WESTERN NORWAY RESEARCH INSTITUTE
VESTLANDSFORSKING
www.vestforsk.no

Vestlandsforsking-rapport nr. 9/2011

Begrepsapparat for reiselivet

Forslag til utvikling av ei felles ”ordliste” for reiselivet

Svein Ølnes og Nils Arne Hove

Vestlandsforsking, Pb 163, 6851 Sogndal • Tlf.: 57 67 61 50 • Faks: 57 67 61 90

Vestlandsforsking-rapport

Tittel Begrepsapparat for reiselivet	**Rapportnummer** 9/2011 **Dato** februar 2011 **Gradering** Open
Prosjekttittel Begrepsapparat for reiselivet – Forslag til utvikling av felles ”ordliste” for reiselivet	**Tal sider** : 25 + 21 **Prosjektnr.** : 6200
Forskar(ar) Svein Ølnes (prosjektleiar) Nils Arne Hove	**Prosjektansvarleg** Ivar Petter Grøtte
Oppdragsgivar NCE Tourism VisitNorway	**Emneord** reiseliv begrepsapparat emneord

Samandrag

I denne forprosjektrapporten er det skissert forslag til utvikling av eit felles begrepsapparat med utgangspunkt i tellUs sine eksisterande kategoriar. Begrepsapparatet omfattar den **tematiske** delen av reiselivsinformasjonen, ikkje andre dimensjonar som t.d. geografi, tid m.m.

Forslaget byggjer på å erstatta dagens hierarkiske kategorisystem med **emneord**. Hierarkiet vil ikkje bli borte, men det er emneorda som vil stå i sentrum, den tematiske strukturen over emneorda er sekundær. Til emneorda føreslår vi å kopla **hjelpeord** for lettare å navigera seg fram til korrekt emneord. Hjelpeorda vil vera til hjelp både for dei som legg inn informasjon, og for sluttbrukarane i søkesamanheng.

Vi føreslår vidare at emneorda blir publiserte i ”nett-skya”, lett tilgjengeleg for alle som vil bruka dei. Emneorda må kontrollerast og ha eit system for jamleg vedlikehald (= kontrollert vokabular), medan hjelpeorda er friare og bør kunna leggjast til av alle som ønskjer det.

Dette blir det første, viktige steget i ei utvikling mot *lenka data* (*Linked Data*). Men for å realisera lenka data, må også informasjonsressursane emneorda blir knytte opp mot, publiserast ope. Det blir ei naturleg vidareføring når det kontrollerte vokabularet er på plass. Til saman vil dette bli viktige byggjesteinar i etableringa av eit digitalt **økosystem for reiselivet**.

ISBN: 978-82-428-0309-2	Pris:

Forord

På eit møte i 2010 mellom sentrale reiselivsaktørar i Norge, vart det bestemt at det skulle setjast i gang eit arbeid for å rydda i dagens begrepsapparat i reiselivet. Initiativet kom frå NCE Tourism og motivasjonen var eit ønskje om å skapa eit økosystem for digital reiselivsinformasjon der også informasjon frå andre kjelder enn dei "offisielle" frå reiselivet også kan integrerast på ein enklare måte.

Vestlandsforsking fekk i oppdrag å utføra eit forprosjekt og komma med tilrådingar til utviking av eit felles begrepsapparat. Vi takkar NCE Tourism og VisitNorway for oppdraget og håpar at forslaga som er skisserte i denne rapporten vil bli gjennomførte.

Vestlandsforsking, februar 2011

Innhald

Samandrag

Dette er ikkje første gangen det blir gjort forsøk på å etablera eit felles begrepsapparat for reiselivet, og det blir kanskje ikkje den siste heller. Det er ein vanskeleg jobb å setja namn på ting, og endå vanskelegare når det er fleire uavhengige aktørar med i biletet. Likevel er det ein nødvendig del dersom ein skal oppnå betre integrasjon og lettare informasjonsutveksling.

I denne forprosjektrapporten er det skissert forslag til utvikling av eit felles begrepsapparat med utgangspunkt i tellUs sine eksisterande kategoriar. Begrepsapparatet omfattar den **tematiske** delen av reiselivsinformasjonen, ikkje andre dimensjonar som t.d. geografi, tid m.m.

Forslaget byggjer på å erstatta dagens hierarkiske kategorisystem med **emneord**. Hierarkiet vil ikkje bli borte, men det er emneorda som vil stå i sentrum, den tematiske strukturen over emneorda er sekundær. Til emneorda føreslår vi å kopla **hjelpeord** for lettare å navigera seg fram til korrekt emneord. Hjelpeorda vil vera til hjelp både for dei som legg inn informasjon, og for sluttbrukarane i søkesamanheng.

Denne måten å ordna begrep på, er prøvd ut i andre prosjekt, på andre område. Mest nærliggjande er kategoriseringssystemet *Los*, eit system for kategorisering av offentlege tenester. Det er eit system eigd av Difi, og brukt i bortimot 150 kommunar for å binda saman nær-slekta informasjonsressursar. Det viser at metoden duger i praksis, og er skalerbar (handterer ein aukande brukararmasse).

Vi føreslår vidare at emneorda blir publiserte i "nett-skya", lett tilgjengeleg for alle som vil bruka dei. Emneorda må kontrollerast og ha eit system for jamleg vedlikehald (= kontrollert vokabular), medan hjelpeorda er friare og bør kunna leggjast til av alle som ønskjer det.

Dette blir det første, viktige steget i ei utvikling mot *lenka data* (*Linked Data*). Men for å realisera lenka data, må også informasjonsressursane emneorda blir knytte opp mot, publiserast ope. Det blir ei naturleg vidareføring når det kontrollerte vokabularet er på plass.

Til saman vil dette bli viktige byggjesteinar i etableringa av eit digitalt **økosystem for reiselivet**. Ved hjelp av byggjesteinane omtalte ovanfor, skal tredjeparts-aktørar lettare kunna utvikla nye system (applikasjoanr, nettsider) for reiselivet og lettare kunna integrera reiselivsinformasjon med tilgrensande informasjon (kulturbasert informasjon, brukargenerert informasjon m.m.).

Vi tilrår at det vidare arbeidet blir drive av **tellUs** i tett samarbeid med reiselivsaktørar som VisitNorway, FjordNorway/NCE Tourism, Visit Sognefjord, Book Norway m. fl. Arbeidet bør integrerast i alt igangsette prosjekt som tellUs' Skattefunn-prosjekt, og eventuelt prosjektet CREAPURE om det blir finansiert av Noregs Forskingsråd.

1. Bakgrunn og mandat

Etter initiativ frå FjordNorge/NCE Tourism v/Anders Waage Nilsen vart det halde eit møte i Oslo 4. mai i år med representantar frå VisitNorway[1], FjordNorge/NCE Tourism, Nasjonal Booking, tellUs, VisitSognefjord og Vestlandsforsking. Bakgrunnen var eit innspel om behovet for ein nasjonal standard for strukturerte data i reiselivet. Som Anders Waage Nilsen [1] har uttrykt det:

> ”Et overordnet mål er å utvikle en sterkere delings- og utvekslingskultur innenfor en økologi av offisielle reiselivsnettsteder. Denne kulturen må dyrkes. En av fordelene med en felles infrastruktur muliggjør erfaringsdeling, nasjonale kompetansetiltak og samfinansiering av applikasjoner.”

Ønskjet er altså ei ny organisering av strukturert reiselivsinformasjon (”begrepsapparat”) for å møta dei nye utfordringane vi ser i den opnare veven kombinert med den sosiale veven.

2. Utgangspunktet for arbeidet

Reiselivet er ein svært samansett bransje med mange små aktørar og med relativt lite samspel og samhandling på tvers. Dette speglar også nettsatsinga. Det finst etter kvart mange reiselivsportalar og fellestrekket ved dei er at dei ikkje snakkar saman. Det er ikkje unikt for reiselivet, men utfordringane i reiselivet er kanskje endå større enn i andre bransjar med meir likearta og likeverdige aktørar.

