

Sluttrapport til Semicolon

Institutt for informatikk, Universitet i Oslo

Martin Giese Audun Stolpe Arild Waaler

19. januar 2011

Innhold

1 Introduksjon	2
1.1 Utfordringer og samarbeid	2
1.2 Åpne data og teknologi-bakgrunn	3
1.3 Oversikt over demonstratorer	4
2 Ifi-gruppen i semantisk teknologi	5
2.1 Personene som har bidratt til Semicolon	5
2.2 Publikasjoner og forskningsaktiviteter	6
3 Undervisning og foredrag	8
3.1 Kurs INF3580 – Semantiske Teknologier	8
3.2 Eksterne foredrag	9
3.3 Kurs ved Semantiske Dager 2010	10
4 Mastergradsprosjekter	10
4.1 Evaluering av spørringer over enveis-datakilder	10
4.2 Generering av XML fra semantiske datalagere	11
4.3 Åpne offentlige data for mobile håndholdte enheter	11
4.4 Transformering av tabulære data til RDF	12
4.5 Visualisering av offentlige data i Oris/Synapse	12
5 Demonstratorer	13
5.1 Brønnøysundregisteret i skyen	13
5.2 Oris/Synapse	14
5.3 Visualisering og analyse av makrodata fra SSB	21
6 Oppsummering og veien videre	24

1 Introduksjon

Denne rapporten oppsummerer bidragene og aktivitetene relatert til deltagelsen i Semicolon fra Institutt for informatikk (Ifi). Ifi har bidratt langs flere akser i prosjektet:

- Oppbygging av aktivt forskningsmiljø i semantisk teknologi ved Ifi
- Undervisning, eksterne foredrag og veiledning av studenter
- Tre kjørbare demonstratorer for tilgjengeliggjøring, visualisering og analyse av åpne data implementert på moderne plattformer for semantisk teknologi

Denne rapporten oppsummerer vår aktivitet de viktigste resultatene.

1.1 Utfordringer og samarbeid

Ifi har gjennom hele prosjektperioden deltatt aktivt i møter med partnere. Vi har samarbeidet spesielt tett med DNV og SSB, men har i løpet av prosjektet også brukt ressurser mot Skattedirektoratet og Brønnøysundregistrene, i tillegg til aktiv deltagelse på Semicolon-møtene.

Gjennomføring av caser som skal fremme semantisk interoperabilitet og samhandling krever stor modenhet og motivasjon hos partnere, siden slike caser typisk sikter på å sammenstille og integrere data fra ulike enheter og avdelinger i en organisasjon, eller sikter på utveksling og deling av informasjon mellom ulike organisasjoner. Betydelig tid i prosjektperioden har blitt brukt sammen med partnerne på undersøkelser for å avklare mulige caser. Dette arbeidet har i flere tilfeller ikke ledet til resultater som omtales i denne rapporten, men det har likevel bidratt til større forståelse både for Semicolon og for egne problemstillinger hos partnerne, og dermed bidratt til å sikre prosjektet bedre forankring. Samtidig har dette arbeidet gitt oss en innsikt i hva som kreves av samhandling i slike typer prosjekter som vi tar med oss videre.

Et eksempel på utfordringer vi har møtt er et forsøk på et prosjekt rundt mikrodata hos SSB. Vi støtte da på juridiske så vel som organisatoriske utfordringer. Det er for eksempel 18 ulike data-eiende seksjoner på SSB med til dels ulik juridikk, og disse har i utgangspunktet ikke tilgang til hverandres data. Utfordringene med å få tilgang til ulike mikrodata for sammenstilling

og analyse ble til slutt så omfattende at vi bestemte oss for se etter andre case enn mikrodata fra SSB.

1.2 Åpne data og teknologi-bakgrunn

Som et resultat av utfordringer med datatilgang bestemte vi oss midt i prosjektperioden for å fokusere på åpne offentlige data. Deltagende partnere forvalter flere interessante data i denne kategorien, og utfordringen er å legge til rette for at disse åpne data kan bli tatt i bruk, og gjerne også lede til innovasjon.

Med åpne offentlige data mener vi her informasjon som er insamlet og bearbeidet av offentlige instanser for diverse forvaltningsformål, og som det ikke hefter juridiske restriksjoner ved som hindrer at de kan tilbys digitalt over internett på en måte som fasiliterer analyse og gjenbruk. Flere åpne offentlige data på nett er et vesentlig steg i enhver overordnet strategi for demokratisering av informasjon, der siktemålet er en livsyklus bestående av insamling, strukturering, syndikering og analyse, etterfulgt av betimelig distribusjon for gjenbruk, innovasjon, verdiskapning og samfunnskontroll.

Den økonomiske og politiske betydningen av å tilgjengeliggjøre åpne offentlige data i maskinlesbare formater er forholdsvis velstudert allerede. Det er bred enighet om at en slik tilgjengeliggjøring bl.a.

- vil bidra til å sikre ansvarlighet i forvaltningen
- vil bidra til å opprettholde en velinformert offentlighet
- er vesentlig for overvåkingen av samspillet mellom miljø og helse
- understøtter analytisk/kritisk journalistikk
- stimulerer gründervirksomhet og økonomisk vekst

Internasjonalt har forskningen på tilgjengeliggjøring av åpne offentlig data fått betydelig momentum de senere årene, og er i dag i ferd med å konvergere i W3Cs (The World Wide Web Consortium, en standardiseringsorganisasjon sammenliknbar med f.eks. ISO) *Linking Open Data* prosjekt. *Linking Open Data*, heretter kalt LOD, betegner en metode for å publisere strukturerte data som gjør at data kan lenkes og sammenstilles på tvers av etater, institusjoner og faggrenser og således øke informasjonsverdien til hver enkelt kilde. LOD bygger på standard protokoller slik som HTTP, men snarere enn å benytte HTTP for å publisere presentasjonsorienterte websider, utveksles informasjon på maskinleselige formater. LOD prosjektet

er i dag en del av et større initiativ kalt *Semantic Web* som arbeider for en mer betydningsorientert/mindre presentasjon- og layoutorientert web.

Det er viktig å være klar over at Semantic Web-initiativet ikke kun er en agenda, men også et sett av standarder og teknologier som ofte refereres til som W3C-søylen eller Semantic Web-søylen. Kort fortalt er dette en suite av teknologier som forutsetter hverandre på en måte som kan visualiseres som en søyle eller kake. Nederst i denne kaken ligger RDF (*Resource Description Framework*) som er et grunnleggende representasjonsformat for å beskrive data og forholdet mellom data. RDF ledsages av spørrespråket SPARQL. Høyere opp i kaken finner man mer uttrykkskraftige formater slik som OWL og SKOS som er designet for å beskrive *klasser* av data for automatisk resonnering over- og/eller validering av- RDF-datasett. Jamført med andre standarder og teknologier (eksempelvis Topic Maps) er Semantic Web-initiativet spesielt velutviklet og godt teoretisk fundert.

Selv med fokus på åpne data har vi støtt på utfordringer med tilgang på data. Når en enhet publiserer åpne data er det vanlig at enheten trekker ut data fra sine databaser og publiserer en del av disse på en nettside. Når vi skal demonstrere mulighetene som ligger i semantisk teknologi, forutsetter dette i flere tilfeller tilgang på datakildene. Vi støter da igjen på utfordringer idet data som er åpne ofte ligger i samme database som data som ikke er det, og dermed ikke er åpent tilgjengelig fra sin kilde.

