS og Per Myrseth, Det Norske Veritas AS

Begge understående artikler baserer seg på arbeid rundt foretaksregisteret og åpne av disse data for gjenbruk.

- Utilizing Ageing Public Sector Information. Scandinavian Workshop on e-Government, January 2010. Per Myrseth, Jon Atle Gulla, Veronika Haderlein, Geir Solskinnsbakk and Olga Cerrato. http://www.electronicgovernment.se/sweg2010/

- Utilizing Ageing Information. Accepted at the 10th European Conference on eGovernment. June 2010. Per Myrseth, Jon Atle Gulla, Veronika Haderlein, Geir Solskinnsbakk and Olga Cerrato. http://academic-conferences.org/eceg/eceg2010/eceg10-home.htm

Papers som er relevant for, men ikke avgrenset til kun å gjelde åpne data er:

- A data quality framework applied to e-government metadata. Per Myrseth, Jørgen Stang and Vibeke Dalberg. The International Conference on E-Business and E-Government (ICEE2011). http://www.icee-meeting.org/2011/
- Information Governance as a basis for cross-sector e-services in public administration. Terje Grimstad and Per Myrseth. The International Conference on E-Business and E-Government (ICEE2011). http://www.icee-meeting.org/2011/

(*) er også listet i [Giese11]. Akademiske papers som er listet i [Giese11] er ikke listet i oversikten over.

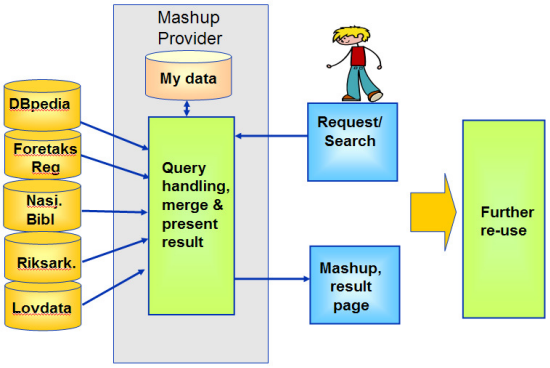
Medieomtale:

- Norwegian Semantic Web Project Latches On To Linked Open Data's Possibilities. Intervju med Per Myrseth på semanticweb.com. Publisert 27. oktober 2010. http://semanticweb.com/norwegian-semantic-web-project-latches-on-to-linked-open-datas-possibilities_b829

I tillegg kommer en rekke presentasjoner på prosjektinternt møter.

10 APPENDIKS: LEGAL ASPECTS OF MASHUPS

(This appendix is based on project discussions in both the Longrec project [HAD+10] and the Semicolon project. There is a short reference list local to this chapter.)



General Mashup architecture

In web development, a mashup is a web page or application that combines data or functionality from two or more sources to create a new service[13].

As an information provider on the internet, a LOD provider needs to identify legal aspects related both to own mashup offerings and, and a third-party-mashup using data from the LOD provider. The figure is a conceptual view of the components making up a mashup.

In our work we have discussed the following topics:

1) Protecting own rights and needs as a mashup and information provider
2) Protecting rights of 3rd parties (sub contractors / information providers), to "my" mashup.
3) Limiting own legal responsibility

These topics are listed in the checklist below.

10.1 Checklist of relevant topics to consider when establishing a mashups

10.1.1 Protecting own rights and needs as a mashup and information provider

In the intersection between a mashup vendor and its users, we have discussed the following topic:

- Pricing issues
- Own Intellectual Property Rights (IPR)
  - Limit what others can do with data/ licences
  - Is "My value added mix", subject to a new IPR? E.g. because of new metadata, linking information sources, quality lifting etc.
- Freedom to reuse, change, mix and sell others data
- Avoid multiple IPR on input to your mashup, relevant for:
  - Record, collection or database level
  - Types and layers of metadata
- Avoid multiple IPR on output
- Limit monopolies on input

[13] http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid)

- Avoid types of tech lock-in on input
  - E.g. avoid Digital Rights Management technologies on input
- Assure high quality guaranties from input providers
- Assure good provenance information and governance regimes on input
- State clear service level agreements towards sub vendors
- Assure that sub vendors are compliant with needed legislations

User generated / citizen generated content used in public sector mashups has not been discussed in depth. Example could be electronic material used for teaching where teachers and students collectively make the material based on public sector information.

10.1.2 Protecting rights of 3rd parties to "my" mashup

The following topics have been discussed as relevant for mashup vendors utilizing data supplied by sub vendors:

- Handle sub vendors IPR, Service level agreements and business model by
  - Transforming them to be part of own agreements towards your customers or
  - Assure transitive properties of sub vendors IPR, SLA and business model etc.

10.1.3 Limit own legal responsibility

The following topics have been discussed as relevant to limit own responsibility:

- Limit quality guaranties and consequences of errors in data
- State clear service level agreements
- Limit types and details of provenance information
- Limit requirements on information governance and compliance
- Limit other party's ability to intervene with internal: business, IT systems architecture and evolution, economy and business strategies.
- Assure own ability to change your mashup functionality, types of data offered, set of sub contractors, service level agreement, IPR etc
- Assure that users are compliant with needed legislations

10.2 Intellectual Property Rights (IPR)

Most jurisdictions recognize the following forms of IPR: trade secret, trademark, patent and copyrights.

For data openly accessible on the web it is in practice not possible/ or not easy to claim trade secrets, trademark or patent protection.

Internationally, it is quite uniform that one cannot claim copyright protection for individual entries of facts stored in a database.

In 1996, EU Introduced the database directive, to grant database creators a sui generis right towards unauthorized extraction of all or a substantial part of the database. This regulation is relevant if such extraction substantially harms the database creator's investments.

1996-2004 US: six bills suggest implementing a "US-database directive". None of them has passed into laws so far.

10.3 Conclusions and further work

While moving towards an information society, with more open information and where public sector as information provider has an active role, there is an increasing need for a sound juridical understanding of the rights and limitations of using public information in mashups. This topic is of high relevance to Norwegian public sector and to EU member countries based on the EU directive 2003/98/EF and the Norwegian implementation of this directive.

As part of further work there is a need for further literature studies and clarifications related to both Norwegian public sector as supplier of open data and as public sector as users of open data.

10.4 REFERENCES local to this appendix

This appendix is based on project discussions and the following literature.

1. Lehmberg, T., Rehm, G., Witt, A., & Zimmermann, F. (2008). Digital Text Collections, Linguistic Research Data, and Mashups: Notes on the Legal Situation. *Library Trends*, 57(1), 52–71.
2. Finding new uses for information. MIT Sloan Management Review, summer 2009. Vol 50 no. 4. Hongwei Zhu and Stuart E Madnick.
3. Does Current copyright law hinder innovation? MIT Sloan Management Review Winter 2009. An article by Jimmy Guterman discussing the book “Remix: Making Art and commerce thrive in the hybrid economy, New York: Penguin, October 2008.
4. One size does not fit all, Legal protection for non-copyrightable data. H. Zhu and S. Madnick. Communication of the ACM September 2009.