Reiselivet er også i ein situasjon der nettet på kort tid har vorte den heilt dominerande kanalen for kundane til å skaffa seg informasjon om reisemål før, under og etter reisa. Kundane ligg solid framfor bransjen, og gapet mellom bruksmønsteret til kundane og tilbodet frå reiselivs-tilbydarane ser berre ut til å berre auka. Internettutviklinga og bruksmønsteret hjå dei reisande bør eigentleg få reiselivstilbydarane til å fryda seg. Men i staden heng bransjen att i gamle vanar og evnar i for liten grad å utnytta det store potensialet som ligg i teknologiskiftet.

Det første steget til betre samhandling for å tilby dei reisande det dei er på jakt etter, er å snakka same språket. Då snakkar vi ikkje om norsk eller engelsk, men om å bruka den same beskrivelsen av reiselivsprodukta. For å samla relatert informasjon om kajakkpadling på fjorden, må reiselivs-produkta utstyrast med like metadata (ord som beskriv innhaldet). Om ein tilbydar kallar sine produkt ”fjordpadling” og ein annan ”fjordkajakk”, blir det ingen samanheng mellom desse. Og dersom reiselivslaget skriv ein artikkel om dei fantastiske mulegheitene for kajakkpadling på Sognefjorden, kan ikkje desse to tilboda automatisk koplast opp utan at nokon seier at både ”fjordpadling” og ”fjordkajakk” er relatert til det felles ordet ”kajakk”.

[1] VisitNorway er brukt som namn på VisitNorway.com og tilhøyrande reiselivsportal som Innovasjon Norge har ansvaret for.

Dette forprosjektet gir ei tilråding til ei ny etablering av eit felles begrepsapparat for reiselivet (=ordliste for reiselivet). Rapporten analyserer no-situasjonen, gir ein modell for eit felles begrepsapparat, korleis arbeidet bør leggjast opp og korleis ordlista bør vedlikehaldast.

Sentrale premissar for arbeidet vårt:

- dagens organisering av begrepsapparat (les: kategoriar i tellUs-databasen) er ikkje tilstrekkeleg for å møta dei nye utfordringane i reiselivet
- begrepsapparatet må støtta deling mellom ulike reiselivsportalar og –tenester
- begrepsapparatet må også kunna integrera interessant og utfyllande informasjon tilbydd av andre enn reiselivet (t.d. kulturinformasjon)
- det må kunna integrera brukarskapt informasjon (omtalar i bloggar, ulike sosiale vevtenester)
- det må vera ope og fleksibelt
- det må vera relativt enkelt å vedlikehalda
- det må vera eit tydeleg ansvar for eigarskap og vedlikehald

Sentrale begrep (ord) brukte i rapporten:

- Begrep (en.: 'Concept'): Eit begrep er ein abstraksjon av eit fenomen. Dersom vi vel begrepet 'hotell' er det eit uttrykk for den mentale førestillinga vår av kva som er eit hotell. Eit konkret hotell, t.d. Kviknes Hotel, er ein *referent* for begrepet 'hotell'. Ordet 'hotell' er *termen*, eller namnet på begrepet. Eit begrep kan ha fleire namn ('hotell', 'hotel'). I denne rapporten er begrep den tematiske karakteriseringa av reiselivsprodukt – i vid forstand.
- Begrepsapparat: Ei samling strukturt informasjon (strukturerte metadata). Kan også noko forenkla omtalast som ei strukturert ordliste. Vokabular blir også brukt i denne rapporten om begrepsapparat
- Semantikk (gr.: "betydningslære" .. "gjelder ordenes betydning"): Læra om ordas betydning. Betydningen blir gitt i metadata som beskriv innhaldet, men også gjennom konkrete definisjonar ("eit hotell er ein overnattingsstad med døgnopen resepsjon, minimum 30 senger…)
- Semantisk vev ("semantic web"): W3C sin standard for å beskriva innhald på nettet og relasjonar mellom innhaldselement
- Emnekart ("Topic Maps"): ISO-standard som i stor grad er ein parallell til den semantiske veven. Både semantisk vev og emnekart er semantiske teknologiar
- Linked Data/Linked Open Data: Publisering av opne data på nettet og utstyrt med standardisert metadata-informasjon som følgjer reglane for den semantisk veven. Sjå vedlegg 1 for nærmare forklaring.
- Ontologi: Eit samla system med definisjonar av begrep og samanhengen mellom dei. Ein reiselivs-ontologi vil ha definisjonar av alle dei reiselivsobjekta som inngår i systemet og samanhengane mellom desse

Ordbruk:

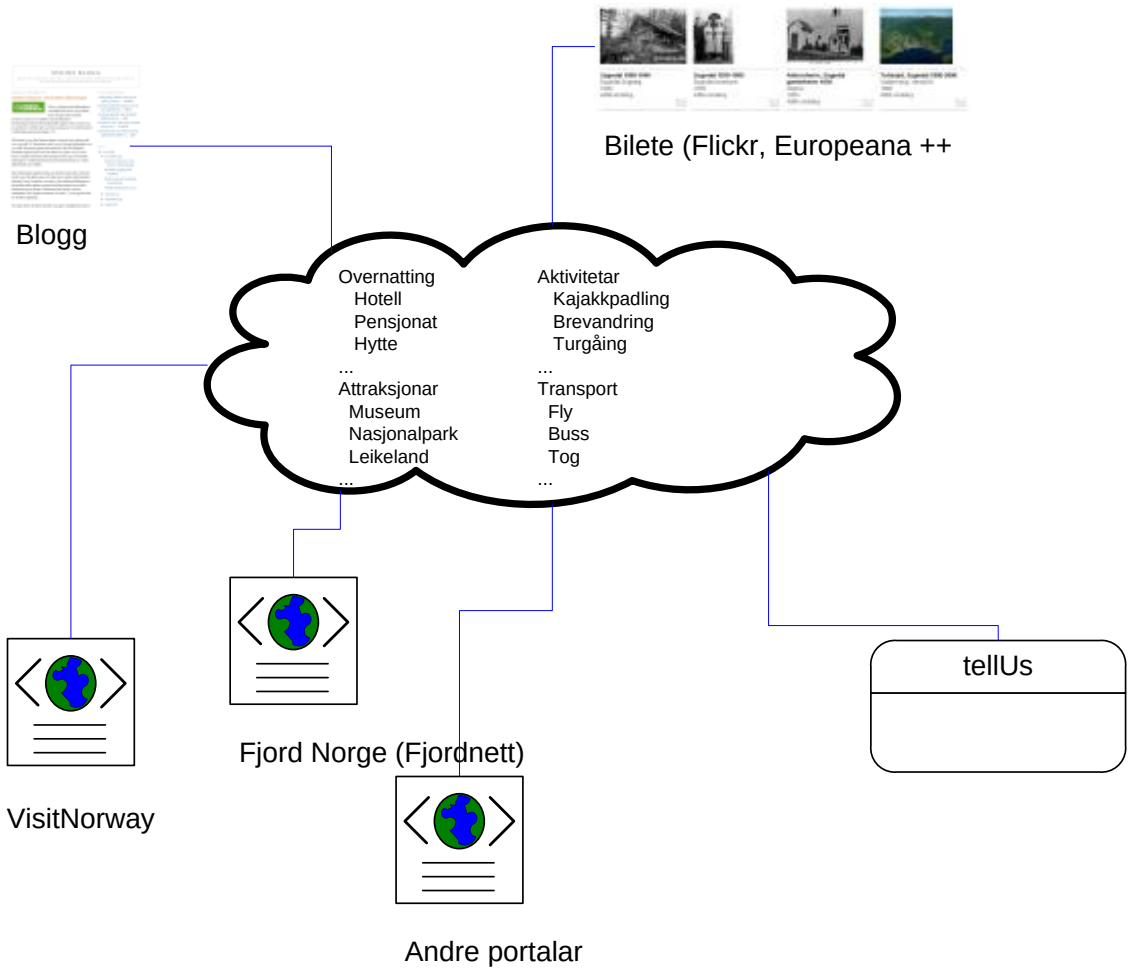
Mange "dett av lasset" så fort ord som 'ontologi', 'semantisk web', 'taksonomi' osv. blir nemnde. Denne rapporten vil i minst muleg grad bruka slike ord og heller bruka 'begrep' og 'begreps-

apparat' og dette igjen må forståast som "ei strukturert ordliste for reiselivet". I vedlegg 4 er det gjort greie for den teoretiske forankringa av arbeidet. Sjå også kap. 8 for nærmare omtale av samspelet mellom dei ulike delane av økosystemet.

