FFI-RAPPORT

17/00012

Semantikkbasert informasjonsforvaltning for Forsvaret

sluttrapportering av aktiviteten på semantiske teknologier i FFI-prosjekt Informasjons- og integrasjonstjenester i INI

Audun Stolpe
Bjørn Jervell Hansen
Jonas Halvorsen

Semantikkbasert informasjonsforvaltning for Forsvaret

sluttrapportering av aktiviteten på semantiske teknologier i FFI-prosjekt Informasjons- og integrasjonstjenester i INI

Audun Stolpe
Bjørn Jervell Hansen
Jonas Halvorsen

Forsvarets forskningsinstitutt (FFI) 19. desember 2016

Emneord

Informasjonsintegrasjon
Informasjonsinfrastruktur
Semantisk web
Kunstig intelligens

FFI-rapport
FFI-RAPPORT 17/00012

Prosjektnummer
1277

ISBN
P: 978-82-464-2872-7
E: 978-82-464-2873-4

Godkjent av
Bjørn Jervell Hansen, *forskningsleder*
Anders Eggen, a*vdelingssjef*

Sammendrag

Rapporten oppsummerer forskningsaktiviteten på semantiske teknologier i FFI-prosjekt 1277 – Informasjons- og integrasjonstjenester i INI. Delaktivitetens oppgave var å studere hvordan semantiske teknologier kan brukes til å bidra til å løse problemet med å integrere informasjon fra heterogene kilder.

Et overblikk over fagfeltet, grunnleggende konsepter og standarder blir først presentert. Her beskrives semantiske teknologier generelt, før det dykkes ned i teknologier og standarder assosiert med Semantic Web, en gruppe semantiske teknologier som har blitt lingua franca for feltet og som er bygget på web-arkitekturen. Rapporten fortsetter så med å legge frem lovende anvendelsesområder for teknologiene med tanke på bruk i forsvarssammenheng. Anvendelsene skissert er følgende: ontologibasert dataintegrasjon der ontologier fungerer som et abstraksjonslag over heterogene kilder, reaktiv hendelsesdeteksjon over strømmer av data, semantiske teknologier som stordatarammeverk, samt bruk av RDF (en standard under Semantic Web paraplyen) som metadatastrategi.

Deretter skiftes fokus over til å beskrive forskningsproblemene som prosjektet konsentrerte seg om, eksemplifisert med to konkrete militære problemstillinger, henholdsvis planlegging av evakueringsflygninger i en konfliktsituasjon og et konseptuelt informasjonssystem for helse- og hendelsesregister. Fellesnevneren for eksemplene er behovet for å hente inn informasjon fra en rekke heterogene informasjonskilder. Basert på de ovennevnte problemstillingene, ble det å løse problemer relatert til innsamling av relevant informasjon fra heterogene informasjonskilder, under de særskilte behov og begrensninger som det militære domenet fordrer, prosjektets hovedfokus. Dette førte til forskningsresultater som er av teoretisk og prinsipiell natur, men som er appliserbare i en videre forstand både innenfor og utenfor det militære domenet. I tillegg ble det, basert på de teoretiske resultatene, utviklet programvare for praktisk informasjonsintegrasjon. Rapporten dokumenterer resultatene som er oppnådd samt lærdommene som ble trukket av dem.

Til slutt forelegges konkrete anbefalinger med tanke på å tilrettelegge for bruk av teknologien både på kort og lang sikt. Mer konkret er det prosjektets oppfatning at dataorientering ved bruk av RDF har et stort potensial for Forsvaret. For at dette skal kunne være realiserbart, har prosjektet kommet frem til anbefalinger som dekker både strategi og retningslinjer og mer tekniske aspekter. For førstnevnte gruppe, anbefales det å utvikle en strategi for standardisert navngiving av ressurser i en dataorientert setting, samt tilhørende metadatastrategi, for bruk i Forsvaret. De mer tekniske anbefalingene er å utvikle kompetanse på grafdatabaser samt å utvikle datafederering som et alternativ til datavarehus. Grafdatabaser er nyttige teknologier som egner seg for analyse av store datamengder, mens datafederering er svært nyttig som et alternativ til datavarehus der informasjon må hentes utenifra og juridiske barrierer kan forhindre lokal lagring av data.

Summary

This document summarizes the research undertaken and recommendations produced in the FFI-project 1277 regarding semantic technologies. The mandate for the activity was to study how semantic technologies can contribute in solving the problem of integrating information from heterogeneous sources

First, a high-flying overview of the subject area, the basic concepts and related standards are presented. The general field of semantic technologies is introduced before diving into the technologies and standards that are associated with the Semantic Web, a group of semantic technologies built on top of the web-architecture, that have become the lingua franca for the field. The document then proceeds to highlight some promising military application areas for the technologies. More specifically, the areas are: ontology-based data access where ontologies act as an abstraction layer over the heterogeneous sources, event detection over streaming data, semantic technologies as a big-data framework, and finally the use of RDF as a metadata strategy.

The focus of the document then shifts over to describing the research questions that were in focus, exemplified by using two concrete cases from the military domain. More specifically, a case involving planning military evacuation flights in a conflict zone and a case conceptualizing an information system involving accessing medical and event-based data. The commonality between the two cases is the need to collect and integrate information from various heterogeneous sources. Based on the aforementioned, the concrete research activities undertaken focused on solving problems related to federated information gathering from heterogeneous data sources, under the distinctive constraints and requirements that the military domain imposes. This produced research results of a theoretical and principal nature, yet are applicable in a wider context both within and outside the military domain. Furthermore, software was produced that facilitates practical information integration based on the theoretical results. The results and lessons learned are subsequently documented.

Finally, a set of recommendations are produced regarding what activities that should be undertaken in order to make the technology ready for use both in the short and the long term. More concretely, the project believes that a data-oriented approach, using RDF, has great potential for military use. However, in order to ensure that it is realizable, the project has produced a set of recommendations regarding strategy and best-practices, as well as technologies that that warrant further focus. With respect to the former group, it is recommended that an URI strategy for naming resources in a data-oriented setting for use in the Norwegian defence is developed, together with a matching metadata strategy. With respect to technology areas, the project recommends acquiring and developing knowledge regarding graph databases as these are well-suited for analyzing large amounts of data. Another recommendation is to develop federated information gathering and querying as an alternative to traditional data warehouse strategies, which is especially useful in situations where information must be collected from sources outside the home organization and/or judicial barriers prevents local storage of data.

Innhold

1 Innledning

Forsvaret har mange og krevende oppgaver, og er avhengig av å utnytte sine tilgjengelige ressurser maksimalt for å løse disse oppgavene på en god måte. Forsvarssjefen har i sitt Direktiv for operative krav besluttet at dette skal gjøres ved hjelp av fellesoperasjoner der tilgjengelige militære elementer settes sammen til en styrke som er tilstrekkelig for å løse oppdraget. Dette skal gjøres uavhengig av det enkelte elements organisatoriske tilhørighet.

En av de viktigste forutsetningene for å få en slik sammensatt styrke til å fungere etter hensikten, er at beslutningstakerne i organisasjonen har en høy grad av felles situasjonsforståelse. For å bygge opp en slik felles situasjonsforståelse, kreves det utstrakt deling av informasjon i organisasjonen.

Utstrakt informasjonsdeling kommer imidlertid med nye utfordringer. For at beslutningstakere skal kunne nyttiggjøre seg de store informasjonsmengdene som kommer som et resultat av den utstrakte informasjonsdelingen, kreves det at informasjonen integreres og sammenstilles. Denne utfordringen blir forsterket av at det ofte er behov for informasjon fra kilder man ikke selv har eierskap til. Da vil man risikere å være i en situasjon der man verken vet nøyaktig hva slags informasjon kildene kan bidra med eller hva slags format eventuell interessant informasjon vil bli levert på. Eksempler på slike kilder er åpne kilder og kilder kontrollert av militære eller sivile samarbeidspartnere. I tillegg er det en utfordring at viktige kilder som det ikke var planlagt å hente informasjon fra kan dukke opp under en operasjon.

1.1 Formålet med FFI-prosjekt 1277

Formålet med FFI-prosjekt 1277 – Informasjons- og integrasjonstjenester i informasjonsinfrastrukturen, var å støtte Forsvarets arbeid med å utvikle et nettverksbasert forsvar gjennom å utforske teknologiske løsninger som bidrar til utviklingen av en felles informasjonsinfrastruktur (INI).

> Forsvarets informasjonsinfrastruktur (INI): IKT-infrastrukturen som sørger for at alle Forsvarets enheter er i stand til å samhandle og utveksle nødvendig informasjon.

Et av fagområdene i dette prosjektet var semantiske teknologier, en familie av informasjonsteknologier med egenskaper som gjør dem velegnede til å løse integrasjonsutfordringer som beskrevet over. Prosjektet skulle i dette fagområdet fokusere på hvordan nye løsninger og standarder skal kunne bidra til å løse utfordringen med å integrere informasjon fra heterogene kilder slik at informasjonen kan brukes som beslutningsunderlag.

Denne sluttrapporten oppsummerer funnene gjort i faggruppa i løpet av prosjektet, og gir anbefalinger som kan trekkes ut av disse funnene.

2 Bakgrunn

2.1 Semantiske teknologier generelt

Begrepet semantiske teknologier benyttes av mange ulike fagmiljøer og står i dag for en ganske forskjelligartet gruppe forskningsområder. Mange av dem er allerede veletablerte tradisjoner som strekker seg flere tiår tilbake. Noen eksempler på semantiske teknologier er ekspertsystemer, automatisk klassifisering og resonnering, maskinlæring og semantisk søk. Det er dog mulig å si at alle disse teknologiene deler en familielikhet, forsåvidt som alle er designet for å destillere mening fra data i større eller mindre grad vha. kunnskapsrepresentasjon og automatisk resonnering. Kunnskapsrepresentasjon må her forstås som en kombinasjonen av fagfeltene logikk og beregnbarhet: logikk tilbyr et språk og et sett av formelle slutningsregler for å beskrive typer av data og sammenhengen mellom dem, mens beregnbarhetsteori studerer hvilke slutninger det er mulig for en maskin å beregne i rimelig tid. For eksempel:

Ekspertsystemer. Ekspertsystemer bruker sofistikerte resonneringsmodeller for å besvare spørsmål og tilby beslutningsstøtte basert på en representasjon av et eller annet domene, eksempelvis medisinsk diagnostikk. Disse systemene inkluderer ofte maskinlæringsalgoritmer som forbedrer systemets ytelse over tid.

Klassifiseringssystemer. Klassifiseringsteknologier bruker heuristikkregler (tenk på dem som erfaringsbaserte tommelfingerregler) for å merke dataelementer i henhold til kategorier som gjør det lettere å gjennomsøke og analysere store datasett.

Semantisk søk. Idéen bak semantiske søketeknologier er å gjøre det mulig og praktisk å filtrere informasjon i henhold til et eller annet begrep, snarere enn kun i henhold til nøkkelord eller -frase. Med semantisk søk skal maskiner selv kunne å avgjøre hvorvidt det er landet eller skipet Norge en person ønsker informasjon om.

Mange andre teknologer kan også kalles semantiske. De deler ikke nødvendigvis mye mer enn en familielikhet bestående av et logisk fundament. Det er snarere regelen at forskjellige algoritmer er implementert i ulike programmeringsspråk, at forskjellige semantiske systemer aksepterer og produserer data på mange mange ulike og innbyrdes uforenlige formater og at de underliggende logiske beskrivelsesspråkene reflekterer ulik filosofi og vektlegging. Ulike kunnskapsbaserte systemer vil derfor sjelden kunne kombineres uten at det investeres betydelig tid i manuell integrasjon.