1.3 Oversikt over demonstratorer

I samråd med Per Myrseth (DNV) definerte vi tre demonstratorer for tilgjengeliggjøring av åpne data, designet over W3C-suiten av semantiske teknologier. De to førstnevnte demonstratorene ble gjennomført av postdocer ved Semicolon, mens den sistnevnte ble gjennomført av DNV i samarbeid med to sommerstudenter ved Ifi, og er også omtalt i sluttrapporten fra DNV.

Den første, kalt Oris/Synapse er en ment å være en publiseringportal og analyseredskap for åpne offentlige data. Idéen om en slik portal har vesentlige forløpere i den amerikanske portalen data.gov og den britiske data.gov.uk. Her hjemme har vi nå fått FADs data.norge.no. Oppdraget med å utvikle denne portalen ble gitt til Computas. Semicolon representert ved Ifi har, gjennom arbeidet med Oris/Synapse, vært vesentlig involvert i diskursen rundt utformingen av denne portalen. Oris/Synapse er nærmere beskrevet i 5.2.

Den andre demonstratoren er designet for å vise hvordan man kan gjøre

allerede eksisterende publiseringskanaler kompatible med LOD protokollen, og således synlige i den betydningsorienterte veven. Mer spesifikt ble det i samarbeid med Brønnøysundregistrene utviklet et LOD-adapter for Enhet-sregisterets webtjeneste som gjør at registerdataene kan behandles som RDF. Denne demonstratoren er nærmere beskrevet i 5.1.

Den siste demonstratoren utnytter allerede eksisterende programvare for å analysere, visualisere og publisere makrodata fra SSB. Hensikten er å si noe om hva det er mulig å gjøre, allerede i dag, med eksisterende hyllevare. Denne demonstratoren er nærmere beskrevet i 5.3.

2 Ifi-gruppen i semantisk teknologi

2.1 Personene som har bidratt til Semicolon

Budsjettrammen for Ifis deltagelse har vært i overkant av 4 årsverk. Disse midlene ble i sin helhet brukt i 2009 og 2010, slik at Ifis bidrag har blitt utført i de to siste årene av Semicolons prosjektperiode. Arbeidet som har blitt finansiert av Semicolon har fordelt seg slik:

- Audun Stolpe: Ansatt som postdoc på Semicolon fra 1.1.2009–31.12.2010. Stolpe har vært ansvarlig for demonstratoren omtalt i seksjon 5.2
- Martin Giese: Ansatt som postdoc i 75% stilling på Semicolon fra 1.3.2009–1.7.2010, resterende har blitt finansiert av Ifi. Giese har vært ansvarlig for demonstratoren omtalt i seksjon 5.1
- Espen Lian: Ansatt 6 måneder i 2009 og 9 måneder i 2010 (som utvidelse av sitt engasjement som PhD-student ved Ifi). Lian har arbeidet som programmerer på demonstratoren omtalt i seksjon 5.2
- Lars Erik Bruce og Håvard Otterstad: Ansatt på timebasis fra juni-september 2010. Bruce og Otterstad utviklet demonstratoren mot åpne data ved SSB. Demonstratoren er beskrevet av Lars Erik Bruce i seksjon 5.3 i denne rapporten.

I tillegg har Prof. Arild Waaler vært aktivt deltagende i flere av arbeidene. Waalers bidrag har blitt gitt utenfor Semicolon-budsjettet og dette er til dels blitt kompensert for ved at postdocene har avlastet Waaler i arbeid knyttet mer til oppbygging av forskingsgruppen enn til Semicolon direkte.

Semicolon-midlene har vært avgjørende for oppbyggingen av gruppen: vi har faktisk bygget et forskningsmiljø ved Ifi i semantisk teknologi rundt aktivitetene til Semicolon. I tillegg til forskerne som har vært finansiert av Semicolon, har gruppens aktiviteter blitt styrket gjennom tre PhD-stipendiater finansiert av Ifi:

- Henrik Forssell, med et prosjekt på integrasjon av databaser
- Martin Skjæveland med et prosjekt på utveksling av data mellom ulike databaser
- Kjetil Kjernsmo, med et prosjekt på føderering av SPARQL-spørringer

Alle disse prosjektene er relatert til aktivitetene i Semicolon.

I tillegg har vi i prosjektperioden skrevet veiledningskontrakt med 7 mastergradsstudenter med prosjekter relatert til Semicolon og to av disse har presentert foreløbige resultater på Semicolon-møter. To av studentene har ikke påbegynt selve arbeidet, de resterte prosjektene er nærmere omtalt i seksjon 4.

2.2 Publikasjoner og forskningsaktiviteter

Gruppen har balansert sin forskningsinnsats mellom anvendt forskning mot Semicolon-partnere og mer grunnleggende teoretisk forskningsaktivitet. Den anvendte forskningen har ledet til demonstratorer og foredrag som beskrevet i denne rapporten. Vi har imidlertid ennå ikke nådd å bringe resultatene fra disse prosjektene frem til internasjonal publisering. Dette har sin naturlige forklaring siden tidsperspektivet fra anvendt forskning mot partnere til ferdig publiserte resultater ofte er langt. For vår del tok det nesten ett år med "forundersøkelser" før vi konkluderte med konkrete prosjekter, og det resterende år er for kort tid for å bringe resultater med prosjekter som involverer data fra partnere og utvikling av eksekverbar kode frem til publisering.

Den teoretiske forskningsaktiviteten har ledet til publiseringer. Prosjektene er kun i beskjedne grad finansiert av Semicolon, det er heller slutføringen av publiseringsaktivitetene som har falt innenfor Semicolon-perioden. Disse prosjektene er relatert til Semicolon ved at de avklarer og gir nye bidrag til det teoretiske grunnlag for teknologiene og metodene som anvendes i demonstratorene.

To av publikasjonene i denne kategorien gir bidrag til grunnlaget for OWL, ontologispråket som brukes i Semicolon. Disse er:

- Lian og Waaler: *An Only Knowing Approach to Defeasible Description Logics (Extended Abstract)*. DL workshop 2009 [5].
- Lian, Johnsen og Waaler: *Confluent Term Rewriting for Only-knowing Logics*. STAIRS 2010 [4].

Tre andre publikasjoner utforsker premisser for normbaserte informasjoninfrastrukturer, en viktig rettsinformatisk side ved problemstillingene adressert av Semicolon:

- Audun Stolpe: Norm-system revision: theory and application. *Artif. Intell. Law* 18(3): 247–283 (2010) [6].
- Audun Stolpe: Relevance, derogation and permission: A case for a normal form for codes of norms. *DEON 2010 10th International Conference on Deontic Logic in Computer Science*, 7–9 July 2010, European University Institute, Fiesole (Firenze) [7].
- Audun Stolpe: A theory of permission based on the notion of derogation. *J. Applied Logic* 8(1): 97–113 (2010) [8]

Forskningsgruppen har avholdt seminar fast to timer i uka gjennom hele Semicolons prosjektperiode. Vi har også bidratt aktivt til det internasjonale forskningsfellesskap gjennom deltagelse i programkomiteer for viktige konferanser. Spesielt arrangerte vi konferansen *Tableaux 2009* ved Universitetet i Oslo, der Giese og Waaler ledet både programkomiteen og organisasjonskomiteen. Konferansepublikasjonen er:

- Giese og Waaler: *Automated Reasoning with Analytic Tableaux and Related Methods, 18th Intl. Conf., TABLEAUX 2009, Oslo, Norway* [2]

Waalder og Giese redigerer for tiden en spesialutgave av et tidsskrift som inneholder utvidelser av de beste arbeidene presentert i *Tableaux*-konferansen:

- Baader, Giese og Waaler: *Special Issue of the Conference “Automated Reasoning with Analytic Tableaux and Related Methods 2009”*. *Logical Methods in Computer Science*, publiseres 2011. [3]

Giese har også deltatt i redigeringen av en spesialutgave i den renommerte *Journal of Symbolic Computation*:

- Giese, Ireland, og Kovacs: *Special Issue on Invariant Generation*. *Journal of Symbolic Computation*, 45(11), 2010. [1]

3 Undervisning og foredrag

Øvrige programkomiteer vi har deltatt i har vært: IJCAR 2008 (International Joint Conference on Automated Reasoning), IJCAI 2009 (International Joint Conference for Artificial Intelligence), FTP 2009 (First Order Theorem-Proving), WING 2009 (Workshop on Invariant Generation), IJCAR 2010, KR 2010 (Principles of Knowledge Representation), WING 2010, LfSA 2010 (Logic for Systems Analysis), SCCS 2010 (Symbolic Computation in Software Sciences), EGOVIS 2010 (International Conference on Electronic Government and the Information Systems Perspective).