3. Digitalt økosystem for reiselivet

Forslaget i denne rapporten byggjer på etablering av eit nytt begrepsapparat i "nett-skya" som skal danna basis for eit digitalt økosystem i reiselivet. Begrepa blir navet i eit økosystem der samhandling på tvers av system blir gjort muleg fordi ein brukar same namnet (begrepet) når ein snakkar om det same.
Begrepa skal ha klare definisjonar og dei skal vera lett tilgjengelege både for menneske og maskin. Det betyr at dei må presenterast både på ein lettfatteleg måte for vanlege brukarar, og på ein standardisert måte for maskiner.

Publisering av begrepsapparatet i skya må sjåast i samanheng med tilgjengeleg-gjering av begrepsapparat og data som *Linked Data (lenka data)*. *Lenka data* er ein standard for å gjera tilgjengeleg nettressursar med eit minimum av semantisk informasjon. Sjå vedlegg 1 for nærmare omtale.



Figur 1: Digitalt økosystem med begrepsapparatet i sentrum

Som figuren over viser, skal det vera ei lang rekkje ulike brukarar av begrepsapparatet. På den måten kan det skapast eit økosystem som vil gjera det muleg å utveksla informasjon på tvers uavhengig av underliggjande system. I dag er det berre muleg med utveksling om ein brukar det same publiseringssystemet. Eit felles begrepsapparat vil i seg sjølv ikkje vera nok til å kopla saman relatert informasjon, men det er eit nødvendig første steg på vegen.

Det er grunn til å understreka at opninga av begrepsapparatet i første rekkje er for å styrka arbeidet med "lausare koplingar", dvs. koplingar mellom ulike kjelder både innafor og utanfor reiselivet. Til andre oppgåver, som t.d. booking, er det behov for meir detaljert og presis informasjonsoverføring.

Begrepsapparatet blir den første byggjesteinen i økosystemet for reiselivet, og bør følgjast opp med publisering av sjølve informasjonen (dvs. omtalen av kvart enkelt reiselivsprodukt som i dag er lagra i tellUs).

4. Dagens organisering av reiselivsbegrep

Det er ikkje første gangen det blir gjort forsøk på å standardisera begrepa i reiselivet. BIT Reiseliv har arbeida med standardisering i snart 10 år, og bransjen sjølv har etablert det som kan kallast de facto-standardar (ikkje offisielle standardar, men standardar i kraft av omfattande bruk).

4.1 BIT Reiseliv og standardisering

Vestlandsforsking utarbeida i 2004 rapporten "Standardiseringsarbeidet i BIT Reiseliv" [2] på oppdrag frå BIT Reiseliv. Rapporten ga ein oversikt over status i standardiseringsarbeidet som vart starta i 2000, og kom også med tilrådingar til gjennomføring av eit hovudprosjekt.

Hovudtilrådinga i rapporten var forenkling av det felles kategori-systemet; dei føreliggjande forslaga var tenkt altfor detaljert. Det vart også tilrådd å bryta opp den strenge hierarkiske ordninga av begrep og laga ei lausare koplingssystem der eit begrep (emneord) kan høyra til fleire overkategoriar. Det er denne tankegangen vi har ført vidare i forslaget i denne rapporten.

Hovudprosjektet vart starta i 2004 og utført som ein del av fase 1 (2004-2006) av programmet. Arbeidet vart utført av Avinet as med Arne Glenn Flåten som prosjektleiar for BIT Reiseliv. Det vart utarbeidd ei xml-basert kategorisering i to nivå. Skjematisk er modellen også vist i eit UML-diagram. Dessverre har ikkje dette forslaget fått særleg gjennomslag og standardiseringsarbeidet i regi av BIT reiseliv kan ikkje seiast å vera vellykka i den forstand at næringa har teke i bruk eit felles begrepsapparat.

4.2 Standardisering og ”standardisering”

Noko av det som blir kalla standardisering i reiselivet ikkje standardisering på begrep, men på system. Gjenbruk og deling blir gjort muleg fordi systema (i stor grad portalar) er bygde på det same systemet, same leverandør. Dette er ei skinn-standardisering. Dersom ein trur ein kan få heile reiselivsbransjen til å basera alle tenester på det same systemet, frå den same leverandøren, kan tett integrering oppnåast. Men det er liten grunn til å tru at ein slik strategi vil fungera i ein større samanhang.

Fjordnett-satsinga som omfattar eit 15-tal portalar under Fjord Norge, fungerer på denne måten. Men trass i denne ”standardiseringa” er Fjord Norge/NCE Tourism likevel dei som har ivra mest for eit uavhengig, standardisert begrepsapparat for reiselivet. Det er truleg fordi dei har sett at gjenbruk og deling på bakgrunn av same IT-system ikkje er vegen å gå i den store samanhangen. Det kan fungera i mindre omfang, og kan også vera kostnadseffektivt her, men det er ikkje ein framtidsretta modell med tanke på integrasjon av ulike informasjons-kjelder.

4.3 Dagens begrepsapparat i tellUs

Dagens kategorisering (= begrepsapparat) av informasjon i tellUs-databasen er prega av ein organisk og ukontrollert vekst. Det som truleg starta med eit relativt godt definert sett av begrep, har vakse til eit uhandterleg system. Det fører til at det blir vanskeleg å kategorisera informasjonen for dei som legg inn, og det blir vanskeleg å bruka kategoriane for dei som skal ta informasjonen vidare som kundar av tellUs.

Dagens system har desse hovudkategoriar (øvste nivå):

1		Overnatting
	110	Stad
2		Servering
	203	Reisepakke
3		Forretningar – Service
36		Aktivitetar
39		Arrangement
4		Tenester
	400	Utleige
	402	Produksjon
	403	Prosjekt
	404	Matkultur
	405	Generelle typar
69		Attraksjonar
99		Transport

Det kan diskuterast om det er 8 eller 15 hovudnivå, men det er ikkje det viktigaste. Det viktigaste er at det er ei blanding av tematisk kategorisering (overnatting, servering, aktivitetar) og andre dimensjonar som *stad*, *pakking av tilbod (reisepakke)*, *administrative*

aktivitetar (prosjekt). Det blir veldig vanskeleg når så ulike kategoriar blir rista saman. Denne samanblandinga held fram også på dei to neste nivåa.

På nivå to i dagens system er det heile 800 kategoriar og på siste (tredje) nivå er det 1133.

4.4 Andre system for kategorisering

Brukarar av tellUs-informasjon kan vanskeleg gjera bruk av kategoriseringa og har utvikla sine eigne system. Dei må difor laga oversetjing (mapping) mellom tellUs og eiga kategorisering. Sjå vedlegg 2 for eit eksempel på ei oversetjing mellom VisitNorway sine kategoriar og tellUs.

Forslaget til ny organisering av begrepsapparatet gjer det lettare å innføra nye strukturar utan at den underliggjande samanbindinga blir broten. Det er fordi **bindemiddelet er emneorda**, ikkje sjølve strukturen.

5. Aktuelle standardar, begrepsapparat og verktøy

Her viser vi til ei kartlegging som vart gjort i Forskingsråds-prosjektet *Sesam4* [4], eit prosjekt for bruk av semantiske teknologiar i små og mellomstore bedrifter. Dei viktigaste standardane omtalte i denne rapporten, er:

5.1 Utvekslingsstandardar

- OTA (Open Travel Association)
- WTO Thesaurus
- ISO 18513:2003
- CEN/TC 329
- Enjoy Europe/MDS (Minimum Data Set)
- HEDNA
- IFITT RMSIG (Reference Model Special Interest Group)

Av desse er OTA den mest brukte og mest aktuelle. Open Travel Association er ein non-profitt-organisasjon etablert av reiselivsbedrifter i 1999. Føremålet er å utvikla og vedlikehalda eit bibliotek av XML-skjema (XML Schemas) til bruk i reiselivet. Desse skjema utgjer OpenTravelXML-spesifikasjonane.