2.2 Semantic Web-konseptet

Semantic Web-teknologier er en familie av helt spesifikke teknologistandarder fra *The World Wide Web Consortium* (W3C). De er designet for å beskrive data i et web-miljø på en måte som

gjør tolkningen av dem uavhengig av opprinnelsessystem og fysisk plassering. Disse standardene inkluderer:

- en abstrakt datamodell kalt *The Resource Description Framework* (RDF). RDF kan også forstås som et formelt språk for å representere kunnskap om dataelementer og relasjonene mellom dem,
- ontologispråket *The Web Ontology Language* (OWL) og dets dialekter, en logikk for å beskrive begrepsmodeller,
- et spørrespråk og protokoll kalt SPARQL som definerer søk i web-orienterte databaser, dvs. i databaser som er tilgjengelige over HTTP,
- et reglespråk (RIF) for å uttrykke slutningsregler i et XML-basert utvekslingsformat,
- et språk for å beskrive dataelementer i websider (RDFa),
- og mer.

The Semantic Web:

En utvidelse av den tradisjonelle webben, med fokus på maskinlesbare data. Bygget ovenpå den eksisterende web-arkitekturen, og består av et sett av standarder og formater som skal gjøre det lettere å dele samt gjenbruke data, både av maskiner og mennesker.

Hva gjelder forholdet mellom semantiske teknologier generelt og Semantic Web-teknologier spesielt, så kan man betrakte sistnevnte som en tilpasning av førstnevnte til åpne HTTP-baserte miljøer. Semantic Web-teknologier er i bunn og grunn tradisjonell regelbasert kunstig intelligens sjøsatt i et web-miljø. Som sådan hviler det hele på tre grunnleggende idéer:

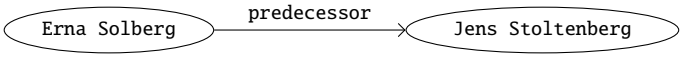
- en standard for utvetydig navning av dataelementer i et omfang som i prinsippet omfatter hele verdensveven,
- databaser som kan nås direkte gjennom web-protokoller, snarere enn å fordre tilkoplingsobjekter i programkode,
- et språk for å uttrykke sammenhenger og overlapp mellom typer av data i forskjellige kilder.

Punktene samsvarer med henholdsvis RDF, SPARQL og OWL. Disse standardene er derfor spesielt sentrale komponenter av Semantic Web-suiten, og vil derfor bli beskrevet litt nærmere nedenfor.

2.2.1 Datarepresentasjon i RDF

Fra en logisk-grammatisk synsvinkel er RDF et svært enkelt kunnskapsrepresentasjonsspråk. Alle uttrykk i RDF er enkle påstandssetninger på formen subjekt-predikat-objekt. Disse uttrykkene kalles bare for tripler i RDF-terminologi (se figur 2.1). RDF-tripler uttrykker således enkle relasjoner mellom objekter. En samling av tripler utgjør tilsammen en RDF-graf. Siden unionen av to RDF-grafer er en graf av samme type, kan en RDF-graf være spredt over flere kilder. Stier fra et objekt til

et annet uttrykker den samlede kunnskapen det er mulig å utlede fra disse kildene om forholdet mellom disse to objektene.



Figur 2.1 *En RDF-trippel*

RDF:

Et enkelt representasjonsspråk for å uttrykke data i form av grafer. Bygget opp av tripler, som er på formen <subjekt, predikat, objekt>. En samling av tripler utgjør en RDF-graf.

RDF er implementert i mange forskjellige filformater, hvilket vil si at den bestemte måten en trippel representeres på varierer fra format til format. Det som er felles for dem alle er hvordan man refererer til objekter og/eller dataelementer på verdensveven og utenfor. RDF-standarden stipulerer at URIer (*Uniform Resource Identifiers*) skal brukes som konstanter, dvs. som navn på ting. Dette er et fundamentalt trekk ved RDF-standarden som er slik fordi det skal være mulig å uttrykke påstander som er utvetydige i et web-vidt omfang. Den semantiske veven er designet for å respektere det såkalte AAA-prinsippet (*Anyone can say Anything about Any topic*), så det er nødvendig med en forsyning av identifikatorer som ikke kan forveksles med hverandre uansett hvor på veven de brukes.

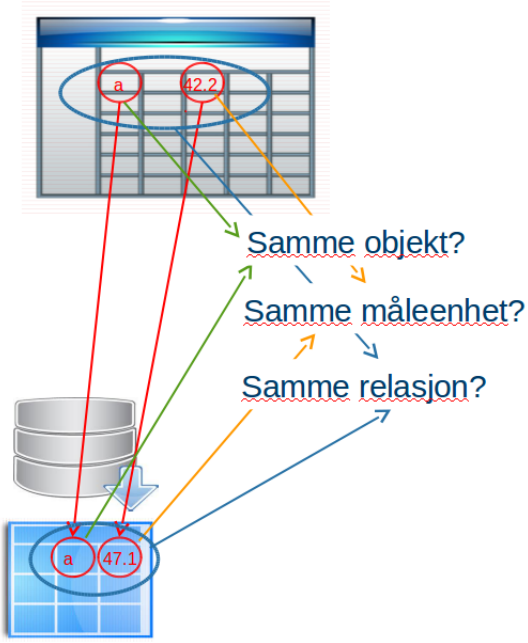
Uniform Resource Indentifier (URI):

En unik identifikator for en ressurs, som tilfredsstiller URI navngivingsskjemaet. Den vanligste formen av gyldige URIer er URLer, bedre kjent som web-adresser.

I RDF-standarden så løses altså dette ved at verdensvevens eget adresseringsmekanisme brukes som basis for navning. Hvert dataelement og hvert objekt "forankres", kan man si, i ett veldefinert punkt innenfor verdensvevens totale navnerom.

URIer er selvsagt designet nettopp for datadeling mellom utvetydig angitte adresser på nettet. I RDF utnyttes dette effektivt til å bryte avhengigheten mellom informasjonsverdien til et dataelement og elementets opprinnelsessystem. Denne avhengigheten kalles gjerne skjemaavhengighet og kan illustreres som i figur 2.2. Problemet kort oppsummert er dette: dersom man ønsker å sammenstille data fra flere uavhengige bakenforliggende lagringssystemer, så er betydningen av navn, relasjoner, måleenheter og annet nødvendigvis bestemt innenfor separate navnerom. Det vil si at det ikke er mulig å fastslå når for eksempel to like navn brukes homonymt eller to ulike brukes synonymt, og ikke mulig å fastslå om relasjoner og innbyrdes forhold mellom dataelementer er forenlige på tvers av kilder.

Når data forankres i URIer, derimot, brytes denne avhengigheten, siden alle URIer tilhører det samme navnerommet. Informasjon som er kodet i RDF vil derfor overflyte applikasjonsbarrierer, og derfor kunne gjenbrukes også til formål som dataene i utgangspunktet ikke ble samlet inn for.



Figur 2.2 *Skjemaavhengighet*

2.2.2 Begrepsmodeller i OWL

Mens RDF uttrykker forhold mellom enkeltobjekter, klatrer OWL (*The Web Ontology Language*) et trinn lenger på abstraksjonsstigen, opp til generelle relasjoner mellom *klasser* av ting. En beskrivelse av generelle forhold mellom klasser er essensielt en logisk teori om vesentlige konseptuelle sammenhenger innenfor et område. Den kalles derfor gjerne en *ontologi* (en teori om det værende) eller, med en muligens mer edruelig terminologi, en *begrepsmodell*.

OWL er et logisk beskrivelsesspråk med en sterk matematisk forankring. Det vil si at det har en formelt spesifisert syntaks og en formelt spesifisert semantikk. Det formelle rammeverket har gjort det mulig å studere forskjellige beregnbarhetsegenskaper nøye. Flere ulike underspråk av OWL er blitt identifisert og skilt ut som egne profiler av standarden. Hver profil er nøye designet for å balanserer uttrykkskraft, dvs. hva det er mulig å si i språket, med kompleksitet, dvs. hva det er mulig å beregne effektivt. Listen under gir en oppsummering:

OWL 2 EL

- spesielt velegnet for begrepsmodeller med et stort antall klasser og egenskaper,
- følgende er effektivt beregnbart:
  - begrepsmodellens konsistens
  - hvilke klasser som sorterer under hvilke andre

- klassetilhørighet for objekter

OWL 2 QL

- spesielt velegnet for ontologibasert grensesnitt over relasjonelle databaser
- uttrykkskraftig nok til å uttrykke UML klassediagrammer og ER diagrammer
- svar på spørringer over begrepsmodellen er effektivt beregnbart

OWL 2 RL

- designet for å uttrykke enkle vilkårssetninger slik som f.eks. Datalog-regler
- implikasjon er effektivt beregnbart

Som oftest brukes en begrepsmodell, uansett OWL-profil, som et abstraksjonslag, eller som en linse over RDF-data. De generelle påstandene i begrepsmodellen uttrykker hvordan data fra ulike kilder forholder seg til hverandre betydningsmessig, og glatter ut semantisk heterogenitet: er denne typen koordinater forenlig med denne typen koordinater? Er disse kodelistene overlappende, eller er felleselementene homonymer? Er denne klassen av kjøretøy pansrede?

Ved at OWL er så nøyaktig designet mhp. beregnbarhetsegenskaper, vil slike påstander beregnes i en automatisert utledningsprosess der nye data implisitt i gamle utledes deduktivt. Man kan derfor si at datasett i RDF ledsaget av en begrepsmodell i OWL er et *selvbeskrivende datasett*: hvert objekt er entydig identifisert i RDF, mens meningen og/eller innebyrden er gitt av begrepsmodellen.

2.2.3 Spørrespråket og protokollen SPARQL

Å søke etter informasjon på den semantiske verdensveven fordrer to ting: et spørrespråk og en standardisert måte å utveksle data på. SPARQL spesifikasjon er designet for å gjøre begge deler. Mer spesifikt definerer SPARQL-spesifikasjonen både et spørrespråk som reflekterer RDF-syntaks, og en protokoll for å utveksle spørringer, svar og RDF-data.

SPARQL:

Et standarisert spørrespråk for RDF-data. Spørrespråket har syntaktiske likheter med SQL, men uttrykker grafmønstre i motsetning til tabellrelasjoner. Evaluering av spørringer invokeres over HTTP i henhold til prinsippene i REST[a].

[a] REpresentatinal State Transfer - Et arkitekturprinsipp for aksessering av web-ressurser

SPARQL som spørrespråk har likheter med SQL forsåvidt som det er designet for å søke i data ved hjelp av å gjenkjenne mønsteret som spørringen uttrykker. SPARQL er svært rikt og uttrykkskraftig. Man kan uttrykke komplekse grafmønstre, vilkår (en. *constraints*) for gyldige svar, valgfrie mønstre og boolske kombinasjoner av dem.

SPARQL-protokollen, på den annen side, definerer et grensesnitt mellom server og klient som tillater klienten å eksekvere spørringer på serveren. Her forutsettes det at serveren implementerer et såkalt SPARQL-endepunkt, som er en HTTP-tilgjengelig RDF-database som samsvarer

med SPARQL-standarden. SPARQL-protokollen er definert som en utvidelse av HTTP, og representerer derfor en utvidelse av verdenveven som er fullstendig forenlig med verdensvevens arkitekturprinsipper forøvrig.

Et SPARQL-endepunkt fremstår på nettet simpelthen som en URI man kan sende spørringer til og som returnerer svar i et avtalt format. Man kan tenke på den som en vevorientert database designet for å utveksle data og spørringer i et HTTP-basert kjøremiljø.

2.2.4 Generelle trekk ved Semantic Web-teknologier

RDF, SPARQL og OWL gjør tilsammen Semantic Web-teknologier ideelt tilpasset dataintegrasjon og -utveksling i et HTTP-basert kjøremiljø. RDF kodifiserer informasjon med bruk av globalt unike URIer som navn, noe som gjør det mulig å uttrykke faktakunnskap utvetydig i et verdensvevsomfang.

SPARQL-endepunkter gjør databaser tilgjengelige direkte over HTTP, mens OWL tilbyr et språk for å beskrive konseptuelle sammenhenger mellom ulike datasett. Sammenhenger som kan beregnes maskinelt.