Waalder ledet i 2010 konferansen *Semantic Days* i Stavanger, en internasjonal konferanse som er blitt en viktig møteplass for offentlig sektor, industri og forskningsinstitusjoner. Semicolon hadde dedikert prosjektspor ved denne konferansen i 2010. Waalder var også nestleder for konferansen i 2009.

Gruppen har arbeidet aktivt for prosjektetablering. I samarbeid med resten av gruppen og bl.a. flere deltagere i Semicolon-konsortiet ledet Waalder et initiativ mot SFI-programmet der vi søkte om støtte til et *Center for Information Quality and Semantics (IQS)*. Søknaden fikk ikke støtte i denne runden, men det er allerede tatt initiativ til å videreføre arbeidet mot andre programmer.

I siste halvår har gruppen vært aktivt samarbeidende med Folkeregister-programmet. En målsetning er å etablere et forskningsprosjekt for å styrke arbeidet med et nytt folkeregister. Det er stor aktivitet i dette arbeidet i skrivende stund.

3 Undervisning og foredrag

3.1 Kurs INF3580 – Semantiske Teknologier

I den grad semantiske teknologier etter hvert tas i bruk er det viktig at informatikkutdannelsen formidler detaljert kunnskap om både teknologien og dens teoretiske fundament. I tillegg er Ifis bidrag i Semicolon, og også i det kommende prosjektet Semicolon II, i stor grad avhengig av å ha tilgang til en bred basis med studenter med relevant kunnskap.

Vi har derfor laget et helt nytt kurs "INF3580 – Semantiske Teknologier" ved Universitetet i Oslo. Kurset dekker alle de grunnleggende teknologiene som utgjør kjernen i dagens semantiske apparat. Dette omfatter RDF data-modellen, spørrespråket SPARQL, og modelleringsspråket OWL. Det ble lagt stor vekt på å undervise både den teknologiske siden og det teoretiske

grunnlaget for teknologien; det førstnevnte gjør studentene i stand til å skrive programmer som løser konkrete problemstillinger, mens kjennskap til grunnlaget er nødvendig for å forstå hva RDF data betyr. Kurset ble avrundet med noen forelesninger om innovative metoder som ikke ennå er del av standardrepertoaret, og en presentasjon av semantiske teknologier i kommersiell bruk ved David Norheim (Computas).

Kurset hadde omtrent 40 påmeldte i vårsemestret 2010, og samme antall i 2011. Dette gjør kurset til et av de mest populære kurs ved Ifi på bachelor-nivå. Studentenes tilbakemeldinger etter kurset i 2010 var meget positive.

3.2 Eksterne foredrag

Vi har holdt en rekke innlegg ved interne Semicolon-møter gjennom hele prosjektperioden. Følgende fem foredrag er inviterte foredrag ved eksterne møter knyttet til vårt arbeid med Semicolon:

- Martin Giese. "Strategier for semantisk åpning av DBMS-datalager". Invitert presentasjon ved Software 2010, Oslo, 10. februar 2010.
- Per Myrseth og Audun Stolpe: Semantiske teknologier og informasjonsforvaltning hos Statistisk Sentralbyrå (SSB)". Invitert presentasjon ved Software 2009, 10.2.2009.
- 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.

3.3 Kurs ved Semantiske Dager 2010

For å spre konkret kunnskap om den praktiske bruken av dagens semantiske teknologier ble det utarbeidet og avholdt et heldagskurs som formidlet grunnleggende kunnskap om:

- Semantiske teknologier og Semantisk Web generelt
- RDF datamodellen og syntaks
- Spørrespråket SPARQL
- Modelleringspråket OWL
- RDF-eksponering av relasjonsdatabaser med D2R.

Kurset var delt opp i fem økter med tre kvarters presentasjoner alternerende med pauser avsatt til praktiske øvelser.

Kursmaterialet er tilgjengelig på kursets hjemmeside.¹

4 Mastergradsprosjekter

4.1 Evaluering av spørringer over enveis-datakilder

Spørrespråket SPARQL tillater å formulere spørringer til et datalager på et "semantisk" nivå. Språket er like generelt som f.eks. SQL. Men mange eksisterende datakilder tillater ikke å evaluere vilkårlige spørringer. F.eks. kan man bruke Enhetsregistret i Brønnøysund til å finne informasjon om en bedrift bare hvis man kjenner til navnet eller organisasjonsnummeret, og ikke ut fra, for eksempel navnet til daglig leder.

Denne oppgaven handler om å lage en automatisk analyse av generelle SPARQL spørringer for å finne ut om og hvordan de kan besvares ved bruk av datakilder med denne type tilgangsbegrensning. Oppgaven veiledes av Martin Giese og utføres av Einar Myre, som nå har oppnådd første resultater.

¹<http://sws.ifi.uio.no/semdays2010/>

4.2 Generering av XML fra semantiske datalagere

Når datalagring og integrasjon basert på semantiske teknologier skal knyttes inn i en eksisterende infrastruktur, er det viktig at de nye komponentene kan generere data i samme format som eksisterende komponenter. I dag blir gjerne XML brukt til datautveksling, og formatet bestemmes nærmere ved bruk av “XML Schema” beskrivelser. Generering av XML som holder seg til et gitt skjema, gjøres ofte ved å skrive ad-hoc kode for eksportering fra datalageret. Det er svært vanskelig å sjekke at koden virkelig garanterer at skjemaet overholdes.

I denne oppgaven skulle det bygges et verktøy som tar et annotert XML skjemadokument, og genererer et XML dokument som er garantert å følge skjemaet. Annotasjonene er SPARQL spørringer som beskriver for de enkelte XML-elementene om og hvor ofte de skal opptre, hvilke attributtverdier skal brukes, o.s.v. Oppgaven ble utført av Trond Anders Sorvoja og veiledet av Martin Giese.

4.3 Åpne offentlige data for mobile håndholdte enheter

Mobile håndholdte enheter blir en stadig viktigere plattform for publisering av offentlige data, først og fremst gjennom utbredelsen av smarttelefoner. En smarttelefon tilbyr avansert funksjonalitet slik som internettilgang og GPS tjenester, og er nå utstyrt med prosessor- og minnekapasitet som gjør dem til troverdige plattformer for fremvoksende lokasjonsbaserte internetteknologier.