OTA vart starta av store tur- og transportaktørar, m.a. dei store flyselskapa, hotella og bilutleige-firma, og har difor hatt stort fokus på flyreiser, leigebil og overnatting. Etter kvart har standardiseringa omfatta fleire område innan reiselivet. Standarden blir vedlikehalden og publisert to gonger i året. Den siste versjonen, 2008B, inneheld ikkje mindre enn 626 ulike XML-filer som beskriv ulike reiselivsområde og –handlingar.

Desse hovudområda er dekka:

airline booking (and checking in etc.)
cruise booking etc.
destination activities
dynamic packages (compound bookings)
hotel booking etc.
golf course
travel insurance
railway bookings

BIT Reiseliv har sidan 2009 arbeida med etableringa av nasjonal bookingteneste, eit arbeid som vart starta tidlegare av andre aktørar. BookNorway AS er no etablert for å driva booking-verksemda på kommersielt grunnlag. Det nye bookingsystemet er ikkje ei ny bookingløysing i seg sjølv, men ei samkøyring av fleire eksisterande løysingar (som City Break, tellUs GuestMaker og Restech). Løysinga skal vera operativ til sesongen 2011, men den offisielle lanseringa skjer først i september i år.

Ein viktig del av arbeidet med nasjonal booking har vore standardisering av omtalen av reiselivsprodukta i tillegg til auka kvalitet på foto. Kategoriseringa av reiselivsprodukta følgjer i stor grad OTA-standarden [3]. Arbiedet med omarbeiding av tellUs-kategoriane må difor sjå nøye på kategoriseringa som er gjort i samband med tilpassing av bookingtenesta.
OTA er skreddarsydd for booking og er som tidlegare nemnt omfangsrikt. Det er eit detaljnivå som er på grensa til det komiske, så i bruken av standarden er det viktig å sjå på dei meir generelle begrepa . Eksempelet under viser detaljnivå i OTA, her frå OTA_CodeTable:

```
<Code Value=”253” CreationDate=”2008-11-04”>
<CodeContents>
<CodeContent Language=”en-us” Name=”DVD player available at front
desk”/>
</CodeContents>
</Code>
```

5.2 Ontologiar for reiselivet

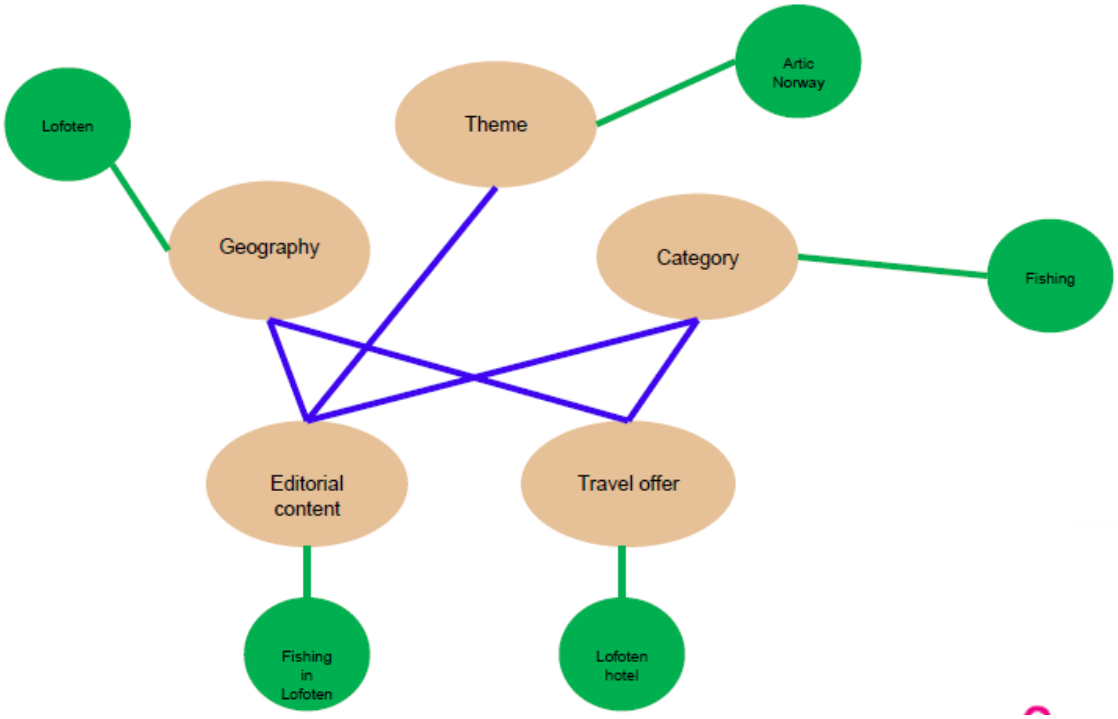
- Harmonise/HarmoNet Tourism Ontology
- Semantically Interlinked Online Communities Project (SIOC)

Hovudinntrykket er at ingen av dei ovannemnde standardane har fått særleg utbreiing med unnatak av OTA. OTA er på si side svært detaljert og bransjespesifikk, prega av hovudaktørane bak initiativet (flyselskap, bilutleigefirma og hotell).

Forskingsrådsprosjektet SeSam4 har utarbeidd ein oversikt over informasjonsressursar og standardar for reiselivet [4].

VisitNorway

VisitNorway sin portal er basert på teknologi-standarden emnekart (Topic Maps). Det er utvikla ein enkel ontologi for portalen :



Figur 2: Ontologi for VisitNorway (henta frå Networked Planet som saman med MakingWaves er firma bak nettsidene)

Ontologien har *kategori, tema* og *geografi* som dei sentrale elementa. Kategori er dei oversette (mappa) tellUs-kategoriane, medan tema er t.d. 'Arctic Norway', 'Family holiday' og 'National parks'. Geografi er sjølvsagt lokalisering, eit heilt avgjerande element. *Travel offer* er det aktuelle reiselivs-produktet (det aktuelle hotellet, den aktuelle aktiviteten osv.) og er sjølvsagt også sentralt i modellen. *Editorial content* er artiklar produserte av VisitNorway-redaksjonen sjølve.

Sesam4-ontologi

I prosjektet *Sesam4* har reiseliv vore det sentrale utprøvingsområdet for semantiske teknologiar. Også her har vi utvikla ein enkel ontologi der vi har henta inspirasjon frå eksisterande ontologiar som VisitNorway og andre. Sjå vedlegg 3 for detaljar i Sesam4-ontologien.

5.3 Verktøy for arbeid med begrep

For å unngå å hamna i same problema som dagens tematiske kategorisering, er det viktig å arbeida med begrepsorganisering på ein systematisk måte. Då er det viktig å kjenna til dei grunnleggjande byggjesteinane og ein framgangsmåte for arbeidet.

Språkrådet har gitt ut "Termlosen – Kort innføring i begrepsanalyse og terminologiarbeid" [6]som ein kort rettleiar i slikt arbeid. Denne bør vera utgangspunktet for begrepsarbeid.

6. Forslag til ny organisering av begrepsapparatet

I forslaget til ny organisering av begrepsapparatet for reiselivet, har vi teke utgangspunkt i dagens tellUs-kategorisering, samanlikna med andre system, og til slutt komme fram til ein modell vi trur både er framtidsretta og realistisk å gjennomføra. Prinsippa for den føreslåtte organiseringa, er:

1. **Forenkling**
   Tal kategoriar og nivå bør reduserast.

2. **Kontrollert vokabular med vekt på emneord**
   Vi føreslår eit kategorisystem basert på *emneord* og med ein tilhøyrande overliggjande struktur i tillegg til eit nivå under kalla *hjelpeord*. Denne modellen er inspirert av systemet *Los*[2], eit system for kategorisering av offentlege tenester utvikla av Vestlandsforsking [5] for Norge.no (no Difi). Det er eit system for kategorisering av offentlege tenester og er i bruk i bortimot 150 kommunar. Det er såleis eit gjennomprøvd system som også har vist at det kan skalerast med tanke på brukarar. Difi (Direktoratet for ikt og e-forvaltning) er ansvarleg for Los, og syter for oppdatering av vokabularet i tillegg til utvekslingsrutinar.