Kombinasjonen av RDF, SPARQL og OWL gir derfor verdensveven noen av de samme trekkene som karakteriserer tradisjonelle kunnskapsbasert systemer, slik som f. eks. ekspertsystemer. Dette inkluderer:

Deklarativitet: Informasjonsmodellering er basert på et deklarativt beskrivelsesspråk, hvilket innebærer at data og datatyper beskrives på et abstraksjonsnivå som konsentrerer seg om dataelementenes konseptuelle betydning snarere enn om formater eller lagringsdetaljer.

Logisk fundament: Standardteknikker fra formallogikken tilbyr presise metrikker for databehandling. Det kan dreie seg om hvorvidt et datasett er konsistent, om hvorvidt en spørring returnerer korrekte og komplette svar, osv.

Informasjonsbehandling er resonnering: Analyse og behandling av datasett foregår som regel som automatisert resonnering eller deduksjon. Naturlige sammenlikninger her er programmeringsspråk som Prolog eller Answer Set Programming. Den viktigste fordelene ved dette er at implisitte fakta kan utledes fra de som er eksplisitt gitt. Dette gjør det typisk mulig å oppdage sammenhenger mellom fakta og derfor å trenge dypere ned i dataene enn det som ellers ville vært mulig.

Takler endrende informasjonsbehov: Spørsmålene som kan stilles er ikke begrenset til predefinerte mønstre slik som tradisjonelle API- og protokollbaserte systemer er. Dette er et resultat av den deklarative tilnærmingen fremfor en proseduralsk oppbygging.

Dataorientering: Semantic Web-teknologier er designet for å gjøre data minst mulig applikasjonsavhengige og mest mulig selvbeskrivende. Tanken er er at informasjon skal overflyte applikasjonsbarrierer og være forståelig for enhver programvare som er i stand til å forstå RDF og OWL. Med andre ord, mer av logikken flyttes over i datasettene; dataene blir rikere, applikasjonene mer generiske.

Komprimert i en enkelt setning, kan man si at Semantic Web teknologier dreier seg om å støtte intelligent, innholdsorientert informasjonbehandling på tvers av programvarebarrierer og kilder ved hjelp av globalt entydige navn og referanser samt et språk for å beskrive logiske forhold mellom dem.

2.3 Om forutsetninger for å ta i bruk Semantic Web-teknologier

Det er ikke rimelig å forvente at offentlige forvaltningsorganer skal erstatte sine nåværende lagringsløsninger — det være seg relasjonelle databaser, regneark eller liknende — med RDF-databaser. De utfordringene en offentlig institusjon står overfor når det gjelder innsamling, forvaltning og spredning av informasjon er gjerne spesifikke for den institusjonen. Dette gjør at de fleste offentlige organer utvikler sin egen distinkte byråkratiske kultur som som oftest er tett sammenvevd med valg av programvaresuiter og lagringsløsninger, både driftsmessig og økonomisk såvel som teknologisk og kompetansemessig.

RDF kan betraktes som en abstraksjon over tabulære data; både regneark og databasetabeller kan enkelt oversettes til RDF. Dette har gitt opphav til RDF-*eksponering*, som tilbyr RDF-basert aksess til innholdet i databaser og regneark m.m. uten at dataene repliseres i en selvstendig RDF-database.

Forskjellige eksponeringsteknologier er basert på forskjellige typer regelspråk som lar en beskrive forholdet mellom et RDF-vokabular (dvs. et sett med URIer som beskriver objekter av interesse) og f. eks. et databaseskjema. Et slikt regelsett brukes for å oversette SPARQL-spørringer til spørringer over det underliggende grensesnittet, f. eks. SQL. Det brukes mao. til å få databasen til å oppføre seg som om den skulle være en RDF-database. Det finnes flere modne og velprøvde eksponeringsteknologier på markedet, slik som:

- D2RQ: Et system for å manipulere databaser som RDF-grafer. Med D2RQ kan man:
  - eksekvere SPARQL-spørringer mot konvensjonelle databaser,
  - eksponere innholdet i databaser som lenkede data på verdensveven og
  - dumpe innhold til rene RDF-databaser.
- XLWrap: XLWrap gjør noenlunde det samme for Excel regneark og kommaseparerte lister som D2RQ gjør for databaser.
- BootOx: Er en eksponeringsteknologi som lar deg benytte OWL begrepsmodeller som eksponeringsgrensesnitt. BootOX er basert på en tilnærming til eksponering som av W3C er standardisert under navnet *Direct Mapping*: hver tabell i databasen koples til en OWL-klasse i begrepsmodellen, hver attributt i en tabell til et RDF-predikat etc.

RDF-eksponering gjør terskelen for å ta i bruk Semantic Web-teknologier lav. Det er ikke nødvendig å erstatte eksisterende lagringsløsninger. RDF-data kan leve side om side med mer tradisjonelle databaser og regneark uten konflikt. Eller mer presist; gjennom eksponering vil RDF-dataene til enhver tid reflektere dataene de eksponerer og endre seg i takt med dem. Det er derfor ikke nødvendig med separat vedlikehold av RDF-dataene etter at oversettelsen mellom RDF og tabulære formater er definert. Mange eksponeringsteknologier vil kunne generere en slik oversettelse automatisk, så den totale kostnaden ved å ta i bruk Semantic Web-teknologier må sies å være svært lav.

3 Anvendelsesområder

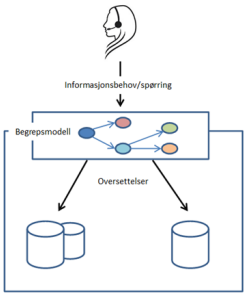
Denne seksjonen skisserer noen typiske anvendelser av Semantic Web-teknologier basert på listen over karakteristiske egenskaper fra avsnitt 2.2.4. Den er ikke ment å være uttømmende, men består av et utvalg eksempler som vi mener best illustrerer den potensielle nytteverdien for Forsvaret.

En spesielt stor utelatelse er forskningen omkring såkalte lenkede åpne data (Linked Open Data/LOD). Denne forskningen konsentrerer seg om å utvikle en beste praksis for å publisere gjenbrukbare, fritt tilgjengelige RDF-data under åpen lisens. Selv om dette miljøet har bygd opp betydelig momentum de seneste årene, har vi vurdert spesielt 'O'en i LOD som perifer i forhold til Forsvarets behov. Vil vil derfor ikke gå nærmere inn på *data openness*, og alt det fordrer, i denne rapporten.

3.1 Ontologibasert dataintegrasjon

Ontologibasert dataintegrasjon (eller *Ontology Based Data Access* – OBDA – som det gjerne kalles) handler om å forene informasjon med overlappende betydning og relevans under et felles spørregrensesnitt. Det vil som regel involvere flere kilder, typisk, men ikke nødvendigvis, relasjonelle databaser. Det er ikke en forutsetning at disse er utviklet eller vedlikeholdt med tanke på hverandre.

I et OBDA-system er integrasjonen kun virtuell og derfor løst koblet til de underliggende datakildene. Kjernen i et OBDA-system består av en begrepsmodell (ontologi) som kan betraktes som en beregnbar spesifikasjon av hvordan typer av data i de underliggende kildene forholder seg til hverandre konseptuelt. Denne spesifikasjonen fungerer som et abstraksjonslag som legges over de underliggende kildene for å presentere dem for brukeren som om de skulle være én kilde (se figur 3.1).



Figur 3.1 *Ontologibasert dataintegrasjon.*

Begrepsmodellen designes gjerne slik at den reflekterer brukerens foretrukne vokabular, noe som tillater en analytiker å uttrykke sitt informasjonsbehov med begreper som reflekterer hans eller hennes kompetanse. Selve datainnsamlingen foregår ved at analytikeren formulerer sitt informasjonsbehov i SPARQL der spørremønsteret defineres ved bruk av vokabular fra eksterne begrepsmodeller. Siden begrepsmodellen uttrykker forhold mellom typer av data i de underliggende kildene, gjør dette i sin tur det mulig å beregne hvilken informasjon som må hentes fra hvilke kilder og hvordan den må kombineres for å svare på analytikerens informasjonsbehov. Denne oversettelsesprosessen gjør at et OBDA-system relativt enkelt kan tilpasses et vilkårlig antall kilder uten at kompleksiteten øker i brukerens øyne.

Ontologibasert dataintegrasjon er per i dag et forholdsvis velstudert forskningsområde der resultatene har begynt å finne sin vei inn i programvarebiblioteker og kommersielle produkter. Det

finnes også grafiske utviklingsmiljøer for å jobbe strukturert med integrasjon på tvers av kilder og for å analysere resultatet. Noen eksempler er:

- Anzo Enterprise: Programvare for dataintegrasjon, søk, analyse og visualisering. Kommer med en egen IDE.
- Stardog: et dataintegrasjonsplatform bygget på smart grafteknologi; spørringer, søk, automatisk resonnering og datavisualisering. Tilbyr et visuelt webgrensesnitt for utvikling.
- Ontop: En plattform for å søke i relasjonelle databaser vha. begrepsmodeller og SPARQL. Bruker Protegé som utviklingsmiljø.

Ontology Based Data Access (OBDA)

En teknologi for å søke i data ved hjelp av begrepene i en ontologi. Skjema og formater i de underliggende kildene, samt heterogenitet mellom de respektive kilden, blir abstrahert bort fra brukeren som forholder seg til ontologien som grensesnitt.

3.2 Hendelsesdeteksjon

For at det skal tjene noen hensikt å sammenstille data i stor skala, må det være mulig å reagere raskt nok og intelligent nok på informasjonen som presenteres. Dette er en stor utfordring for Forsvaret siden data i stor grad er ferskvare som leveres i høy hastighet i form av strømmer. Eksempler er AIS-systemer, blåprikksystemer og sanntids hendelseslogger.

Hendelsesdeteksjon (en. *Event Recognition*) er derfor et viktig område som faller naturlig inn i forlengelsen av dataintegrasjon. Som navnet tilsier, dreier det seg om å gjenkjenne mønstre som indikerer at en type hendelse har funnet sted i en datastrøm under overvåkning. Hendelsesdeteksjon er grunnlaget for reaktive systemer som selv kan respondere på trusler på adekvat vis. Eksempler omfatter angrep i computernettverk, avviksdeteksjon i maritime overvåking, fremvoksende trender i sosiale medier, o.a.

Hendelsesorientert databehandling er allerede et veletablert beregningsparadigme innenfor applikasjonsområder som spenner fra finans og medisin til logistikk og analyse av nettverkstrafikk, og er uavhengig av Semantic Web-visjonen som sådan. I senere tid har det imidlertid blitt svært vanlig å bruke Semantic Web-teknologier for å forklare konseptuelle sammenhenger i dataene.

Sosiale mediestrømmer er et spesielt velstudert eksempel. Overvåkning av sosiale mediestrømmer dreier seg, på individnivå, om slike ting som å fastslå brukeres digitale identitet og aktiviteter. På gruppenivå kan det dreie seg om demografi, interesser og kollektiv adferd. Det finnes allerede flere OWL-derivater som er spesialdesignet for å beskrive on-line samfunn og sammenhengen mellom dem. Et utvalg av disse er:

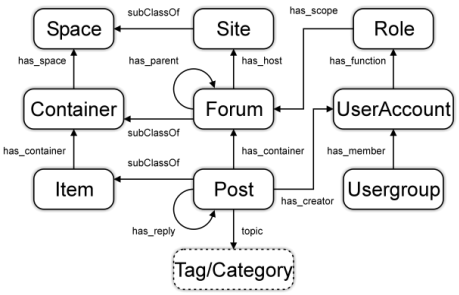
SIOC (akronym for *Semantically-Interlinked Online Communities*): et forsøk på å bruke Semantic Web til å beskrive strukturen i et on-line-samfunn; hvilke typer innhold som finnes, og hvilke

relasjoner som eksisterer mellom innholdsobjekter og andre typer samfunnsobjekter slik som grupper, fora, etc. Et eksempel fra SIOC er vist i figur 3.2.

FOAF (akronym for *Friend of a Friend*): en ontologi for å beskrive personer, deres aktiviteter og deres relasjoner til hverandre.