Denne oppgaven tar utgangspunkt i at en smarttelefon også kan være en hensiktsmessig plattform for integrering av offentlig informasjon. Mer spesifikt dreier det seg om å bygge en applikasjon for operativsystemet Android som publiserer offentlig data i tilknytning til det stedet en bruker befinner seg på på et gitt tidspunkt. Applikasjon, som nå er ferdigstilt i første versjon, henter data fra ulike kilder, eksempelvis byantikvarens gule liste, oversikt over ladestasjoner for el-biler, byggesaker og annet, og markerer dem i mobiltelefonens karttjeneste (Google Maps). Brukeren kan filtrere data etter kategorier samt berøre angitte punkter for mer informasjon.

Arbeidet er et samarbeid mellom Ifi og Computas, der Audun Stolpe er hovedveileder og David Norheim biveileder. Arbeidet utføres av Jens Kilde Mjelva og skal være ferdig i mai 2011.

4.4 Transformering av tabulære data til RDF

Mye data i offentlig sektor ligger lagret i formater som med en fellesbetegnelse kan kalles *tabulære*. Det kan dreie seg om Excel regneark, kommaseparerte filer, konvensjonelle databaser eller HTML tabeller. Det er videre slik at RDF—selve ‘grunnstenen’ i nettsemantikken—kan betraktes som en abstraksjon over tabellformen. Med andre ord egner RDF seg utmerket for å harmonisere og republisere allerede eksisterende kilder som adresserbare instansdata i en semantisk verdensvev. En avgjørende forutsetning er selvsagt at man har verktøystøtte for selve transformasjonsprosessen.

Denne oppgaven tar i utgangspunktet for seg transformering fra Excel til RDF som et første skritt på veien mot et mer generisk verktøy. Oppgaven innebærer å spesifisere et templatformat som angir regler for transformasjon av et gitt Excel ark til RDF samt å utvikle programvare for å utføre selve transformasjonen. Arbeidet inngår som en del av Oris/Synapse, jfr. 5.2 nedenfor, og er tema for mastergradsoppgaven til Fabian Alknes. Oppgaven skal være ferdig i mai 2011.

4.5 Visualisering av offentlige data i Oris/Synapse

Oris/Synapse er en portal som genererer og sammenstiller data fra offentlig sektor, og assisterer aktører i offentlig sektor med å konvertere tabulære data til RDF (jamfør forrige oppgave). Systemet er nærmere beskrevet i 5.2. En sentral komponent i Oris/Synapse er visualiseringsverktøyene som hjelper brukeren med å sammenstille, analysere og visualisere offentlige data fra ulike kilder. Per i dag tilbys flere slike verktøy; kakediagrammer, taksonomier, kart og spredningsdiagrammer. Flere er ønskelig, eksempelvis tidslinjer og grafer. Disse verktøyene styres gjennom spørrespråket SPARQL som velger ut variabler og verdier for visualisering.

Denne oppgaven vil videreføre arbeidet med datavisualisering i Oris/Synapse ved å gjøre hovedsakelig to ting: 1) legge til nye verktøy som gir andre blikk over dataene enn de vi allerede har, 2) se nærmere på hvilke typer spørringer som bør kunne standardiseres og tilbys brukeren i form av et valg mellom knapper, noe vi tror vil oppleves mer ‘intuitivt’ enn SPARQL.

Arbeidet utføres av Diyar Amin og skal være ferdig høsten 2011.

5 Demonstratorer

5.1 Brønnøysundregisteret i skyen

Ifi-gruppen har implementert et pilotsystem som eksporterer deler av Brønnøysundregistrenes (BRreg) enhetsregister som LOD (linked open data).

Enhetsregisteret inneholder for alle norske foretak informasjon som postadresse, forretningsadresse, daglig leder, telefonnummer, osv. Enkelte manuelle oppslag i enhetsregisteret har lenge vært mulig gjennom Brønnøysundregisterets nettsider. Ikke all informasjon blir vist, f.eks. får nettbrukeren ikke personnummeret til daglig leder.

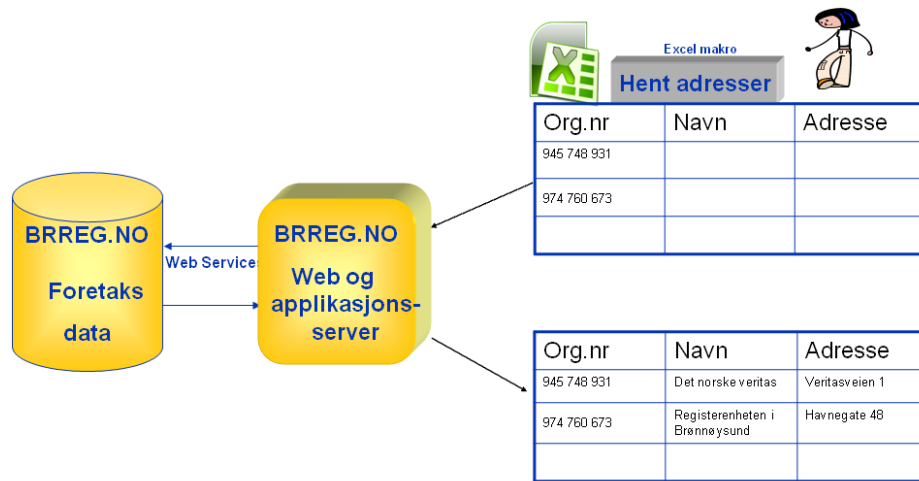
Brønnøysundregisteret selger også tilgang til enhetsregistret via en så kalt Web Service. Dette betyr at kundens programvare kan ta kontakt med tjener hos Brønnøysundregistrene for å få tilgang til all informasjon om et foretak lagret i enhetsregistret, gitt organisasjonsnummeret. Informasjonen blir levert i form av et XML dokument, formatert i henhold til visse, faste skjemaer utarbeidet av Brønnøysundregisteret. Til forskjell fra det manuelle enkelttoppslaget er resultatene dermed *maskinlesbare*. Det er derfor forholdsvis lett å utvikle programvare som bruker dataene.

Systemet, som er utarbeidet ved Ifi i samarbeid med Brønnøysundregisteret, tillater gratis tilgang til deler av enhetsregistret i maskinlesbar form. Informasjonen som blir gjort tilgjengelig omfatter for tiden bare postadressen til foretak, men systemet er laget slik at det med enkle midler også er mulig å inkludere annen informasjon, som for eksempel alt som er tilgjengelig gjennom manuelt oppslag.

For å gi en universell og interoperabel representasjon av foretaksinformasjonen, blir denne representert i RDF, som er det foretrukne kunnskap-representasjonsformatet i semantisk web-miljøet. Protokollen for å hente informasjonen betegnes som “linked open data”, LOD: Hver bedrift representeres ved en URI. For eksempel representeres et foretak med organisasjonsnummer 974760673 gjennom URIen <https://ws.brreg.no/loд/enhet/974760673>. Ønskes informasjon om foretaket, sendes det en vanlig HTTP forespørsel² for denne URIen til tjeneren `ws.brreg.no`, som svarer med et RDF dokument som beskriver foretaket.

Teknisk er løsningen realisert med en servlet som ble utviklet ved Ifi og kjøres på tjenerne i Brønnøysund. Når servleten skal behandle en forespørsel

²HTTP protokollen brukes til å hente nettsider fra tjener i den vanlige WWW.



Figur 1: Oppslag av foretaksadresser i et regneark via LOD tilgang (bilde av Per Myrseth)

sender den en forespørsel til den opprinnelige Web Servicen, og oversetter dataene som blir levert til RDF formatet.