   Formelt sett er denne organiseringa av begrep ein *thesaurus*[3]. Ein thesaurus er ei vidare-utvikling av den enklare *taksonomien* (dagens tellUs-kategorisering er ein slags taksonomi), med fast definerte koplingar/relasjonar. Dei viktigaste relasjonane i ein thesaurus er:

   BT (Broader Term): Meir generell term (breiare)
   NT (Narrower Term): Meir spesifikk term (smalare)
   RT (Related Term): Relatert term (dvs. emneord som er relaterte på eit vis)
   SN (Scope Note): Ein definisjon av termen (emneordet)
   U (USE): Eksempel: 'biltilsyn' USE 'trafikkstasjon'
   UF (USED FOR): Den omvendte vegen: 'trafikkstasjon' UF 'biltilsyn'

   - Emneorda er sentrale og nettressursar blir kopla til desse. tellUs får eit spesielt ansvar for vedlikehald av emneord.
   - Hjelpeorda er knytte til emneorda og kan opprettast fritt. Hjelpeord kan vera synonym, utgåtte begrep, smalare begrep o.l. Hjelpeorda vil styra brukaren mot det rette emneordet.
   - Temastrukturen over emneorda kan også variera frå brukar til brukar, litt slik som dagens strukturar varierer frå tellUs til VisitNorway, til FjordNorge og til andre. tellUs sitt forslag til temastruktur kan brukast av andre, men dei kan definera sin eigen som kan vera betre tilpassa situasjonen (den aktuelle portalen, applikasjonen m.v.). Emneorda blir dei same, det er på dette nivået deling av informasjon føregår.

3. **Alt kan ikkje løysast med kategorisering**
   Vi trur ikkje det er muleg å definera ei kategorisering eller eit begrepsapparat som løyser alle

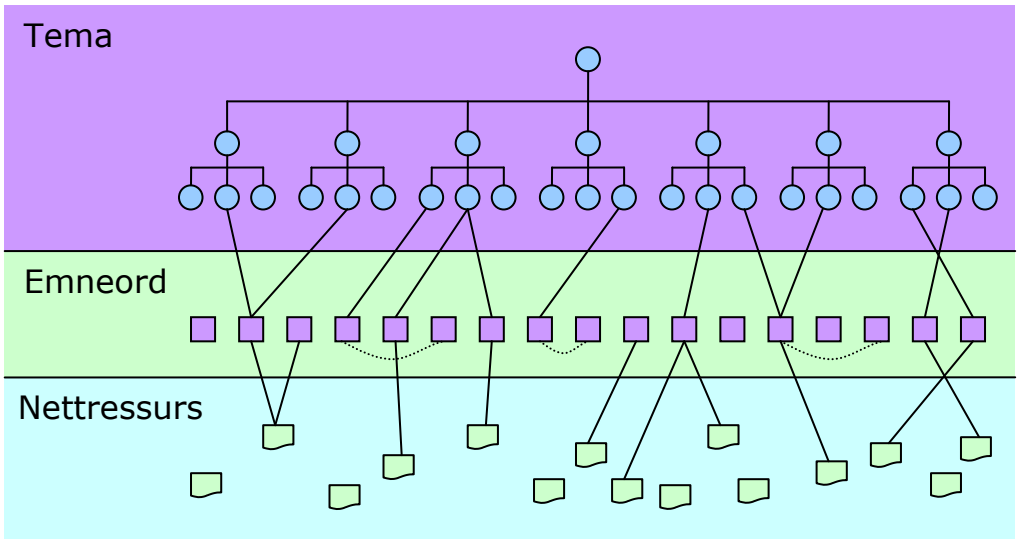
[2] Meir informasjon om Los på http://los.difi.no
[3] ANSI/NISO Z39.19 frå 2005 er ein standard for ein eittpråkleg (monolingual) thesaurus. For meir informasjon, sjå http://bit.ly/z39_19

problem. Eit begrepsapparat som skissert her, er **nødvendig, men ikkje tilstrekkeleg**. Tilrettelegging for søk, bruk av sosiale medium m.m. er nødvendige tilleggsfunksjonar for ei godt fungerande teneste.

4. **Publisering av emneord i "nett-skya"**
   Forslaget inneber at tellUs må opna opp begrepsapparatet og dela det med andre. Vi føreslår at heile begrepsapparatet blir publisert i "skya", tilgjengeleg for kven som helst. Det er viktig å ha klart for seg at deling av ressursar ved hjelp av kategoriane som blir publiserte i "skya" vil vera av typen "laus kopling". Dei systema som krev større presisjon og fleire detaljar, t.d. booking, må bruka fastare koplingar med meir detaljert informasjon (som dagens overførins-metodar). Vedlikehaldsrutinane må likevel vera tydeleg definerte, sjå slutten av kapitlet.

5. **Fleire format**
   Vokabularet (begrepsapparatet) må vera tilgjengeleg i fleire format. Det bør vera tilgjengeleg i menneskeleseleg format som HTML, i maskinlesbart format som t.d. XML, og helst også uttrykt i semantiske standardar som SKOS/RDF eller XTM.

7. **Publisering av informasjon i "nett-skya"**
   For å få full effekt av den nye strategien bør også informasjonsressursane i tellUs-basen publiserast i skya. Då vil heile dagens lukka tellUs-database bli "vrengt ut" og synleg for alle. Det vil bli fundamentet i det digitale reiselivs-økosystemet. Det vil kunna skapa eit godt grunnlag for ei rik og kreativ utvikling med mange applikasjonar og tenester som vi i dag ikkje har tenkt på. Det blir ein parallell til den rivande utviklinga som skjer innan opne data ("Linked Open Data"[4]).



Figur 3: Modell for ny organisering av tellUs-begrep (NB! At berre enkelte nettressurar er kopla til emneord i eksempelet, er tilfeldig)

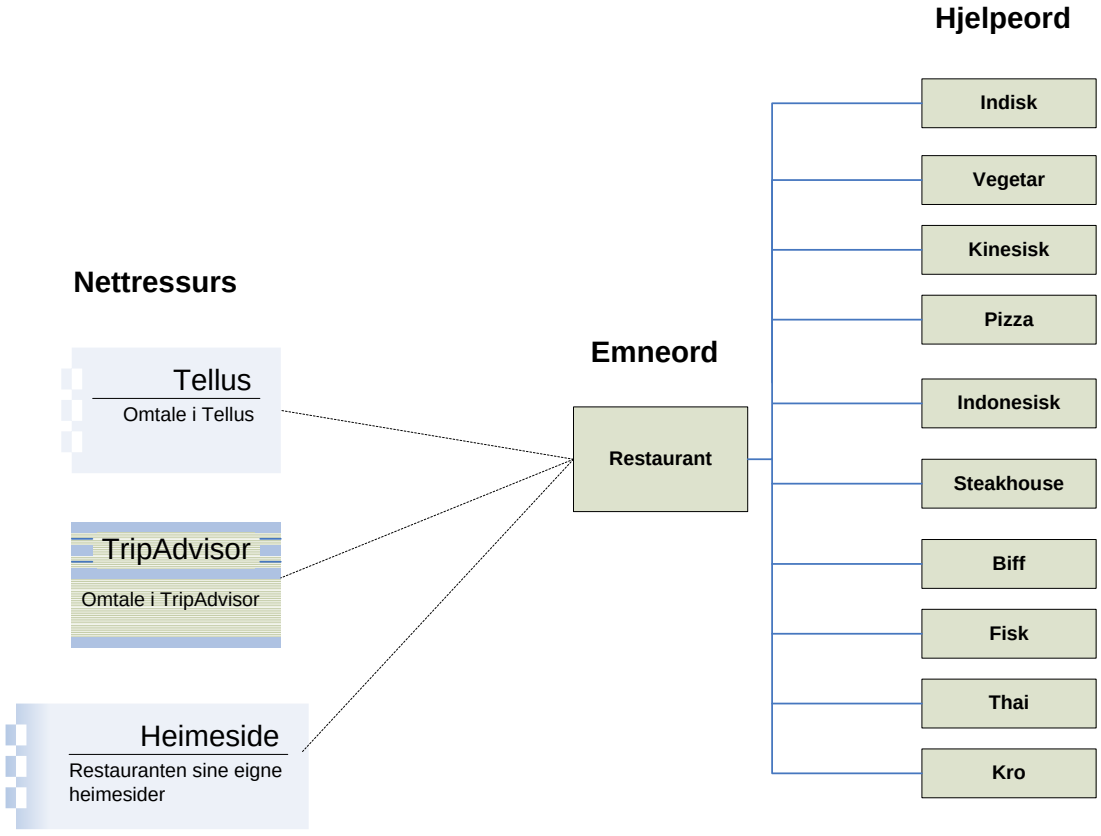
[4] Sjå http://linkeddata.org

Illustrasjonen over viser systemet Los, men reiselivsbegrepa både kan og bør organiserast på same måten. Sentralt i modellen er **emneorda**, det er dei som er bindemiddelet (limet) i samhandlinga. Det er dei som må delast og gjenbrukast skal ein få til samhandling.