GUMO (akronym for *The General User Model)*): en ontologi som inneholder et sett av modeller for brukere av on-line-samfunn. GUMO klassifiserer og samler forskjellige dimensjoner av informasjon som kan knyttes til en digital identitet. Det kan være biologiske data, slik som puls, kjønn og alder, det kan være sosiale data slik som yrke, eller det kan være data om interesser og evner, slik som svømmeferdigheter og musikksmak.

DPLO (akronym for *Digital.Me LivePost Ontology*): en ontologi som forsøker å representere tidstemplet, dynamisk personlig informasjon. Eksempler er oppfatninger, kommentarer og tilstedeværelse i sosiale nettverk og andre on-line fora.



Figur 3.2 *SIOC konsepter.*

Automatisert hendelsesgjenkjenning blir viktigere og viktigere i det vi beveger oss inn en tidsalder hvor det ikke bare er tilgjengeligheten av informasjon, men også vår evne til å analysere den og handle på bakgrunn av den som gir det vesentlige konkurransefortrinnet.

3.3 Semantiske teknologier som stordatarammeverk

For å få full effekt av selvbeskrivende data er det nødvendig å kunne sammenlikne, kombinere og analysere data fra forskjellige kilder, kilder som potensielt er utviklet for forskjellige formål av forskjellige dataeiere. Det kan være nye sensorer, data fra eksisterende sensorer, eller informasjon fra sivile kilder. *Open Source Intelligence* er et eksempel på en type informasjon som nyttiggjør seg informasjon fra sivile kilder.

Med en stadig økende grad av digitalisering blir mengden tilgjengelige data raskt uhåndterlig for militære beslutningstakere og analytikere. Generelt kalles dette fenomenet gjerne *Big Data*, heretter *stordata*. En vanlig måte å oppsummere hva stordata dreier seg om, er ved å henvise til de tre v'ene (i engelsk nomenklatur):

- *Volume*: mengden av data er så stor at tradisjonelle lagringsteknologier og analyseverktøy ikke strekker til.
- *Variety*: typen av data varierer, både gjennom at det kan være tekst, bilde, lyd osv, og gjennom at de er kodet i forskjellige formater og i henhold til forskjellige standarder.
- *Velocity*: dataene strømmer i sann, eller nær sann, tid.

Semantic Web-teknologier har egenskaper som passer disse utfordringene godt. Problemet med variasjon kan langt på vei løses med RDF og denne standardens bruk av URIer som unike navn for dataelementer (jamfør kapittel 2.2). I dette scenariet vil grunnlagsdataene forme en stor sammenhengende graf som kan omfatte et vilkårlig antall fysiske separate kilder.

Stordata:

Data av *forskjelligartet natur*, som kommer i *store mengder* og har *hyppig oppdateringsfrekvens* (Volume, Variety, Velocity).

Hva gjelder volum, så har det vært vanlig å benytte distribuerte RDF-databaser bygget over konvensjonelle stordatarammeverk slik som Hadoop MapReduce. Disse systemene bruker det underliggende MapReduce APIet for å eksekvere spørringer og for å koordinere mellomresultater på tvers av en klynge av servernoder der de faktiske dataene ligger lagret. Imidlertid har flere uavhengige studier vist at dette gir betydelig ytelsesforskjeller avhengig av hvor ofte data fra forskjellige kilder må kombineres for å sammensette et endelig svar. MapReduce er typisk svært effektivt når mellomresultater er små og kan behandles parallelt, dvs. uavhengig av hverandre. Når mellomresultatene er store og avhenger av hverandre, derimot, så kan kostnadene ved MapReduce-modellene, og koordineringen den fordrer, være uoverkommelig høye.

I den senere tiden har dette ført til fremveksten av distribuerte filsystemer som forsøker å utnytte RDFs grafmodell mer direkte. Dette er grafdatabaser som er spesielt optimalisert for traversering og søk. Fordelene med et stordatarammeverk som er bygget på RDF fra bunnen og opp er mange. Her er noen:

- Mer komplekse spørringer innenfor akseptable ytelsesrammer: Grafer er en naturlig datastruktur for å beskrive sammenhengen mellom objekter innenfor et hvilket som helst domene, noe databaseforskningen selvsagt har vært klar over i flere tiår allerede. Konvensjonelle stordatarammeverk kommer derfor ofte med tilleggsbiblioteker for grafanalyse. Et eksempel er GraphX. GraphX er en abstraksjon over såkalte *Resilient Distributed Datasets* som utgjør lagringsmodellen i bl. a. Hadoop og Spark. Ytelsesmessig er den dog begrenset av den underliggende batch-orientert eksekveringsmodellen til det underliggende stordatarammeverket: den gir typisk akseptabel ytelse når mellomresultater er løst sammenkoblet. Et stordatarammeverk som er bygget på RDF fra bunnen og opp, er derimot optimalisert for grafanalyse fra starten av. Dette gjør at mer komplekse spørringer kan uttrykkes i RDFs eget spørrespråk, og at de (med en tilbørlig vaktsom utforming) returnerer svært hurtig.
- Dypere søk gjennom resonnering: Som nevnt i Seksjon 2.2.4 er RDF en deklarativ, logikkbasert datamodell som støtter begrepsmodeller og automatisert resonnering. Flere RDF-baserte stordatarammeverk implementerer støtte for ontologispråk, typisk OWL 2 RL (som igjen er en variant av Datalog). Et eksempel er RDFox som støtter datalog-resonnering over RDF-data ved hjelp av avanserte evalueringsalgoritmer. Denne resonneringen fungerer som en slags forsterker for SPARQL-spørringer. Spørringene formulerer et informasjonsbehov ved hjelp av abstrakte, domenenære konsepter, slik de er karakterisert i den datalogbaserte begrepsmodellen. Resonneringmotoren håndterer omskrivingen av disse spørringene til en form som gjør at de kan evalueres direkte mot de ulike datakildene uten at brukeren trenger å forholde seg til detaljer som har å gjøre med formater og representasjon.
- Gjenbrukbare data: Som nevnt i kapittel 2.2.1 er RDF designet for å bryte avhengigheten mellom informasjonsverdien til et dataelement og elementets opprinnelsessystem, dvs. for å redusere generell skjemaavhengighet. RDF-data er selvbeskrivende, overflyter applikasjons-

barrierer, og kan gjenbrukes utenfor sin opprinnelige kontekst.

Det finnes allerede flere kommersielle stordatarammeverk for RDF på markedet. To eksempler er Blazegraph og Stardog. Begge er høyytelses grafdatabaser som er optimalisert for RDF og SPARQL.

Markedsledende Blazegraph benytter en Multi-GPU-teknologi som er i stand til å analysere grafer med opp til 50 milliarder tripler pr. maskin i nær sanntid. Blazegraphs GPU-akselerasjon er flere hundre ganger raskere enn sammenliknbare CPU-teknologier, og størrelsesordener raskere enn konvensjonelle stordatateknologier basert på Hadoop slik som HBase, Titan og Accumulo.

3.4 RDF som metadatastrategi

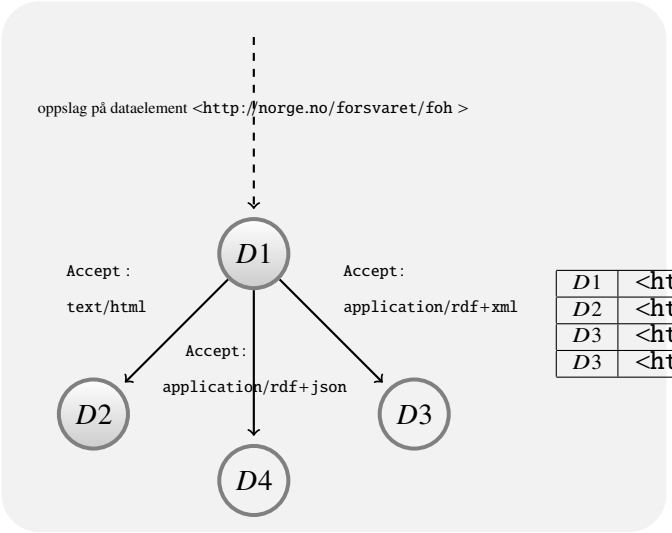
I offentlig forvaltning der ulike organer skal samhandle, er metadata av avgjørende betydning. Behovet for gode løsninger på dette området er presserende også pga. arkiveringsplikten som alle offentlige organer er underlagt. Metadata dokumenterer slike ting som opphavet til et dataelement, hvilke juridiske restriksjoner som hefter ved et datasett, eierskap, graderingsnivåer og ideelt sett all annen type informasjon som dataelementets livssyklus fordrer. Det er vanlig å dele metadata i tre typer: strukturelle, deskriptive og administrative metadata. Vi regner en mer detaljert beskrivelse av disse som utenfor skopet for denne rapporten.

Det er en forutsetning for at et dataelement skal kunne gjenbrukes på tvers av systemer og til ulike formål at disse metadataene ikke erodere når informasjon vandrer fra ett register til et annet. Hvilket register kommer informasjonen opprinnelig fra? Hvordan var den merket og beskrevet der? Hvilken type gjenbruk er egentlig tillatt? Osv.

Et eksempel fra Forsvaret er saniteten og helsedata. Forsvarets systemer for medisinske opplysninger består av flere registre, eksempelvis SANDOK, TANDOK og Forsvarets Helseregister (FHR). Et informasjonssystem med både hendelses- og helsedata for personell i Forsvaret vil gi kunnskap om sykdom, skade og andre helseproblemer samt hvilke forhold som er årsaken til problemene. Sivile registre som Reseptregisteret, Norsk pasientregister og KUHR vil også kunne være aktuelle. Det er selvsagt klare juridiske vilkår for å bruke informasjonen i disse registrene. All behandling av helsedata er strengt regulert i lover og forskrifter som ivaretar den enkeltes personvern. Det er derfor avgjørende viktig, dersom man skal kunne se for seg at denne informasjonen skal kunne gjenbrukes og sammenstilles, at det finnes gode mekanismer for forvaltning av metadata.

Dette er selvsagt et problem som ikke har en rent teknologisk løsning. Man må forutsette metadatastrategier som også omfatter organisasjons- og regelverksutvikling. Med RDF er det imidlertid mulig å tilby en teknologisk løsning for å knytte et sett av metadata, når det engang er bestemt, til et dataelement på en måte som som gjør at informasjon og metainformasjon ledsager hverandre gjennom hele livsløpet til dataelementet: som nevnt tidligere bruker RDF URIer for å identifisere dataelementer og objekter på en utvetydig måte. En opplagt idé som etterhvert har blitt standard praksis er å dokumentere et dataelement ved å sørge for at det er mulig å gjøre oppslag på URIen som navngir det.

Oppslag må her forstås vidt. Verdensveven er ment som et informasjonsrom for både menneskelig og maskinell utnyttelse. Det bør følgelig være mulig å gjøre oppslag som tilpasser innholdet etter



$D1$	<http://norge.no/forsvaret/foh>
$D2$	<http://norge.no/forsvaret/foh.html>
$D3$	<http://norge.no/forsvaret/foh.rdf>
$D3$	<http://norge.no/forsvaret/foh.json>

Figur 3.3 *Innholdsforhandling*

den som spør, typisk HTML for mennesker og RDF/XML for maskiner. Det finnes støtte for dette i en mekanisme i HTTP-protokollen som kalles *innholdsforhandling* (*content negotiation*). Innholdsforhandling går ut på å bruke HTTPs responskode **303 See Other** for å knytte ulike beskrivelser til en URI, , se figur 3.3. Klienten velger mellom disse beskrivelsene ved hjelp av `Accept`-feltet i HTTP-headeren. Dersom `Accept`-feltet indikerer at klienten foretrekker HTML, så vil serveren respondere ved å sende et HTML-dokument. Dersom klienten foretrekker RDF, så vil serveren sende eksempelvis en JSON- eller XML-representasjon av RDF avhengig av hva klienten ber om.