Løsningen blir også omtalt på nettsidene til Brønnøysundregisteret.³ Der kan det også lastes ned et regneark som slår opp adressene til en hel liste med foretak gjennom den nye tjenesten.

5.2 Oris/Synapse

Bakgrunn. Idéen bak Oris/Synapse ble først unnfanget høsten 2009 i samarbeid mellom Per Myrseth (DNV) og Ifi. En lærdom vi på det tidspunktet hadde trukket av Semicolon tilsa at det både er mulig og hensiktsmessig å sondre mellom to ulike aspekter av interoperabilitetsproblemet. Det første omhandler *datautveksling etater imellom* der dataene typisk er unndratt offentligheten. I slike transaksjoner betinger ofte kompleks juridikk hvilke tekniske løsninger som er akseptable. Allikevel, siden hensikten med transaksjonen som regel er kjent, er det ofte mulig å komme frem til spesifikke løsninger for spesifikke problemer. Det andre aspektet av interoperabilitetsproblemet dreier seg isteden om *publisering av data* der man forutsetter at informasjonen er juridisk åpen og i prinsippet tilgjengelig for allmenheten. I motsetning til utvekslingsproblemet er juridikken ikke like mye av et hinder her, og de tekniske løsningene må være generiske snarere en spesifikke siden det ikke er kjent hvordan dataene vil bli brukt.

³http://www.brreg.no/nyheter/2010/05/semicolon_tilgang_data.html

Overordnet hensikt. Oris/Synapse adresserer den andre av disse to utfordringen og er således ment å være et publiseringsverktøy for åpne offentlige data som tilbys allmenheten for vilkårlig gjenbruk. Mer spesifikt er Oris/Synapse en portal som genererer og sammenstiller semantiske data, der den overordnede hensikten er todelt:

1. Å hjelpe allmenheten med
 - (a) å finne og analysere offentlige datakilder, og
 - (b) å gjenbruke ev. republisere data for egne formål
2. Å hjelpe etater og dataeiere med
 - (a) publisere RDF data, og
 - (b) eksplisere og gjøre konkret bruk av sine begrepsdefinisjoner.

Designprinsipper. Oris/Synapse er konstruert i henhold til et sett av designprinsipper som er ment å fremme ovennevnte målsetning, samtidig som de ivaretar vesentlig interesser hos dataeierne, slik som kontroll over data og datakvalitet. De vesentligste av disse prinsippene er de følgende:

1. Distribuert lagring: Oris/Synapse inneholder ikke selv data, kun henvisninger til datakilder. Datakildene forblir hos dataeierene som derfor beholder kontroll. Dette gjør forøvrig at Oris/Synapse uten videre er kompatibelt med andre registre, eksempelvis CKAN.
2. Ett felles format: Dataene forutsettes å foreligge i ett felles format designet for utveksling og viderebruk, nemlig RDF. Dette gjør at dataene kan flettes og analyseres i kombinasjon, og gjenbrukes utenfor sin opprinnelige kontekst.
3. RDF-hjelp til dataeierene: Oris/Synapse inneholder en komponent som gjør det enkelt for dataeiere å transformere tabulære data (eksempelvis Excel og CSV) til RDF, og oppføre det resulterende datasettet i registeret.
4. Gjenbrukbarhet av alle ressurser: *Alle* informasjonsressurser er kodet i RDF. Dette gjelder metadata om datasettene, *klassifisering* av datasettene samt datasettene selv. Fra et teknisk synspunkt er det ingen forskjell på disse ressursene, som derfor kan behandles med én og samme verktøyportefølje.

5. Full maskinleselighet: Oris/Synapse antar ingenting om brukeren av systemet. All funksjonalitet som tilbys i det grafiske brukergrensesnittet (visualisering unntatt, selvfølgelig) er også tilgjengelig for maskinell prosessering. Hensikten er at Oris/Synapse skal kunne støtte utviklingen av nye datadrevne tjenester og nettapplikasjoner.
6. Semantisk skalerbarhet: Oris er bygd for å være semantisk skalérbar i følgende forstand: Lite eller ingen kjennskap til semantiske teknologier forutsettes på dataeiersiden for å tilknytte seg registeret. Som et RDF-basert register er imidlertid Oris/Synapse 'vennlig innstilt' til semantiske beskrivelser av datasett skulle noen finne det hensiktsmessig å benytte slike. Det vil si, ingen programvareutvidelser er i utgangspunktet nødvendig for å ta i bruk ontologispråk slik som OWL og SKOS.

Det er vår oppfatning at disse prinsippene vil bidra til å komme vekk fra 'kikkehullene' som i dag er den dominerende mekanismen for datapubliserings i offentlig sektor. Med kikkehull mener vi her webtjenester som implisitt betrakter enkelte typer informasjon som primære og dermed, igjen implisitt, også gjør antagelser om hva som er viktig for *brukeren* av dataene. Et typisk eksempel er Enhetsregisterets webtjeneste der tilgang til data om bedrifter og foretak forutsetter et kjent enhetsnummer. Dette utelukker bruk av enhetsregisteret for f.eks. stedsbaserte søk—noe som stadig blir mer aktuelt gitt fremveksten av mobil GPS teknologi—der man vil være nødt til å søke etter enhetsnummere basert på lokasjon og/eller adresse snarere enn omvendt.

Rent generelt er informasjon en avgjørende faktor i innovasjon og verdiskapning, og potensialet i et datasett forsterkes gjennom innovativ gjenbruk. Det er derfor tilrådelig å ikke anta noe om hvilke data en sluttbruker først og fremst vil være interessert i, slik man i realiteten gjør når man tilbyr et kikkehull.

Systemkomponenter og virkemåte i Oris/Synapse Synapse. Registreringsgrensesnitt og konverteringshjelp:

- konverterer Excel regneark (etterhvert også CSV, XML o.a.) til RDF, basert på et enkelt templatformat som definerer regler for transformasjonsprosessen.
- laster opp metadata om det resulterende datasettet til Oris
- tilbyr i tillegg et registreringsgrensesnitt for datasett som allerede foreligger i RDF

5 Demonstratorer

SEMICOLON » Synapse
En nevrotransmittor for åpne offentlige data

Konvertér Registrér

Tittel		Dekning	
Proveniens		Utgiver	
Kilde		Plassering	
Språk		Utgivelsesdato	
Type	Naturressurser og miljø	Rettigheter	

Beskrivelse

Registrér

Om Semicolon

Samhandling mellom etater, næringsliv og borgere er en nødvendig forutsetning for å effektivisere og forbedre offentlig tjenesteproduksjon, samt for å stimulere verdiskapning basert på gjenbrukbarhet av åpne data. Semicolon (I) er et NFR-støttet forsknings- og innovasjonsprosjekt som pågår ut 2010. Semicolon har fokus på økt elektronisk samhandling i offentlig sektor, både etater imellom samt mellom borgere og forvaltning.