Til emneorda er det knytt ein tonivå-struktur (**tema**). Temastrukturen kan ha så mange nivå ein vil, men det beste vil vera å prøva å halda den nokså flat. Strukturen kan også variera frå brukar til brukar, om lag som i dag der VisitNorway og FjordNorway har andre, eigenutvikla strukturar enn tellUs, sjølv om data er like. Men i motsetnad til dagens tellUs-struktur, er det ikkje strukturen som blir den sentrale, men emneorda. Så lenge dei er like, spelar det inga rolle kva struktur som er brukt over.

Emneorda blir knytte til nettressursar, som er omtale av reiselivsprodukt. I dag ligg nettressursane i ein database, og er ikkje direkte referer-bare. Det må endrast for å få full nytte av den nye organiseringa.

Til emneorda er det knytt ei rekkje **hjelpeord** (ikkje vist i figuren over, men figuren under). 'Hjelpeord' er eit samlebegrep for ord som 'synonym', 'utgåtte termar' osv. Hjelpeorda skal som ordet seier, hjelpa brukaren til å finna fram til det rette emneordet, og her er brukaren både sluttbrukaren som leitar etter reislivsinformasjon og den som legg inn informasjon i tellUs og kategoriserer denne. Hjelpeorda vil også vera viktige i søk, men det føreset at dei følgjer nettressursen, dvs. blir lagt inn i nettressursen saman med emneordet/emneorda.



Figur 4: Samanhengen mellom hjelpeord, emneord og nettressurs (eks. frå dagens org.)

I eksempelet over er 'restaurant' emneordet som blir knytt til nettressursane frå tellUs, TripAdvisor og restauranten si eiga heimeside. Til emneordet 'restaurant' er det knytt ei rekkje hjelpeord. Ved søk på eitt av desse orda, vil brukaren bli leia til emneordet med tilhøyrande ressursar.

Medan emneorda må haldast under kontroll (det kontrollerte vokabularet), kan og bør hjelpeord sleppast fritt. Kvar brukar bør kunna leggja til sine eigne hjelpeord til dei felles emneorda som blir henta frå "skya". På den måten får ein fanga opp lokale spesialitetar og særmerkje.

Utvikling av nytt begrepsapparat

Utviklingsarbeidet må ta utgangspunkt i dagens tellUs-kategoriar og halda det opp mot andre aktuelle begrepsapparat og variantar av tellUs-begrepa. Dette bør styrast av tellUs, men utførast i lag med sentrale brukarar, både på innlegging av data og distribusjon av data.

Arbeidet bør følgja dei prinsippa og forslaga som er føreslegne over. Arbeidet med begrepsapparatet bør skje uavhengig av teknologisk fundament, dette kan godt gjerast med Excel eller liknande verktøy.

Vedlikehald av begrepsapparatet

Når forslaget er å organisera begrepa som eit kontrollert vokabular, ligg det i namnet at det må kontrollerast av nokon. Nokon er i dette tilfellet tellUs. Slik dette oppdraget er forma og slik situasjonen er i dag, må tellUs ta ansvaret for vedlikehaldet av begrepsapparatet fordi det først og fremst blir eit hjelpemiddel for dei.

Men med opning av begrepsapparatet må næringa elles også bidra. Det betyr at det må vera mulegheiter for å komma med endringsforslag frå alle andre som brukar begrepsapparatet.

I tilknyting til publisering av begrepsapparatet i skya, bør det difor også etablerast rutinar for tilbakemelding om nye emneord og endring av emneord.

7. Distribusjonsmodell for begrepsapparatet

Begrepsapparatet må formidlast på fleire format, leseleg både for menneske og maskin. For menneskeleg tilgang er publisering som HTML det beste, men det bør i tillegg også føreliggja som rein tekst slik at det lett kan takast i bruk i andre samanhengar. For maskinell lesing er XML det føretrekte formatet. Utvekslinga kan skje både via filer og via web services.

7.1 Eksport – import av XML-filer

Den enklaste forma for distribusjon av begrepsapparatet er ved hjelp av eksport eller import av XML-filer som blir tilgjengelege via nettet (via ein URL). Filene kan omfatta delar eller heile begrepsapparatet. Det er naturleg å kunna velja berre emneorda, sidan brukarane kan ha eigne temastrukturar og eigne hjelpeord knytte til emneorda.

Formatet på filene bør kunne støtte ulike standarar som Topic Maps (XTM) og RDF-triplar avhengig av brukaren sine behov.

7.2 Web services

Eit alternativ til import av xml-filer, som skissert over, er å henta ut begrepsapparatet ved hjelp av web services. Dette alternativet krev ikkje spesiell installering eller tilrettelegging lokalt, berre at brukaren gjer bruk av definert web service og resultata frå den.

Web service-en tek seg av alle oppslag mot den sentrale "Begrepsdatabasen". Web service-en vil ha fleire prosedyrar/funksjonar som returnerer tema, undertema, emneord og hjelpeord. Eksempel på funksjonar er:

- HentTema (som returnerer tema med tilhøyrande undertema)
- HentEmneord (returnerer alle emneord tilhøyrande eit undertema)
- HentHjelpeord (returnerer alle hjelpeord tilhøyrande eit emneord)

Men den største gevinsten får vi når nettressursane også blir tilgjengeleg i "skya". Då vil vi også kunna ha spørjingar som hentar alle nettressursar knytte til eit emneord, eller nettressursar med gitt emneord i eit geografisk område.

Web service-en bør støtta REST-prinsippet[5]. Det vil sei at URL-en til Web service-en blir lesbar og fortel kva vi *filtrerer* på, og kva som skal *returnerast* .

8. Samspelet mellom elementa i "økosystemet"

Forslaga i denne rapporten er forankra i idéen om å skapa eit digitalt økosystem for reiselivet. Begrepsapparatet, og helst også etter kvart publiserte informasjonsressursar frå tellUs-databasen, skal danna basisen i systemet. Tanken er at dette fundamentet skal gi næring til nye og kreative løysingar, både tradisjonelle web-baserte løysingar og mobile applikasjonar ('apps'). Til saman vil dette utgjera eit digitalt økosystem.

Til grunn for forslaget om etablering av eit digitalt økosystem for reiselivet, ligg nokre grunn-leggjande tankar om web-en generelt, og om opne data spesielt.

8.1 7 råd for å "tenkja web"

Med uttrykket å "tenkja web" meiner vi her å få ei djupare forståing for korleis web-en fungerer og korleis ein kan støtta dei grunnleggjande prinsippa i den. Råda spesifiserte under, er henta frå Jon Udell[6], Microsoft. Fleire av desse råda heng saman med råda om opne data omtala litt

[5] REST = Representational State Transfer; enkelt sagt ein beskrivelse av korleis web-en slik vi kjenner den, fungerer (sjå meir forklaring på http://en.wikipedia.org/wiki/Representational_State_Transfer

[6] http://blog.jonudell.net/2011/01/24/seven-ways-to-think-like-the-web/
Jon Udell er ein kjend web-ekspert som m.a. lenge var ein viktig spaltist i tidsskriftet BYTE.

seinare. Dei er også sentrale for motivasjonen som må liggja til grunn for dei forslaga vi set fram i denne rapporten.