I RDF utnyttes innholdsforhandling til å knytte dokumentasjon og metadata til et navngitt dataelement og på den måte sørge for at metadata flyttes "oppstrøms" til dataelementets opprinnelse. Ved å sørge for at URIer (*Uniform Resource Identifier*) er URLer (*Uniform Resource Locators*) som kan slås opp, kan beskrivelsene som knyttes til en URL vha. en 303-henvisning inneholde standardiserte maskinleselige metadata om elementet. De senere årene har det blitt investert mye i slike standarder. Noen av de mer velkjente initiativene på dette området er:

- Dublin Core: Er et lite RDF-vokabular for å beskrive åndsverk generelt. Beskrivelsen kan angi slike ting som rettighetshaver og utgiver, men også informasjon om åndsverkets struktur og format.
- VoID (*Vocabulary of Interlinked Datasets*): Inkorporerer og utvider Dublin Core-vokabularet, med begreper som er nødvendige for å beskrive hele datasett; deres tema og innhold, oppdateringsrate, bidragsytere, skjemastruktur, o.a.
- PROV Ontology: Er en begrepsmodell uttrykt i OWL. Den tilbyr et sett av klasser, egenskaper og begrensninger som kan benyttes for å representere og dele proveniensinformasjon generert av forskjellige systemer i forskjellige kontekster.

W3C forfekter disse standardene som beste praksis for metadata i RDF.

Innholdsforhandling:

Innholdsforhandling i RDF er en måte å knytte flere dokumenter til én og samme URL. URLen fungerer som en identifikator for et eller annet objekt eller dataelement, mens de assosierte dokumentene inneholder metadata. Noen av dokumentene vil bestå av maskinleselige versjoner av metadataene, mens andre vil bestå av HTML eller tekst.

4 Forskningsfokus

Semantikkarbeidet i prosjekt 1277 har fokusert på ontologibasert dataintegrasjon som en interessant og potensielt svært anvendelig tilnærming til dataintegrasjon i Forsvaret. Spesielt har vi vektlagt

- kompatibilitet med legacy-systemer: ontologibasert dataintegrasjon kan gjøres uten å erstatte nåværende lagringsløsninger, jamfør kapittel 2.3,
- evne til å utnytte betydningssammenhenger i dataene: utveksling av informasjon mellom militære systemer vanskeliggjøres ofte av forskjellige typer semantisk heterogenitet. Det kan for eksempel dreie seg om forskjellige standarder for geokoding eller forskjellige kodelister for hendelser. Med begrepsmodeller kan disse betydningssammenhengene uttrykkes og automatiseres, jamfør kapittel 2.2.2 og
- sanntidighet: Ved bruk av RDF-eksponeringsteknologier kan man unngå datavarehusløsninger som må vedlikeholdes og oppdateres separat. RDF-eksponering gjør at integrasjonssystemet til enhver tid reflektere dataene slik de faktisk foreligger i kildene og endrer seg i takt med dem, jamfør 2.3. Dette er en interessant egenskap mhp. situasjonsbevissthet.

Vi ser for oss at ontologibasert dataintegrasjon kan anvendes i to generelle typer scenarier i Forsvaret:

1. I scenarier der nettverkstopologien må forutsettes å være dynamisk, dvs. at utvalget av datakilder ikke er konstant. Dette vil f. eks. være typisk for systemer som produserer situasjonsbilder.
2. I scenarier der nettverkstopologien kan forutsettes å være noenlunde statisk. Dette vil være typisk for fagsystemer som er bygget over det som generelt kan kalles arkiver, f. eks. utstyrsregistre eller helseregistre.

Disse to typene anvendelser stiller ulike krav til et ontologibasert dataintegrasjonssystem. Grovt kan man sin at dynamiske nettverkstopologier fordrer lettere begrepsmodeller med lavere kompleksitetsprofil og mindre uttrykkskraft. Mer av resonnering må gjøres kontinuerlig og nettverket må hele tiden skannes for aktuelle kilder.

4.1 To eksempler

Vil vil ikke gå nærmere inn på forskjellen mellom statisk og dynamiske nettverkstopologier mhp. dataintegrasjon i denne rapporten. Vi nøyer oss med å gi ett litt lenger eksempel på hver av dem, som er ment å antyde deres potensielle nytteverdi for Forsvaret.

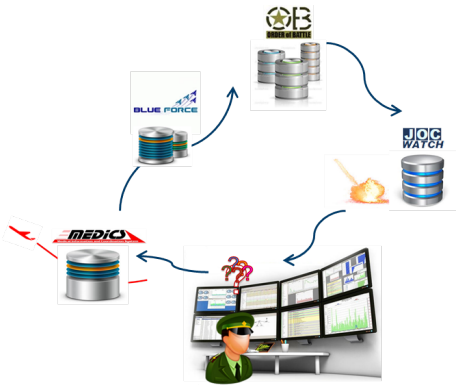
4.1.1 Eksempel 1: Planlegging av evakueringsflygninger

I dette scenariet forestiller vi oss en militær analytiker i et operativt hovedkvarter som har som oppgave å planlegge og overvåke evakueringsflygninger inn i et stridsområde. Det er spesielt viktig

å holde et våkent øye med evakueringsflygninger som er truet av fiendtlig aktivitet. Dersom en flygning er truet, er det analytikerens oppgave å lete etter vennlige styrker som er i stand til å nøytralisere trusselen i nærheten av landingsområdet.

Analytikerens informasjonsbehov kan uttrykkes slik: finn alle evakueringsflygninger som er slik at flygningen kan klassifiseres som truet, og den vennlige enheten som har kapabiliteter til å bekjempe trusselen det er snakk om.

For å besvare dette informasjonsbehovet vil analytikeren vanligvis måtte kombinere informasjon fra flere ulike systemer, som vi normalt vil finne i et operativt hovedkvarter. Analytikeren vil kanskje måtte konsultere en hendelseslogg for å følge utviklingen i stridsområdet og et sporingssystem for evakueringsflygninger for å avgjøre hvilke landingssoner som er truet. I tillegg vil han naturlig nok ha behov for et "blåprikksystem" som viser hvor de vennlige styrkene står, og muligens også en Order of Battle-logistikkdatabase for opplysninger om de vennlige og fiendtlige stridsenhetene. Jamfør figur 4.1.



Figur 4.1 *Dataintegrasjon for evakueringsflygninger.*

Å kombinere alle disse opplysningene manuelt er en tidkrevende og skjør prosess som fordrer at analytikeren kjenner de ulike informasjonssystemene og deres vanligvis ulike datamodeller og spørregrensesnitt.

Ontologibasert dataintegrasjon tilbyr en dynamisk og fleksibel løsning på analytikerens problem med å sammenstille informasjon. I et OBDA-system er integrasjonen kun virtuell og derfor løst koblet til de underliggende datakildene. Dette betyr blant annet at de underliggende datakildene ikke behøver å være utviklet eller vedlikeholdt for å utveksle informasjon med hverandre

Som nevnt i kapittel 3.1 består kjernen i et OBDA-system av en begrepsmodell som kan betraktes som en beregnbar spesifikasjon av hvordan typer av data i de underliggende kildene forholder seg til hverandre konseptuelt. Denne spesifikasjonen fungerer som et abstraksjonslag som legges over de underliggende kildene for å presentere dem for analytikeren i vårt tenkte tilfelle, som om de skulle være én kilde. Begrepsmodellen designes gjerne slik at den reflekterer brukerens foretrukne vokabular, noe som tillater analytikeren å uttrykke sitt informasjonsbehov med begreper som reflekterer hans kompetanse.

Selve datainnsamlingen foregår ved at analytikeren formulerer sitt informasjonsbehov med konsepter og relasjoner som er definert i begrepsmodellen. Siden begrepsmodellen uttrykker forhold mellom typer av data i de underliggende kildene, gjør dette i sin tur det mulig å beregne hvilken informasjon som må hentes fra hvilke kilder og hvordan den må kombineres for å svare på analytikerens informasjonsbehov. Denne oversettelsesprosessen gjør at et OBDA-system relativt enkelt kan tilpasses et vilkårlig antall kilder uten at kompleksiteten øker i brukerens øyne.

4.1.2 Eksempel 2: Et informasjonssystem for et helse- og hendelsesregister

Norske soldater som tjenestegjør i utlandet kan bli utsatt for både fysiske og psykiske skader. Personskader som følge av stridshandlinger forekommer, i ytterste konsekvens tap av liv. For å redusere risikoen for skader, minimere tap av liv, og gi skadde den oppfølging de trenger, er det viktig at Forsvaret har oversikt over skader. Det er viktig at det dokumenteres hvem som utsettes for en hendelse og hvordan situasjonen oppsto. På denne måten vil man kunne jobbe preventivt med skader, man vil kunne følge opp soldater over tid og finne årsakssammenhenger, gjøre epidemiologiske studier, finne sannsynlighet for senskader og følge veteraner inn i alderdommen.

Selv om Forsvaret har gode separate systemer for registrering av helsedata og HMS/hendelsesdata, er det klare juridiske barrierer i veien for å samle disse i et enhetlig medisinsk informasjonssystem. All behandling av helsedata er strengt regulert i lover og forskrifter som ivaretar den enkeltes personvern, og den samlede effekten av disse forskriftene utelukker ethvert tenkelig informasjonssystem basert på et sentralt datavarehus. Denne konklusjonen forblir høyst sannsynlig stående enten man snakker om mikrodata om navngitte personer eller man snakker om ferdigaggregerte forløpsdata og statistikk. Konsekvensene av dette for etableringen av et informasjonssystem for et helse- og hendelsesregister er klare: Enten må det søkes dispensasjon i form av lov eller forskrift, eller så må man forutsette at dataene blir værende der de er under et noenlunde likt forvaltningsregime som i dag.

Det følger at tiltak 52 i regjeringens handlingsplan; “Forsvaret skal (. . .) etablere systemer som kan generere data av god kvalitet til bruk for helsemessig oppfølging, dokumentasjon, forebyggende HMS, statistikk, oversikt og forskning” bør dreie seg om å øke dataflyten på tvers av registre ved å øke gjenbrukbarhetseffekten av dataene der de ligger. En kombinert bruk av

- innholdsforhandling som metadatastrategi (jamfør kap. 3.4),
- RDF-eksponering av fagsystemene (jamfør kap. 2.3) og
- begrepsmodeller for integrasjon (jamfør kap. 3.1).

bør kunne møte mange av disse behovene, og være et interessant alternativ. Pga. at utvalget av registre som skal samordnes kan forutsettes å forbli det samme (statisk nettverkstopologi), er det mulig å se for seg å definere sammenstillingsprosesser som gitt dagens forvaltningsregime og forskrifter er juridisk skalerbare.

4.1.3 Forskjellige begrepsmodeller til forskjellig bruk.

Det er muligens naturlig å tenke at så lenge alle beslutningstakere får tilgang på den samme informasjonen og er kjent med den samme intensjonen vil de konkludere likt. Empiriske studier, blant annet innenfor kommando- og kontrolliteraturen, viser at dette ikke er tilfellet. Informasjon må behandles og framstilles på relevant vis før den kan brukes: informasjon om fienden foredles til etterretning, mens eksempelvis geografisk informasjon må tilpasses et egnet oppløsningsnivå for å sammenstilles med informasjon om egne og andre styrker. Alt dette må i sin tur tilpasses avdelingens nivå og oppdrag. Det er derfor ikke noe enkelt svar på hvordan informasjon bør presenteres, det avhenger av formål, bakgrunnskunnskap og kontekst.

Med ontologibasert dataintegrasjon er det mulig å sette de samme underliggende datakildene i perspektiv på forskjellig vis avhengig av interesse og relevans. Siden integrasjonen av dataene er

virtuell og formidlet av begrepsmodellen, er dette i bunn og grunn et spørsmål om å bytte til en annen begrepsmodell som uttrykker andre sammenhenger. Man kan derfor tenke på en begrepsmodell som en “semantisk linse” som man betrakter dataene gjennom. Forskjellige linser avslører forskjellige sammenhenger i de underliggende dataene.

5 Forskningsaktiviteter og resultater

Basert på forstudier som de i forrige kapittel, har den mer spesifikke ambisjonen for semantikk-aktiviteten i 1277 vært å utvikle et ontologibasert dataintegrasjonsverktøy som egner seg også for dynamiske nettverkstopologier. Dette innebærer bl. a.