W3C-standarder

- RDF
- SKOS
- OWL
- SWRL
- SPARQL

Modelleringsparadigmer

- Unified Modelling Language
- Object Role Modelling
- Topic Maps
- F-Logic
- Description Logics

Øvre ontologier

- DOLCE
- OCHRE
- BFO
- SUMO
- PROTON

Verktøy

- Protogé
- NeOn
- QuOnto
- TopBraid
- OntoBroker

Norske endepunkter

- Semicolon Oris
- Computas LOD server

Figur 2: Synapses registreringsgrensesnitt

Oris. Register og datavever:

- tilbyr søk i mengden av registrerte datasett vha. kategorier og fritext
- Oris inneholder i tillegg en 'handlevogn' som fungerer slik at
 - brukeren legger utvalgte datakilder i vognen
 - og ber Oris om å flette dataene
 - dataene presenteres i et grensesnitt som tilbyr analysefunksjonalitet i form av søk og visualisering
 - brukeren kan til slutt generere nye RDF filer av resultatene, for gjenbruk

Oppsummert kan Oris/Synapses samlede virkemåte altså beskrives slik: Synapse gir anledning til å registrere datasett i Oris, som på sin side best kan beskrives som et distribuert register og analyseverktøy. Et typisk eksempel

SEMICOLON » Oris

Register og datavever

The screenshot shows the Oris web application interface. At the top, there is a search bar with the text 'Søkeresultater: 7 treff'. Below the search bar, there is a 'Kategorier' (Categories) section on the left with a list of categories and checkboxes. The main content area displays a list of datasets under the heading 'Elementær demografi i norske kommuner'. Each dataset entry includes a title, a brief description, and a 'Legg til' (Add) button. On the right side, there is a detailed view of the selected dataset, showing its description, URL, and other metadata. At the bottom, there is a section for 'Uvalgte datakilder' (Selected data sources) with a 'Fjern' (Remove) button and an 'Analysér utvalg' (Analyze selection) button.

Et SEMICOLON-prosjekt. Prosjektansvarlige: Audun Stolpe og Espen Lian

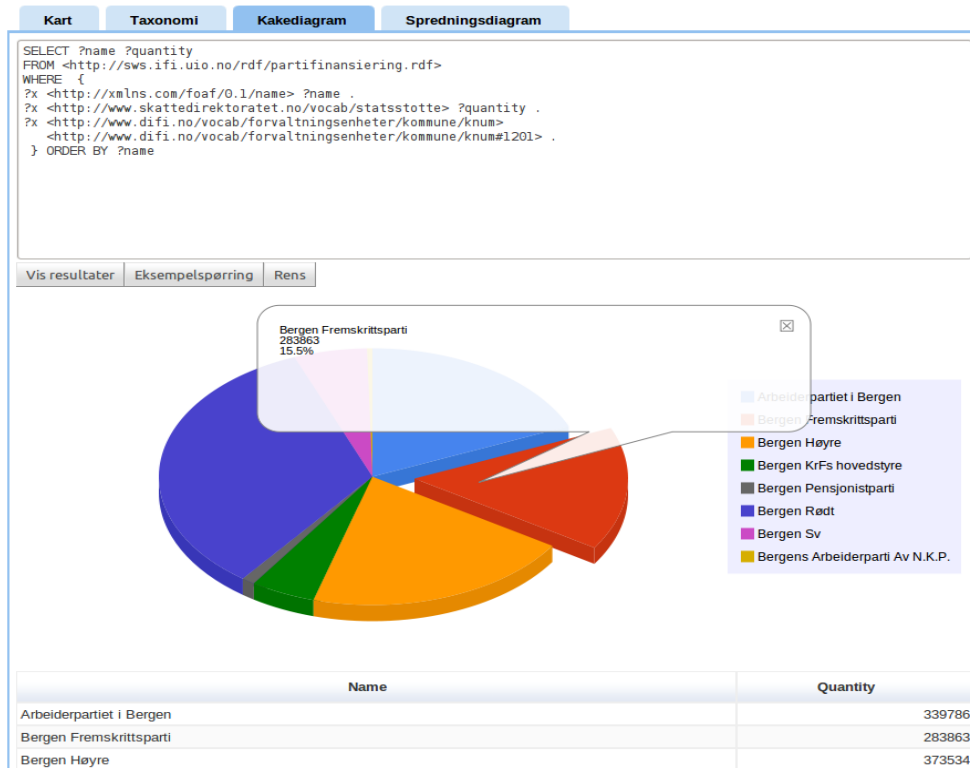
Figur 3: Oris' registergrensesnitt

på anvendelse kan gå som følger:

1. Excel ark og RDF templatfil lastes opp til Synapse: Regnearket transformeres til RDF vha. angivelser i templatet og returneres til bruker for nedlasting.
2. Bruker lagrer den resulterende RDF-grafen på en URL som angitt i templatfilen.
3. Metadata om datasettet samt URL til dets plassering lastes opp til Oris hvor det gjøres tilgjengelig i et søkegrensesnitt.
4. Oris henter datasettet på angitt lokasjon på forespørsel av bruker.
5. Brukeren sammenstiller og analyserer dette med andre registrerte datasett han/hun måtte finne interessante, og gjør eventuelt et uttrekk av resultatene.

SEMICOLON » Oris

Register og datavever



Figur 4: Oris' visualiseringsgrensesnitt: Kakediagram.

Prosjekterte utvidelser. Oris/Synapse er foreløpig å betrakte som en 'proof-of-concept' implementasjon, og er ikke ennå et produksjonsklart system. Vi kommer imidlertid til å forsøke å opprettholde momentum i utviklingsarbeidet, først og fremst i form av tilknyttede mastergradsprosjekter.

Prosjekterte forbedringer og utvidelser er av forskjellig art slik som sikkerhet, administrasjon av den bakenforliggende databasen, videreutvikling av søkegrensesnittet o.a.

Mest prekärt, og derfor verd å nevne her, er behovet for å sette Oris/Synapse i stand til automatisk å oppdatere datasett som publiseres med regelmessige intervaller, slik f.eks. en del datasett i Statistikkbanken, eller byantikvarens gule liste gjør. For å imøtekomme dette behovet, vil Oris/Synapse versjon 2 inneholde en kanalaggregator som 'lytter' til RSS kanaler dedikert datasett med regelmessige generasjonsintervaller. Idéen er at offentlige dataeiere

5 Demonstratorer

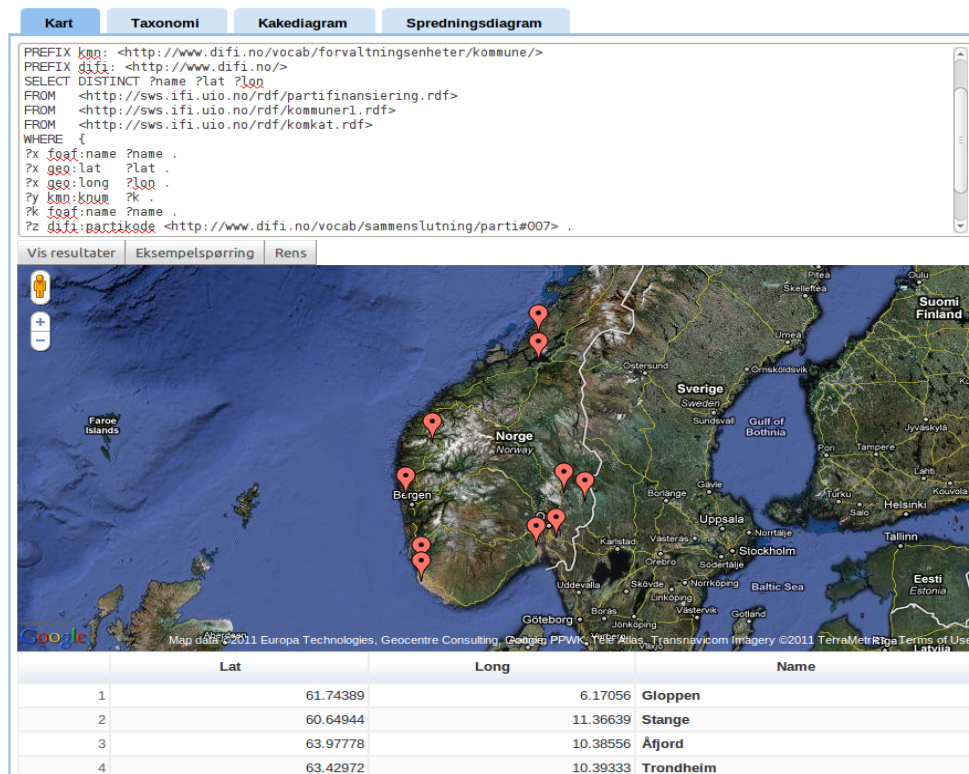
skal kunne annonsere sine data ved å publisere en RSS kanal istedenfor et enkeltstående datasett. Denne kanalen forutsettes

1. å være dedikert én enkelt type data eller datasett
2. å angi en publiseringsfrekvens for datasettet
3. å angi en utløpsdato for datasettet
4. å selv være uttrykt i RDF, eksempelvis som RSS 1.0.