1. **Ver den autoriative kjelda for eigne data**
   (”Be the authoritative source for your own data”)
   Også NRKBeta har lansert eit liknande slagord (kalla *NRKBeta-doktrinen*) i sitt arbeid for å gi slepp på data samstundes som ein ikkje reduserer seg sjølv og sin eigen autoritet. Det handlar om å tora å dela data, men på same tid gjera det tydeleg kven som står bak dei.

2. **Overfør ved referanse, ikkje verdi**
   (”Pass by reference, not by value”)
   Eit eksempel: Når du sender ein URL til nokon, overfører du ved hjelp av referanse, ikkje verdi. Alternativet er å kopiera innhaldet og senda det (som verdi)

3. **Ver merksam på skilnaden på strukturerte og ustrukturerte data**
   (”Know the difference between structured and unstructured data”)
   Dersom du viser ein kalendar som ei rein HTML-side, eller verre: som ei pdf-fil, viser du ustrukturerte data. Dersom du publiserer kalendaren din som iCalendar (standard for kalendarinformasjon), sender du strukturerte data som maskinene kan tolka.

4. **Bruk disiplinerte namne-konvensjonar**
   (”Create and adopt disciplined naming conventions”)
   Tydeleg namngiving lettar seinare bearbeiding og aggregering av data. Eit eksempel er t.d. å oppgi eit tydeleg Twitter-emneord (hashtag) for eit arrangement.

5. **Distribuer data til så mange som muleg**
   (”Push your data to the widest appropriate scope”)
   Sender du ein epost, når den ein eller fleire personar. Skriv du eit blogginnlegg på eit intranett, når den alle i bedrifta. Skriv du eit blogginnlegg på web-en, når den potensielt heile verda.

6. **Delta i nettverk både som bidragsytar og lesar**
   (”Participate in pub/sub networks as both a publisher and a subscriber”)
   Igjen er blogg-sfæren det beste eksempelet. Diskusjonane i kjølvatnet av eit innlegg er det som gir meirverdi til den opprinnelege informasjonen.

7. **Gjenbruk komponentar og tenester**
   (”Reuse components and services”)
   For å realisera web-arkitekturen ”small pieces loosely joined” er dette nødvendig. Mange av dei mest suksessrike applikasjonane/tenestene på nettet er utvikla etter denne modellen.

8.2 7 råd for opne data

Dersom informasjonsressursane frå tellUs-basen også blir publiserte ope for alle, vil det kunna skje med utgangspunkt i standarden *Linked Data* og på den måten koplast til mange andre opne data. Linked Data, eller Linked Open Data (LOD) som det også blir kalla, kan sjåast på som ein lettvariant av semantisk web. Den er langt mindre ambisiøs enn mykje av det ontologiarbeidet som har skjedd innafor den semantiske web-en og som i mange tilfelle har vore i overkant ambisiøst og for lite forankra i dei faktiske it-behova i samfunnet.

Med lenka data har ein teke eit steg tilbake og prøvt å finna eit minimum av tilleggsinformasjon som skal til for å få ei meiningsfull, og likevel automatisk, utveksling av informasjon. Her er 7 gode råd for opning av data og tilnærming til lenka data:

1. **Bruk standard internett-protokollar (http) for tilgang til nettressursar**

2. **Alle objekt må ha ein unik identifikator (URI)**

   Adressa (URI-en) må vera permanent (persistent) og kunna lesast både maskinelt og av menneske

3. **Unngå aggregering av data**

   Data er best som råkost, ikkje aggreger utan at det er heilt nødvendig

4. **Strukturer metadata på ein maskinlesbar måte (t.d. XML-serialisering eller XML/RDF/XTM)**

5. **Bruk internasjonalt teiknsett (UTF-8)**

6. **Bruk minimum Dublin Core som global standard for beskrivelse av data (metadata)**

7. **Tenk på kopling mot andre datakjelder ved å tilretteleggja for Linked Data (sjå vedlegg)**

8.3 Felles begrepsapparat

Med eit fritt tilgjengeleg begrepsapparat og tilkopla informasjonsressursar er utgangspunktet det beste for ei innovativ utvikling av nye tenester og applikasjonar. Eit felles begrepsapparat som blir brukt av fleire aktørar gjer det muleg å utveksla ressursar basert på dei felles emneorda. Dette må kombinerast med informasjon om t.d. geografi for å få den delte informasjonen presis nok.

Utviklinga av eit digitalt økosystem for reiselivet bør starta med ein gjennomgang av dagens begrepsapparat, med utgangspunkt i tellUs-kategoriane. Næringa må finna fram til passande emneord, etter modellen som er føreslått i denne rapporten. Begrepsapparatet må gjerast ope

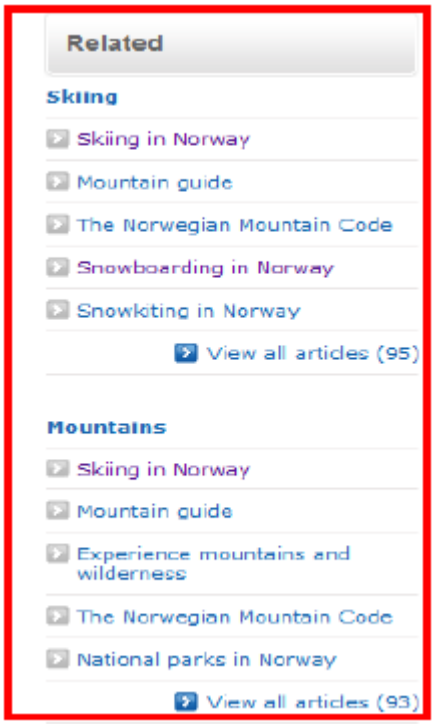
tilgjengeleg slik at det blir lett å bruka for alle interesserte. Det vil også gi nyttige tilbakemeldingar slik at det kan bli kontinuerleg oppdatert og forbetra.

8.4 Portalar og applikasjonar

Det digitale økosystemet skal gjera det lettare å kopla informasjon frå ulike kjelder ved at dei same emneorda blir brukte til å kategorisera ulike ressursar. I ein tradisjonell portal kan det gjerast ved at relatert informasjon blir vist i høgremargen som i eksempelet under:

I applikasjonar for smarttelefonar vil ein kunna utnytta den opne informasjonen ved å kopla tema og lokalisering og presentera på eit kart. Det vil bli veldig lett å laga enkle applikasjonar som t.d. hentar ut relevante aktivitetar i ein viss omkrins frå der du er akkurat i augneblinken.

Byggjesteinane vil stimulera til kreativ utnytting av tilgjengelege ressursar slik at reiselivsinformasjonen kan bli meir brukt og nå ut til fleire potensielle besøkande. Det er å håpa at fundamentet kan skapa eit mylder av applikasjonar som til saman blir det nødvendige økosystemet for digital reiselivs-informasjon.



Figur 5: Relatert informasjon, eks. frå VisitNorway

9. Oppsummering og arbeidet vidare

Dette er ikkje første gangen det blir gjort forsøk på å etablera eit felles begrepsapparat for reiselivet, og det blir kanskje ikkje den siste heller. Det er ein vanskeleg jobb å setja namn på ting, og endå vanskelegare når det er fleire uavhengige aktørar med i biletet. Likevel er det ein nødvendig del dersom ein skal oppnå betre integrasjon og lettare informasjonsutveksling.

I denne forprosjektrapporten er det skissert forslag til utvikling av eit felles begrepsapparat med utgangspunkt i tellUs sine eksisterande kategoriar. Begrepsapparatet omfattar den **tematiske** delen av reiselivsinformasjonen, ikkje andre dimensjonar som t.d. geografi, tid m.m.

Emneord som det sentrale

Forslaget byggjer på å erstatta dagens hierarkiske kategorisystem med **emneord**. Hierarkiet vil ikkje bli borte, men det er emneorda som vil stå i sentrum, den tematiske strukturen over emneorda er sekundær. Til emneorda føreslår vi å kopla **hjelpeord** for lettare å navigera seg fram til korrekt emneord. Hjelpeorda vil vera til hjelp både for dei som legg inn informasjon, og for sluttbrukarane i søkesamanheng.