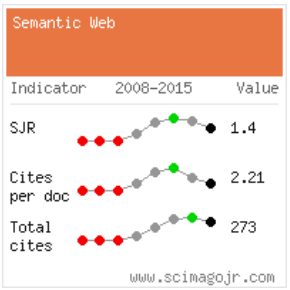
1. at det ikke skal være en forutsetning at kildetilfanget er konstant; kilder kan komme og gå,
2. at det ikke skal være en forutsetning at kildene har samme eier eller er under sentral kontroll,
3. at både sivile og militære kilder skal kunne utnyttes,
4. at integrasjonen skal skje i sanntid uten mellomlagring av data.

Et slikt verktøy vil heretter kalles en *federator*.

5.1 Et matematisk rammeverk for sunn og komplett federering

Det finnes verktøy på markedet allerede som oppfyller kravene i denne listen. Integreringsverktøyet FedX fra selskapet fluidOps, f.eks., er mye brukt. Da det allikevel ble bestemt at vi skulle gjøre et arbeid på å definere et nytt system, skyldtes det at vi etter et omfattende og nær uttømmende litteratursøk ikke kunne finne noen federator som oppfyller Forsvarets krav til etterrettelighet.

Etterrettelighet i denne sammenhengen kan forstås vha. begrepene *sunnhet* og *kompletthet*. Et federeringsverktøy er sunt dersom det ikke "dikter opp svar" dvs. gir falske positiver, og det er komplett dersom det ikke utelater svar dvs. gir falske negativer. Ingen av verktøyene som eksisterer på markedet i dag er teoretisk velfundert i den forstand at de tillater oss å si noe om sunnhet- og kompletthetsegenskaper under ulike betingelser.



Figur 5.1 *SWJ scientometri.*

Betydningen av sunnhet tør være åpenbar. Til spørsmålet om betydningen av kompletthet, er et todelt svar muligens mest dekkende: fra et praktisk synspunkt er kompletthet underordnet brukerbehov. Et ufullstendig svar som returnererer innenfor rimelig tid er som regel å foretrekke fremfor et fullstendig svar som stiller overdrevne krav til bruk av prosessorkapasitet og/eller minne. Fra et forvaltningsperspektiv, er godt kartlagte kompletthetsegenskaper imidlertid svært viktig, da et kompletthetsresultat i bunn og grunn gir en matematisk karakterisering av *oppførselen* til en federator. Sagt annerledes; en teori om betingelsene under hvilke en federator gir sunne og komplette svar er samtidig en teori om når federatoren kan forventes å produsere falske negativer og/eller positiver, under hvilke omstendigheter slike tilfeller kan utelukkes, hva det vil koste i tid og minne å eliminere *alle* feilmarginer, etc.

Forskningsinnsatsen på semantikkområdet i 1277 ble i første omgang fokusert på dette problemet, dvs. på karakteriseringen av sunne og komplette spørringer. Mer spesifikt utarbeidet vi et matematisk

rammeverk som gir et sett av nødvendige og tilstrekkelige vilkår for informasjonssøk over forskjellige konfigurasjoner av RDF-kilder. Dette arbeidet er publisert i tidsskriftsartikkelen

- A. Stolpe & J. Halvorsen: "A Logical Characterization of SPARQL Federation". *Semantic Web Journal* , vol. 6. no. 6, pp. 565-584. IOS Press.

Denne artikkelen inneholder også et evalueringsavsnitt, der vi bruker vårt foreslåtte matematiske rammeverk til å analysere oppførselen til de mest kjente federeringsverktøyene som allerede finnes på markedet. Dette ga oss samtidig et godt grunnlag for å vurdere hvilke verktøy som vil kunne egne seg for militære anvendelser.

Om Semantic Web Journal:

"The Semantic Web journal, brings together researchers from various fields which share the vision and need for more effective and meaningful ways to share information across agents and services on the future internet and elsewhere."

Scientometri fra 2013 (statistikk fra SCImago). Jamfør Figur 5.1:

- Blant alle listede tidsskrifter og konferanser innenfor computervitenskap i verden kommer Semantic Web Journal på 18. plass.
- Blant alle listede tidsskrifter og konferanser innenfor computer nettverk og kommunikasjon kommer Semantic Web Journal på 2. plass.
- Blant alle listede tidsskrifter og konferanser innenfor informasjonssystemer i verden kommer Semantic Web Journal på 3. plass.
- Blant alle listede tidsskrifter og konferanser innenfor anvendt computervitenskap i verden kommer Semantic Web Journal på 6. plass.

5.2 Lærdommen fra UV14

Som et naturlig neste steg i semantikkarbeidet ble det implementert en prototype basert på det teoretiske arbeidet beskrevet i forrige avsnitt. Det ble besluttet å teste denne programvaren ved å forsøke å sammenstille/analysere ISR-data under øvelsen Unified Vision 2014 (UV14) på Ørland i perioden 18-28. mai 2014.

Unified Vision samlet ca. 2000 deltakere fra 18 land, og et stort antall luft-, bakke og sjøplattformer støttet av operatører og systemspesialister. Formålet med øvelsen var å demonstrere innsamling og bruk av informasjon i planlegging og gjennomføring av militære operasjoner (se [4]). Hovedkontingenten fra FFI kom fra prosjektet Nettverksbasert ISR, og deltakelsen fra 1277 ble sikret med støtte fra dette prosjektet.

Operatørstasjonene i UV14 besto av klienter med programvare utviklet i MAJIIC2-prosjektet. Denne programvaren ble benyttet til visning og bruk av funksjoner som kunne konfigureres til

å støtte enhver K2- og/eller JISR-rolle. På UV14 representerte JISR-cellen prosessene assosiert med å samle og håndtere informasjons- og etterretningsbehov. UV14 ble derfor vurdert som en i utgangspunktet god testcase for dataintegrasjonsverktøy generelt, og semantikkbaserte verktøy spesielt.

MAJIIC2: Multi-intelligence All-source Joint Intelligence Surveillance and Reconnaissance Interoperability Coalition

Er et 9-nasjoners prosjektprogram for utvikling av spesifikasjoner og standarder innenfor kapabilitetsområdet JISR som ble initiert i 1997 ved Paris Interoperability Experiment, og videreført i CAESAR (2001), MAJIIC (2005) og MAJIIC2 (2011).

5.2.1 Funn

Gjennom samtaler med representanter for NCIA på UV14, lærte vi at de allerede hadde gått langt i retning av å bruke RDF som en sentral komponent i MAJIIC-infrastrukturen.

Hensikten med MAJIIC2 er å støtte samarbeid gjennom deling av konsistente data på tvers av et upålitelig 'mission network'. Datamodellen til MAJIIC2 definerer strukturen til disse dataene, mens dataene selv lagres i en såkalt *Non-Relational Structured Storage Service*. Denne lagringstjenesten må støtte detaljerte spørringer om MAJIIC-dataene, f.eks.:

- Hvilke RFIer (*Request For Information*) har kommet inn i løpet av de siste 24 timene? Tittel og beskrivelse?
- Hvilke oppgaver er løpende?
- Hva er K2-forholdet mellom to enheter?

For å kunne svare på slike detaljerte spørringer ble MAJJIC-dataene på UV14 utvekslet som XML via WS-Transfer Web-tjenester. XML-meldingene ble så transformert til RDF, lagret i RDF-databaser og deretter analysert vha. SPARQL (se avsnitt 2.2.3).

Det var følgelig en naturlig ambisjon for semantikkaktiviteten i 1277 å forsøke å sammenstille informasjon fra disse og andre datakilder (f.eks. ORBAT-databaser og hendelseslogger) gjennom federering. Som nevnt tidligere, var forventningen å kunne garantere korrekte og komplette svar med utgangspunkt i det teoretiske grunnlagsarbeidet og vår egen prototype, som beskrevet i forrige avsnitt.

5.2.1.1 Problemet med blanke noder i RDF data

Det som imidlertid viste seg å være et problem var at autotransformeringene av XML til RDF produserte mange såkalte *blanke RDF-noder*. En blank node (også kalt en *bnode*) er en node i en

RDF-graf som ikke er identifisert vha. en URI. Den er en anonym node som er å betrakte som en logisk variabel. Blanke noder brukes til å si slike ting som 'x er en artillerienhet' uten å identifisere x.

Det er svært vanlig at ferdigdefinerte transformasjoner fra XML eller tabulære formater til RDF benytter blanke noder for å gi en oversettelse som er mest mulig direkte og konservativ. De fleste hyllevarene på dette området benytter en slik direkte transformasjonsstrategi (se avsnitt 2.3) dersom de ikke fôres med håndkurerte transformasjonsregler. Det er enklest å illustrere hvordan dette fungerer med en tabell. Tabellen 5.1 gir et oppdiktet eksempel på hvordan data kan se ut i en tabell:

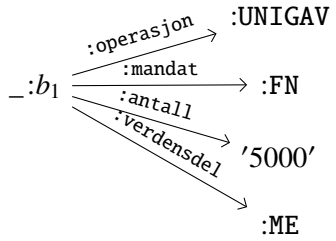
Operasjon	**Mandat**	**Antall soldater**	**Verdensdel**
UNIGAV	FN	5000	ME
IFAK	NATO	321	E
NAGWAG	OSSE	92	NA

Tabell 5.1 *Imaginære tabulære data.*

For å representere denne tabellen i RDF, så må relasjonenene mellom kolonnene bevares. Det fordrer at etterfølgende celler knyttes til samme rad. Raden er imidlertid ikke selv et dataelement, den er kun en entitet som knytter et sett av dataverdier sammen i den ønskede relasjonen. En direkte transformasjon vil derfor måtte representere selve raden, for så å knytte hver av dataverdiene til denne ved å bruke kolonnene som RDF-predikater. Siden raden ikke er et navngitt dataelement, er det vanlig å bruke en blank node for dette formålet. Dersom vi tar første rad i tabell 5.1 som eksempel, vil resultatet av transformasjonen være figur 5.2 der den blanke noden merket $\_:b_1$ representerer raden, mens alle andre navn som begynner med kolon uten understrek foran er URIer, altså entydige identifikatorer for dataverdiene de angir (tall angis med såkalte *literal values*, dvs. som strenger).

Problemet med blanke noder fra et federeringsperspektiv er at siden de er anonyme, altså at de ikke er URIer, så er de nettopp ikke utvetydige utenfor sin opprinnelige kontekst. Det gjør at en blank node ikke kan gjenkjennes på tvers av SPARQL-spørringer, noe som er nødvendig forutsetning for federering.

Vi oppdaget dette problemet da vi testet vår egen prototype mot RDF-databaser med NRS3 data. Det viste seg at de fleste spørringen ikke returnerte komplette svar, og at informasjonstapet var betydelig.



Figur 5.2 *Direkte transformering av første rad i tabell 5.1.*

5.2.1.2 Omfanget av problemet

Dette problemet med blanke noder i federering viste seg å være presserende. Det er to årsaker til dette.

For det første fantes det på det tidspunktet vi ble klar over det, ingen teoretisk beskrivelse av problemet — knapt nok en erkjennelse[1] — og ingen programvare som håndterte det. Etter at vår egen prototype viste seg utilstrekkelig, gjorde vi et grundig søk gjennom forskningslitteraturen og prøvde ut alle eksisterende federatorer inkludert markedslederen FedX. Ingen var i stand til å håndtere blanke noder i dataene.

For det andre er det gode grunner til å anta at dette problemet ville ha møtt oss igjen og igjen, spesielt dersom man ser for seg å utnytte sivile kilder og *Open Source Intelligence* i tillegg til militære — f.eks, for oppgaver slik som etterretning og overvåkning. Blanke noder er en del av RDF-spesifikasjonen og en av de grunnleggende byggeklossene i Semantic Web. De brukes bl. a. for å:

1. representere relasjoner med høyere aritet enn 2,
2. reifisering RDF-tripler, dvs. gjøre dem til gjenstand for andre påstander (angående f.eks. utgiver eller troverdighet),
3. skjule data som ikke skal være synlig [2],
4. beskrive lister av objekter,
5. representere en klasse av objekter som står i en bestemt relasjon til minst ett annet objekt.