På denne måten vil Oris/Synapse kunne bli tilnærmet selvadministrerende mhp. denne typen datasett.

SEMICOLON » Oris

Register og datavever



Figur 5: Oris' visualiseringsgrensesnitt: Kart.

5.3 Visualisering og analyse av makrodata fra SSB

Denne demonstratoren ble utviklet av Ifi-studentene Lars Erik Bruce og Håvard Otterstad sommeren 2010. I samarbeid med DNV skulle de undersøke mulighetene for å anvende W3C-standarder for å visualisere og distribuere innholdet i Statistikkbanken til Statistisk Sentralbyrå (SSB). Med utgangspunkt i programpakken Anzo for Excel og Anzo on the Web, utviklet av Cambridge Semantics, har de brukt enkle ontologier og mange RDF-grafer for å representere statistikk.

Formålet med prosjektet var å utforske nye teknologier og fremgangsmåter for å finne mer effektive måter å distribuere og visualisere innholdet i SSBs statistikkbank. Studentene har produsert en fyldig rapport, og denne forklarer blant annet:

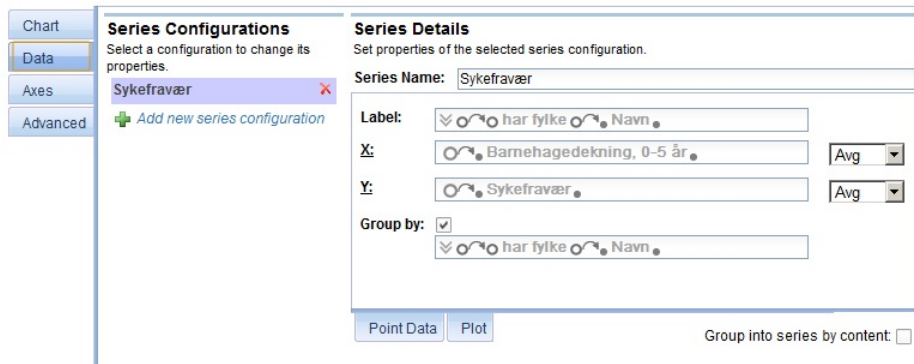
- Hvordan man kan lage ontologier som beskriver forholdet mellom ulike statistikk-variabler.
- Hvordan man kan merke statistikken med vokabular fra ontologien.
- Muligheten for å lagre statistikk som RDF-tripler og hvorfor dette letter oppgaven med å visualisere statistikken i et web-grensesnitt.

Under dette arbeidet har de også utarbeidet en enkel prosedyre for hvordan man kan hente inn, merke og laste opp statistisk materiale fra SSB og andre byråer og undersøkelser. Dette kan igjen benyttes for å lage visuelle representasjoner med en passende GUI-løsning på toppen. Under hele arbeidet har hovedfokuset vært å benytte standardiserte teknologier, da spesielt OWL, RDF og TRIG.

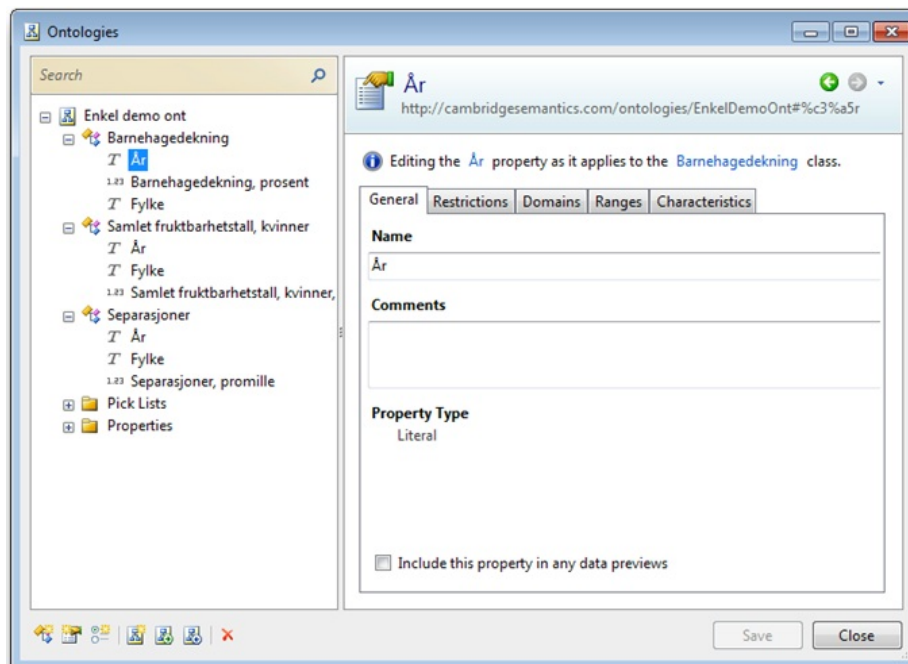
Å benytte standardiserte teknologier, spesielt med tanke på dataformatet, gjør det betydelig lettere å distribuere data til ulike aktører, og benytte data for ulike typer representasjoner. Med verktøy som kan visualisere data som er semantisk lagret, som Anzo on the Web, blir det lett å for eksempel visualisere data for ulike fylker over forskjellige år, mens data i statistikkbanken er lagret på kommune og måned. Man trenger ikke lenger instruere datamaskinen om hva som skal summeres, med litt semantikk og mye data gjøres dette automatisk.

En av fordelene med å ha statistikken representert i RDF-grafer, hvor relasjonene er logisk beskrevet i ontologier, er at sluttbrukeren selv kan bestemme hvilke data som skal sammenlignes. Brukeren kan selv velge hvilket nivå hun ønsker å visualisere dataene på, for eksempel ved å snitte alle kommunene sammen til fylker. Dataene må kun være lastet opp på det

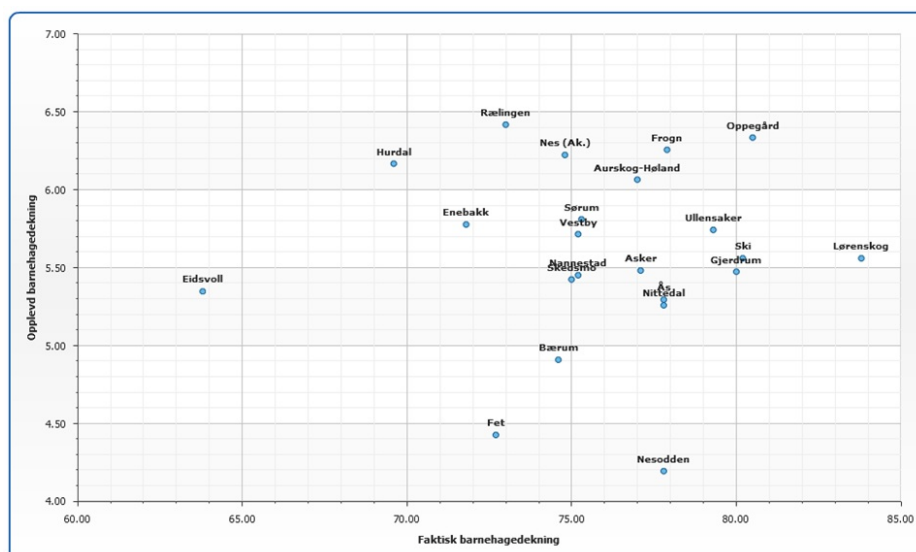
5 Demonstratorer



Figur 6: Grensesnittet til Anzo on the Web gir fullstendig valgfrihet til hvilke statistiske variabler som skal sammenlignes



Figur 7: Ontologi-redigering i Anzo for Excel



Figur 8: Barnehagedekning (SSB) og opplevd barnehagedekning (DIFI)

laveste nivået. Et av problemene som dukket opp underveis er at programvaren, Anzo, kan ikke ta høyde for at dataene er i prosent, og ved å snitte disse prosentdata blir resultatet galt.

De konkrete oppgavene gikk ut på å visualisere statistikk for ulike statistikk hentet fra SSB, men også fra eksterne kilder som DIFIs folkeundersøkelse og data om vannkvalitet hentet fra Folkehelseinstituttet (FHI) for å vise hvordan statistikk hentet fra forskjellige kilder lett kan kobles sammen med et felles vokabular. Data fra SSB gikk på sykefravær, skilsmitter og separasjoner samt barnehagedekning.

Med utgangspunkt i disse dataene, lagret som RDF-tripler og beskrevet med enkle ontologier, fikk studentene med Anzos innebygde visualiseringsverktøy lagd stolpediagrammer, kakediagrammer og spredningsgrafer. Med et solid rammeverk i bunnen, basert på semantiske teknologier, var det ingen begrensninger på hva slags informasjon grafene kunne representere og sammenligne. Alle ulike data kunne hentes inn på samme visualisering, og representeres på forskjellige geografiske og tidsmessige nivåer, ved hjelp av enkle museklikk fra brukerens side.

Vi ser at dette er en helt ny møte for sluttbrukere å tilnærme seg statistikk på. Til nå finner man som regel kun ferdigstilte visualiseringer eller rådata, som legger begrensninger på hva slags informasjon en sluttbruker får tilgang til. Nytteverdien av dette er stor, men må selvsagt veies opp mot muligheten for at uerfarne brukere kan visualisere statistikken på en feilaktig måte.

Konklusjonen er at semantiske teknologier har stort potensial for å gjøre tilgjengelig offentlig statistikk for sluttbrukere. Standardiserte ontologier vil gjøre det lettere å samarbeide om statistikk og metadata på tvers av avdelinger og organisasjoner. Muligheten for å lett merke og laste opp egen statistikk vil gjøre det enklere for blant annet forskere og bedrifter å bedrive forskning, og et slikt verktøy vil også gi større mulighet for aviser og andre medier å fremstille interessant statistikk og kan være med på en demokratiseringsprosess rundt de data som finnes.

6 Oppsummering og veien videre

Ifis viktigste bidrag til Semicolon har vært:

- Utvikling av tre demonstratorer som eksponerer åpne data fra dataeiere som er partnere i prosjektet
- Utvikling av et nytt kurs i semantisk teknologi ved Ifi. Dette kurset er enestående i Norge og er allerede blitt et av de mest populære kurs ved Ifi på bachelor-nivå med ca 40 påmeldte både i 2010 og til det forestående kurset vårsemesteret 2011
- Aktiv deltagelse i prosesser for å øke deltagende etaters forståelse av problemstillingene som ligger til grunn for Semicolon
- Oppbygging av et levende forskningsmiljø i semantisk teknologi ved Ifi med stor kontaktflate både nasjonalt og internasjonalt gjennom engasjement i konferanser (spesielt Semantic Days), nye konsortier (spesielt SFI-søknad) og programkomitearbeid

I vårt arbeid har vi støtt på to utfordringer:

- Vi har erfart at aktivt samarbeid med delagende etater om deres data og infrastruktur er en tidkrevende og omfattende prosess. En betydelig andel av vårt budsjett i prosjektet har blitt brukt på arbeid for å forankre prosjektet og detaljplanlegge aktiviteter.
- Gjennom målrettet innsats i og utenfor Semicolon har vi tilegnet oss relevante state of the art-metoder og teknologi. Vi mener nå vi er det ledende kompetansemiljø i Norge på dette området.

For Ifis del vil aktiviteten i Semicolon videreføres i Verdikt-prosjektet Semicolon II, der vi har blitt tildelt en ressurs på to postdoc-år. Fra vårt arbeid med Semicolon bringer vi flere ressurser inn i Semicolon II:

- Publiserbare resultater om åpne data. Vi regner med at publiseringsaktiviteten vil øke betydelig i løpet av Semicolon II
- Etablert samarbeid med deltagende etater
- Tre PhD-stipendiater (Forssell, Skjæveland, Kjernsmo) finansiert av Ifi med problemstillinger relatert til Semicolon II
- 6 pågående masterprosjekter med prosjekter direkte motivert av Semicolon og med involvering av en eller flere partnere i Semicolon II
- Utdanning av kandidater med kurs i semantisk teknologi (over 100 nye i prosjektperioden til Semicolon II)
- Detaljert kjennskap til forskningsfront
- Et stort internasjonalt nettverk av forskere

Referanser

- [1] Martin Giese, Andrew Ireland, and Laura Kovács, editors. Special issue on invariant generation. *Journal of Symbolic Computation*, 45(11):1097–1250, November 2010.
- [2] Martin Giese and Arild Waaler, editors. *Automated Reasoning with Analytic Tableaux and Related Methods, 18th Intl. Conf., TABLEAUX 2009, Oslo, Norway*, volume 5607 of *LNCS*. Springer, 2009.
- [3] Martin Giese, Arild Waaler, and Franz Baader, editors. Special issue on automated reasoning with analytic tableaux and related methods. *Logical Methods in Computer Science*, 2011. To appear 2011.
- [4] Espen H. Lian, Einar Broch Johnsen, and Arild Waaler. Confluent Term Rewriting for Only-knowing Logics. In Thomas Ågotnes, editor, *Proceedings of the Fifth Starting AI Researchers' Symposium (STAIRS 2010), Lisboa, Portugal, 16-17 August, 2010*, volume 222 of *Frontiers in Artificial Intelligence and Applications*, pages 162–174. IOS Press, 2010.
- [5] Espen H. Lian and Arild Waaler. An only knowing approach to defeasible description logics (extended abstract). In Bernardo Cuenca Grau, Ian Horrocks, Boris Motik, and Ulrike Sattler, editors, *Description Logics*, volume 477 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.

- [6] Audun Stolpe. Norm-system revision: theory and application. *Artif. Intell. Law*, 18(3):247–283, 2010.
- [7] Audun Stolpe. Relevance, derogation and permission. In Guido Governatori and Giovanni Sartor, editors, *DEON*, volume 6181 of *Lecture Notes in Computer Science*, pages 98–115. Springer, 2010.
- [8] Audun Stolpe. A theory of permission based on the notion of derogation. *J. Applied Logic*, 8(1):97–113, 2010.