Det er verd å legge merke til at blanke noder med andre ord er svært nyttige for å uttrykke abstrakte logiske konsepter (1, 4, 5) som er nødvendige i begrepsmodeller.

Blanke noder forekommer i mange W3C-standarder, verktøy og publiserte datasett. En forholdsvis ny empirisk studie konkluderer:

> Of the 783 domains contributing to our corpus, 345 (44.1%) did not publish any blank nodes. The average percentage of unique terms which were blank nodes for each domain — i.e., the average of %bnodes for all domains — was 7.5%, indicating that although a small number of high-volume domains publish many blank nodes, many other domains publish blank nodes more infrequently [2].

Dersom denne studien er representativ betyr det at blanke noder kan forventes å forekomme i omtrent halvparten av alle RDF-datasett. At det hverken finnes teori eller implementasjoner som håndterer dem er derfor ikke bare et problem i vår begrensede UV14 kontekst, men et problem for The Semantic Web som sådan.

[1] Et unntak er [1]: "The scope of blank nodes is limited to the document in which they appear , [...] reducing the potential for interlinking between different Linked Data sources. [...] it becomes much more difficult to merge data from different sources when blank nodes are used, [...] Therefore, all resources in a data set should be named using URI references."

Blanke noder:

En blank node (også kalt en *bnode*) er en node i en RDF-graf som ikke er identifisert vha. en URI. En blank node er derfor en anonym node som er å betrakte som en variabel. De brukes til å si slike ting som 'x er en artillerienhet' uten å identifisere x nærmere.

5.3 Generalisering av det matematiske rammeverket

Erfaringene og etterarbeidet fra UV14 viste at det ligger en spenning i selve idéen om RDF-federering som så langt ikke har blitt håndtert av det vitenskapelig samfunnet forøvrig. Denne spenningen ligger i forholdet mellom, på den ene side, federatorens behov for å re-identifisere RDF-noder fra kilde til kilde og, på den andre, eksistensen av navnløse blanke noder i disse kildene.

Dette problemet trengte å bli gitt en matematisk beskrivelse og løsning dersom det skulle ha noen hensikt å gå videre med implementeringsarbeidet. Derfor antok også semantikkarbeidet i 1277 en mer teoretisk karakter. Resultatet av dette arbeidet er rapportert i tidsskriftartikkelen under:

- A. Stolpe & J. Halvorsen: "Distributed Query Processing in the Presence of Blank Nodes". *Semantic Web Journal* (*forthcoming*).

Artikkelen viser at det er mulig å løse ovennevnte federeringsproblem ved å være omhyggelig med hvordan og i hvilke deler en global spørring sendes rundt til de ulike kildene, men at det forutsetter en ganske dyptgripende generalisering av selve SPARQL-semantikken.

5.3.1 Beregnbarhetsaspektet

På det tidspunktet vi begynte å jobbe med federering, var alle eksisterende federeringsverktøy (og dersom vi ikke regner med vår egen implementasjon, så er det slik fremdeles) basert på å splitte en global spørring i én partisjon, for så å sende hver av cellene i denne partisjonen til alle kildene som kan besvare den. Federeringsverktøyet FedX fra selskapet fluidOps er bygget over denne idéen.

Vårt arbeid viser at denne federeringsstrategien kun vil være sunn og komplett dersom det ikke eksisterer blanke noder i kildene. Dersom blanke noder finnes i kildene, så vil ikke én enkelt partisjon av den globale spørringen være nok fordi de blanke nodene gjøre at svar som kombineres fra flere kilder krever andre oppdelinger av den globale spørringen enn svar som kommer fra én enkelt kilde. Det er derfor nødvendig å prøve mange forskjellig partisjoner av den globale spørringen.

Det er desverre faktisk slik at det kombinatoriske rommet her er formidabelt. Som et eksempel vil en spørring som uttrykker 8 betingelser i *verste fall* trenge å bli oppdelt på 4 140 forskjellige vis (som er er Bell-nummeret til 8), mens en spørring med 10 vilkår øker dette tallet til 115 975.

Konklusjonen her er uunngåelig: RDF-federering, dersom man ikke utnytter konkret kunnskap om grafstrukturene i kildene, er i det verste tilfellet ikke effektivt beregnbart.

Vi har arbeidet mye med å begrense denne kombinatoriske eksplosjonen. Det viser seg at det i de aller fleste realistiske tilfeller sannsynligvis er mulig å gjøre situasjonen praktisk overkommelig. Grovt sett følger vi to spor:

Det ene undersøker hvordan man kan begrense antall nødvendige kombinasjoner av delsvar ved å gjøre mellomresultater (resultater fra én eller flere kilder som sammenstilles med resultater fra en annen) så små som mulig (uten å miste informasjon). Dette arbeidet er mer eller mindre ferdig og er rapporter i to artikler

- J. Halvorsen & A. Stolpe: "On Minimal Intermediate Results in Zero-Knowledge Federation". Under vurdering for *Semantic Web Journal*, og
- A. Stolpe & J. Halvorsen: "Minimal Intermediate Results II: Adjusting the concepts". Denne artikkelen er ferdigskrevet og klar for publisering, men avventer publiseringen av den første.

Det andre sporet vi følger forsøker å holde antallet nødvendige partisjoner av den globale spørringen så lavt som mulig. Partisjonene er nødvendige for å dekke alle muligheter når man ikke vet noe om grafstrukturene i kildene. Det overveiende flertallet av dem vil dog være bomskudd og vil ikke returnere svar. Idéen er å forsøke å utelukke disse tidlig ved å sondere kildene for informasjon som forteller noe om relevante grafmønstre. Foreløpige resultater indikerer at det vil være mulig i de fleste realistiske tester å redusere antall partisjoner av en spørring med omtrent 10 vilkår til under 10 partisjoner. Dette vil dog variere med hvordan data er fordelt i kildene, og en grundig empirisk evaluering vil være nødvendig.

Algoritmisk er dette arbeidet godt utviklet på det tidspunktet denne rapporten skrives. Den empiriske evalueringen gjenstår, så der er det for tidlig å konkludere.

5.4 Programvare

Hovedresultatet i tidsskriftartikkelen [3] gir en sunn og komplett *generell* federeringstrategi for RDF-data. Den er generell i den forstand at den ikke gjør noen antagelser om hvordan dataene ser ut, hvilket også vil si at de kan inneholde blanke noder.

I 2014 publiserte vi første versjon av federeringsverktøyet Hårfagre, som er en implementasjon av denne teorien. Gjennom teoretisk kontroll over sunnhet og kompletthet, er systemet spesielt utviklet for å møte militære krav til etterrettelighet. Programvaren, som er skrevet i Scala, er publisert under en GNU 2 lisens for åpen kildekode og er fritt tilgjengelig på BitBucket.

Programvaren implementerer ikke heuristikkene som er beskrevet i forrige avsnitt, noe som er pågående arbeid. Hårfagre vil ikke ha problemer dersom det ikke er blanke noder i kildene, men den kan foreløpig ikke forventes å yte godt i alle tilfeller av større spørringer i motsatt fall. Vi jobber med å forbedre gjennomsnittsytelsen på bakgrunn av heuristikkarbeidet som ble beskrevet i forrige avsnitt, som vi forventer vil ha stor betydning.

5.5 Oppsummering av resultater

Pga. de oppdagelsene vi gjorde mens vi arbeidet med case-studier, mer spesifikt med data fra UV14, så har semantikkaktiviteten i 1277 i all hovedsak vært konsentrert om å løse grunnlagsproblemer innenfor federeringsteori og SPARQL-semantikk. Til gjengjeld er det ingen overdrivelse å si at vi har klart å flytte dette fagfeltet betydelig, med flere publikasjoner i gode fagfellevurderte tidsskrifter, og flere, håper vi, på vei i samme retning.

Vi har gjennom dette arbeidet skaffet oss et veldig godt utgangspunkt for å anslå hvorvidt disse teknologiene egner seg for Forsvaret, hva som er deres *pros* og *cons* samt hvilke tilpasninger som er nødvendige. Vi går gjennom dette i form av en serie med anbefalinger i neste kapittel.

6 Anbefalinger og konklusjon

Gitt det vi har lært i løpet av 1277 om hvilke utfordringer RDF-federing byr på, så er det opplagte spørsmålet om vi er i posisjon til å anbefale videre aktivitet på dette området. Dette er selvsagt et spørsmål om å veie fordeler og ulemper mot hverandre:

Fordeler:

- RDF-data er i en viss forstand selvbeskrivende data.
- RDF-data har høy gjenbruksverdi.
- Metadata kan flyttes oppstrøms og forankres direkte i dataelementet.
- Identifikatorer i RDF er utvetydige hvor som helst på nett.
- RDF-data er skjemauavhengig, og kan forvaltes som sådan.
- RDF er en ikke-proprietær standard.
- Legacy-data kan eksponeres som RDF uten endring.
- RDF har et standarisert spørrespråk (SPARQL).
- RDF er logikkbasert og støtter avanserte analyseprosesser basert på AI-algoritmer.
- RDF-databaser blir mer og mer vanlige i stordata, og har gode egenskaper sammenliknet med tradisjonelle MapReduce-systemer.

Ulemper:

- Datarepresentasjon i RDF forutsetter en strategi for å prege identifikatorer: hver dataeier i Forsvaret trenger et navnerom for sine URIer, og noen må bestemme hvordan disse navnene skal utformes og organiseres.
- Blanke noder er lite velegnet for federering.

Problemet med blanke noder kan mildnes på minst to måter:

1. man kan sørge for å unngå dem i sine *egne* data,
2. eller man kan benytte heuristikk for å redusere konsekvensene.

Det er riktignok slik at punkt 2 her setter sin lit til et arbeid vi ennå ikke har ferdigstilt, selv om de foreløpige resultatene er oppmuntrende. Uansett så vil det første punktet alene være nok til å nøytralisere problemet for "in-house" data. Problemet vil da kun oppstå dersom man sammenstiller data fra sine egne kilder med andre — sivile eller militære — man ikke selv har kontroll over.

Summa sumarum så er det vår oppfatning at dataorientering generelt (jamfør avsnitt 2.2.4), og RDF spesielt, har et stort potensiale for Forsvaret. Fordelene oppveier ulempene, men ulempene må selvsagt håndteres. Vi forsøker å gjøre det i anbefalingene som følger.

Anbefalingene er strukturert som en modenhetsstige, der hvert trinn gir gevinster som er uavhengig av høyere trinn. Alternativt kan avsnittet betraktes som et grovt omriss av en implementeringsplan for semantikkstøtte i Forsvarets informasjonsinfrastruktur.

Anbefaling 1: en URI-strategi for Forsvaret

Vi anbefaler at det skrives en praktisk veileder for design av URIer i Forsvaret. Det er meningen at dette dokumentet skal definere overordnede prinsipper for datarepresentasjon i RDF, såvel som tilpasninger som er spesifikke for Forsvaret. Disse inkluderer:

- hvilket domene man skal velge for et sett av URIer (eks. `www.forsvaret.no/saniteten/`),
- hvilken struktur URIen skal ha (eks. REST struktur, meningsbærende eller ikke),
- språk og målformer i URIer,
- hvordan man skal unngå blanke noder uten å miste uttrykkskraft,
- hvordan man skal håndtere endring og tid,
- hvordan oppslag på URIer skal gjøres,
- hvilke kvalitetskarakteristikker som alle datasett bør kjennetegnes av,
- hvilke metadataformater som skal være tilgjengelige,
- beskrivelse av det forvaltningsregime som er nødvendig for å øke tilliten til, og derfor gjenbrukbarheten av, dataene.

Den stipulerte effekten av disse retningslinjene vil være:

- konsistent bruk av begreper og terminologi på tvers av domener,
- betydningen av dataelementene presiseres, tvetydighet fjernes,
- gjenbruksverdien øker,
- legacy data kan eksponeres som RDF (jamfør avsnitt 2.3 og avsnitt 2.2.1),
- federering kan gjøres effektivt i *alle* tilfeller

Dokumentet vil være direkte relevant for:

- forvaltningsinstanser som trenger å standardisere begrepsbruk,
- dataeiere innen Forsvaret,
- utviklere av programmeringsgrensesnitt (APIer) inkludert SOA- og REST-grensesnitt

Det finnes tilsvarende dokumenter fra sivil sektor i andre land. Av disse er "Designing URI Sets for the UK Public Sector" det som virker mest velutviklet. Dette dokumentet danner grunnlaget for data.gov.uk som er et britisk statlig prosjekt for å gjøre data fra offentlig forvaltning åpent tilgjengelig på maskinleselige formater.

Det anbefales at "Designing URI Sets for the UK Public Sector" brukes som modell for et tilsvarende dokument for Forsvaret.

Anbefaling 2: en ledsagende metadatastrategi

En konsistent bruk av metadata vil kunne øke verdien til et datasett betydelig. Vi anbefaler at det utvikles en metadatastrategi som ledsager til URI-strategien. Metadataene bør innholde informasjonstypene i tabell 6.1.

Type metadata	Formål
Begrepsdefinisjon	Presisere betydning og unngå forveksling
Relasjoner til andre datasett	Lenke til andre relevant RDF-datasett.
Proveniens	Oppgi kilden og formålet med datasettet
Nøyaktighet	Indikere betydningsfulle feilmarginer
Kompletthet	Indikere dekningsgrad
Tidsriktighet	Beskrive tidsforsinkelsen mellom fakta og data
Lisenser og juridiske vilkår	Angi vilkår for bruk, lagring, utveksling etc.
Levetid	Å angi en garantert gyldighetsperiode og en utløpsdato
Representasjoner	Indikere hva slags beskrivelser av en URI som er tilgjengelige

Tabell 6.1 *Anbefalte metadata*

Den stipulerte effekten av disse retningslinjene vil være:

- at sporbarheten og gjennomsiktigheten av et datasett øker,
- at arkiveringsverdien øker,
- at datautveksling blir juridisk gjennomsiktig,
- at et datasett vil kunne utnyttes med større presisjon,
- at den totale gjenbruksverdien av et datasett øker.

I tilegg til å bestemme hvilke typer av informasjon et metadatasett skal ha, vil en metadatastrategi også standardisere merkingen eller kodifiseringen av dem. Det vil si, den vil bestemme hvilke begrepsmodeller og RDF-vokabularer metadataene skal samsvare med.

Anbefaling 3: Utvikle kompetanse på grafdatabaser

De to første anbefalingene gir et godt utgangspunkt for RDF-eksponering av legacy-systemer slik som relasjonelle databaser. En god URI-strategi fostrer konsistent datarepresentasjon også mellom dataeiere. En metadatastrategi, på den annen side, flytter forvaltningsrelevante data oppstrøms og bidrar til sporbarhet og juridisk transparens. RDF-eksponering er imidlertid begrenset av det underliggende legacy-systemet, som ikke nødvendigvis er godt tilpasset en graforientert prosesseringsmodell.

En graforientert prosesseringsmodell er spesielt naturlig for analyse av store datamengder, for eksempel for etterretning og overvåkning. Slik analyseoppgaver vil ha mye å tjene på å kunne utnytte grafalgoritmer, eksempelvis analyse av stier, identifisering av naturlige klynger av noder, rangering av kandidater, identifisering av relevante subgrafer m.m. Dagens grafdatabaser kommer med et stort antall av slike velstuderte optimaliserte algoritmer.

En graf er en fleksibel og intuitiv datastruktur som ofte kan være et naturlig valg også fra et forvaltningsperspektiv. I kontrast til relasjonelle databaser, er ikke grafdatabaser definert av et predefinert skjema som krever manuell kurering for å endres. De er derfor bedre tilpasset situasjoner der dataene er uforutsigbare, eller kan tenkes å endre seg raskt. Dette kan være situasjonsbilder,

innsamlet etterretningsdata, innsamlede data fra sosiale medier, summen av ORBAT-informasjon i koalisjonsoperasjoner eller *Joint Intelligence Surveillance and Reconnaissance*-informasjon mer generelt. I mange slike tilfeller vil konvensjonelle lagringsløsninger vise seg å være en dårlig match for å lagre og analysere informasjonen som samles inn.

Anbefaling 4: Utvikle federering som et alternativ til datavarehus

Steget fra eksponering av relasjonelle databaser i RDF til å lagre data *som* RDF i grafdatabaser, vil gjøre at analysesystemer kan håndtere mye større datamengder og drille dypere i dataene. Et stordataramme verk er imidlertid fortsatt å betrakte som et lagringssystem som er underlagt sentral redaksjonell kontroll.

Mange av analyseoppgavene Forsvaret står overfor forutsetter bruk av kilder som ikke kan er "in-house". Spesielt gjelder dette overvåkning og etterretning der sporingssystemer og mediestrømmer fra sivil sektor er viktige informasjonsressurser. Det gjelder imidlertid også andre typer data, slik som for eksempel helsedata. Helsedata eksisterer på tvers av samfunnsektorer og juridiske barrierer og bringer en ekstra dimensjon til sammenstillingsproblemet - vi liker å tenke på det som *juridisk skalerbarhet*.

Det er derfor i hovedsak to grunner til at det er interessant å fortsette å utvikle kompetanse på federeringsteknologier.

1. Federering er nødvendig dersom informasjon på tvers av systemer og sektorer skal kunne utnyttes effektivt uten mellomlagring og caching av data.
2. Mellomlagring og caching av data vil som regel ikke være juridisk skalerbart, siden fysisk samlokalisering av data er svært tungvint med tanke på relevant lovgivning. Med en infrastruktur basert på federering, er det mulig å tenke seg at spørringer klareres enkeltvis, innenfor et regulert tidsrom. Siden sammenstilling ikke fordrer samlokalisering, unngår man juridisk sammenstillingsproblematikk utover det som angår den bestemte spørringen det er snakk om.

6.1 Konklusjon

Det er vår oppfatning at disse anbefalingene sammen gir en helhetlig og robust tilnærming til semantikkbasert informasjonsforvaltning for Forsvaret. Hvert av trinnene er en forutsetning for og forsterker hvert av de neste, men har verdi uavhengig av høyere trinn. F.eks. vil det være mange gevinster knyttet til å oppfylle de første to trinnene som innebærer å bestemme seg for en konsistent måte å utforme URIer for Forsvaret på, og å bestemme seg for hvilke metadata som skal knyttes til dem og hvordan de skal representeres. Umiddelbare gevinster vil være gjenbrukbarhet, skjemauavhengighet, arkiveringsverdi, informasjonsinnhold, etc.

Virkelig interessant blir det imidlertid når man tenker seg at informasjonsforvaltning på dette grunnivået brukes til å støtte federering. Det er først da synergieffektene av utvetydige veldokumenterte data vil kunne løse utfordringer som har å gjøre med å krysse ulike byråkratiske kulturer og juridiske barrierer. Det vi foreslår er i bunn og grunn en metode for å gjøre dette trinn for trinn og etter behov, etterhvert som de nødvendige forutsetningene avklares.

Referanser

[1] Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 1st edition, 2011.

[2] Alejandro Mallea, Marcelo Arenas, Aidan Hogan, and Axel Polleres. On blank nodes. In *Proceedings of the 10th International Conference on The Semantic Web - Volume Part I*, ISWC'11, pages 421–437, Berlin, Heidelberg, 2011. Springer-Verlag.

[3] Audun Stolpe and Jonas Halvorsen. Distributed query processing in the presence of blank nodes. *Semantic Web Journal (forthcoming)*, 2016.

[4] Leiv Øyehaug, Trygve Sparr, Torgar Haugen, Sigbjørn Aune, Simen Rustad, Pål Bjerke, and Rune Stensrud. FFIs evaluering av nasjonalt bidrag under Unified Vision 14. FFI-rapport 16/01398, Norwegian Defence Research Establishment (FFI), 2016.

About FFI

The Norwegian Defence Research Establishment (FFI) was founded 11th of April 1946. It is organised as an administrative agency subordinate to the Ministry of Defence.

FFI's MISSION

FFI is the prime institution responsible for defence related research in Norway. Its principal mission is to carry out research and development to meet the requirements of the Armed Forces. FFI has the role of chief adviser to the political and military leadership. In particular, the institute shall focus on aspects of the development in science and technology that can influence our security policy or defence planning.

FFI's VISION

FFI turns knowledge and ideas into an efficient defence.

FFI's CHARACTERISTICS

Creative, daring, broad-minded and responsible.

Om FFI

Forsvarets forskningsinstitutt ble etablert 11. april 1946. Instituttet er organisert som et forvaltningsorgan med særskilte fullmakter underlagt Forsvarsdepartementet.

FFIs FORMÅL

Forsvarets forskningsinstitutt er Forsvarets sentrale forskningsinstitusjon og har som formål å drive forskning og utvikling for Forsvarets behov. Videre er FFI rådgiver overfor Forsvarets strategiske ledelse. Spesielt skal instituttet følge opp trekk ved vitenskapelig og militærteknisk utvikling som kan påvirke forutsetningene for sikkerhetspolitikken eller forsvarsplanleggingen.

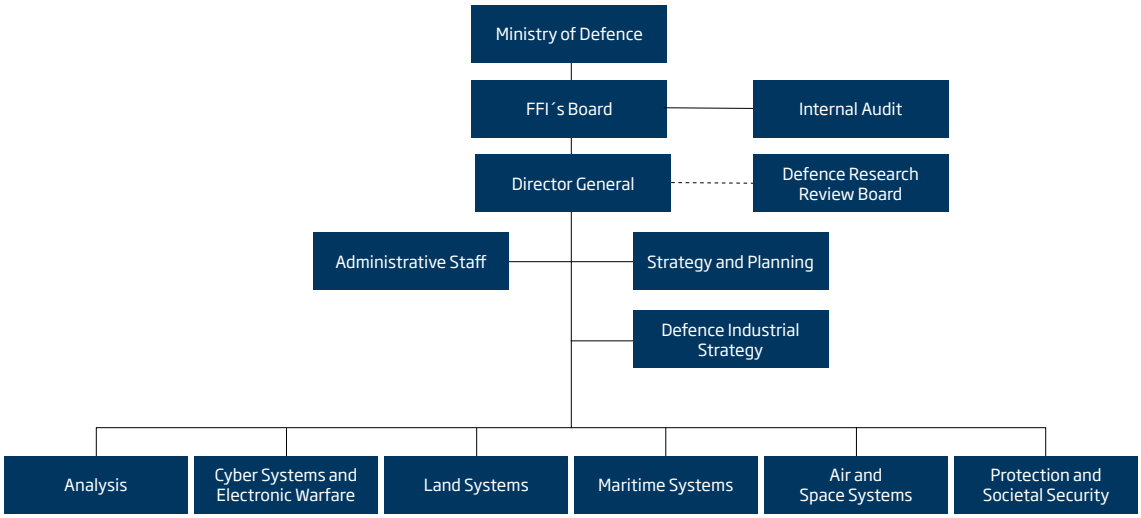
FFIs VISJON

FFI gjør kunnskap og ideer til et effektivt forsvar.

FFIs VERDIER

Skapende, drivende, vidsynt og ansvarlig.

FFI's organisation



Forsvarets forskningsinstitutt
Postboks 25
2027 Kjeller

Besøksadresse:
Instituttveien 20
2007 Kjeller

Telefon: 63 80 70 00
Telefaks: 63 80 71 15
Epost: ffi@ffi.no

Norwegian Defence Research Establishment (FFI)
P.O. Box 25
NO-2027 Kjeller

Office address:
Instituttveien 20
N-2007 Kjeller

Telephone: +47 63 80 70 00
Telefax: +47 63 80 71 15
Email: ffi@ffi.no

FFI Forsvarets forskningsinstitutt
Norwegian Defence Research Establishment