Denne måten å ordna begrep på, er prøvd ut i andre prosjekt, på andre område. Mest nærliggjande er kategoriseringssystemet *Los*, eit system for kategorisering av offentlege tenester. Det er eit system eigd av Difi, og brukt i bortimot 150 kommunar for å binda saman nær-slekta informasjonsressursar. Det viser at metoden duger i praksis, og er skalerbar ved at den kan handtera ein aukande brukarar-masse utan problem.

Vi har brukt 'begrep' og 'begrepsapparat' i denne rapporten som uttrykk for inndeling av reiselivs-produkt i tematiske kategoriar. Vi har vidare gjort ei forenkling og snakka om ei tematisk **ordliste**, medan 'begrep' i vitskapleg forstand er meir komplisert. Vedlegg 4 om "Ontology Construction: Background and Practices" viser ei vitskapleg framstillinga av begrep (jfr. det semiotiske triangelet) og relasjonar mellom begrep forstått som ein ontologi.

Publisering av emneord i "nettskya"

Vi føreslår vidare at emneorda blir publiserte i "nett-skya", lett tilgjengeleg for alle som vil bruka dei. Emneorda må kontrollerast og ha eit system for jamleg vedlikehald (= kontrollert vokabular), medan hjelpeorda er friare og bør kunna leggjast til av alle som ønskjer det.

Dette blir det første, viktige steget i ei utvikling mot *lenka data* (*Linked Data*). Men for å realisera lenka data, må også informasjonsressursane emneorda blir knytte opp mot, publiserast ope. Det blir ei naturleg vidareføring når det kontrollerte vokabularet er på plass.

Til saman vil dette bli viktige byggjesteinar i etableringa av eit digitalt **økosystem for reiselivet**. Ved hjelp av byggjesteinane omtalte ovanfor, skal tredjeparts-aktørar lettare kunna utvikla nye system (applikasjoanr, nettsider) for reiselivet og lettare kunna integrera reiselivsinformasjon med tilgrensande informasjon (kulturbasert informasjon, brukargenerert informasjon m.m.).

Vidare framdrift

Vi tilrår at det vidare arbeidet blir drive av **tellUs** i tett samarbeid med reiselivsaktørar som VisitNorway, FjordNorway/NCE Tourism, Book Norway m. fl. Det er fleire igangsette og planlagde

prosjekt som har delar av forslaga her innebygd. Det gjeld særleg tellUs' Skattefunn-prosjekt som alt er godkjent og sett i gang, og det gjeld prosjektsøknad til nyleg utlysing i Forskingsrådet sitt Verdikt-program.

Vi føreslår at det vidare arbeidet blir integrert i desse prosjekta. Ein stor fordel vil vera at dette delvis alt er igangsette prosjekt og at vidareføringa dermed kan ta til utan opphald. Ein annan fordel er at arbeidet då vil bli godt innpassa i forretningsplanar og dermed godt forankra i næringa sine behov. Dersom prosjektsøknaden til Verdikt blir støtta, vil det også gi eit viktig, ekstra forskingsbidrag til arbeidet.

Figuren på neste side skisserer ein framdriftsplan for arbeidet ut frå denne tankegangen.

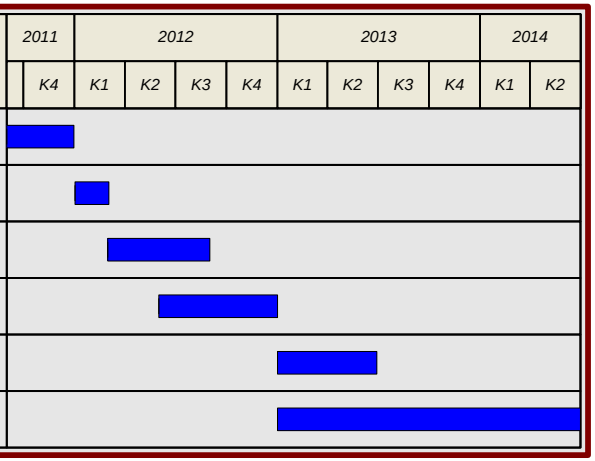
Framdriftsplan

tellUs har planlagt eit arbeid internt der ein gjennomgang og opprydding i begrepsapparatet er sentralt. Det er eit arbeid i tråd med forslaga i rapporten og tellUs-prosjektet er den mest sannsynlege vidareføringa av arbeidet, som nemnt over.

I tillegg til tellUs' internprosjekt er det også søkt om midlar frå Forskingsrådet sitt Verdikt-program til eit prosjekt for å etablera det omtalte digitale økosystemet. Vestlandsforsking har vore sentrale i utforminga av søknaden der tellUs står som søkjar. Andre samarbeidspartnarar i prosjektsøknaden er Norsk Regnesentral, Computas, NCE Tourism/Fjord Norge, Visit Sognefjord og Fylkesarkivet i Sogn og Fjordane. Prosjektet vil ta opp sentrale delar av tellUs' interne prosjekt og arbeida vidare med forslaga i denne rapporten.

Forslag til plan for realisering av det digitale økosystemet for reiselivet følgjer i store trekk tellUs-prosjektet "*Relasjonsbasert konsept for en interaktiv reiselivsnæring*" og prosjektplan for *CREAPURE* (Verdikt-søknad sendt til Forskingsrådet 16.02.11).

ID	*Aktivitetsnamn*	*Start*	*Slutt*	*Lengde*
1	Utarbeiding av nytt begrepsapparat	01.09.2011	30.12.2011	17,4u
2	Publisering av begr.app. i skya	02.01.2012	02.03.2012	9u
3	Utvikling av rammeverk for LOD	01.03.2012	31.08.2012	26,4u
4	Implementering av LOD i tellUs	01.06.2012	31.12.2012	30,4u
5	Testing og validering av LOD + pilot	01.01.2013	28.06.2013	25,8u
6	Innføring av ny forr.modell, tellUs	01.01.2013	30.06.2014	78u



Referansar

[1] NCE Tourism: ”Strukturerte data og reiselivsinformasjon”

[2] Haugen, Oluf, Ingjerd Skogseid og Svein Ølnes: ”Standardiseringsarbeidet i BIT Reiseliv” (VF-notat 7/2004) - http://www.vestforsk.no/filearchive/vf-notat7-2004_1.pdf

[3] OTA: Open Travel Alliance, http://www.opentravel.org/Specifications/Default.aspx

[4] SeSam4: www.sesam4.net

[5] Ølnes, Svein: Nye LivsIT – Forslag til ny informasjonsstruktur for LivsIT, VF-rapport 2/2005, http://www.vestforsk.no/rapport/nye-livsit-forslag-til-ny-informasjonsstruktur-for-livsit

[6] Suonuuti, Heidi: ”Termlosen – Kort innføring i begrepsanalyse og terminologiarbeid”, Språkrådet 2010.

Vedlegg 1: Linked Data

Linked Data og *Linked Open Data* er to uttrykk som blir stadig meir brukte. Det representerer ei utvikling mot enklare semantiske løysingar. Semantisk web har vore spådd som det neste store etter framveksten av web 2.0 og sosiale media og har då også fått tilnamnet *web3.0*. Men det har teke lang tid for semantiske teknologiar å få fotfeste, og mykje av grunnen er at det blir for komplisert for mange å setja seg inn i logikk og ontologi-utvikling. LD/LOD tilbyr ein enklare måte å oppnå mykje av det som den semantiske web-en har lova.

Linked Data, på norsk *lenka data*, er ei samling beste praksisar for publisering og kopling av strukturerte data på web-en. Til saman utgjer desse initiativa *the Web of Data*. Det som skil desse koplingane (lenkene) frå tradisjonelle hyperlenker, er at dei er utstyrte med informasjon om *type*. Ei vanleg lenke på web-en seier ikkje noko om kva type informasjonsressursen som lenka peiker til, er. Med *lenka data* vil vi i tillegg få opplysningar om kva *type* data det blir lenka til. Fordelen med denne teknologien er at data lenka på denne måten blir maskinleseleg og det opnar for ei rekkje nye bruks-mulegheiter.

Lenka data brukar standarden RDF[7] for å uttrykkja *type* data. Resultatet blir *Web of Data*, eller meir presist *web of things in the world, described by data on the Web*[8]. RDF er ein standard for å uttrykkja semantisk innhald, og det blir gjort ved hjelp av såkalla *triplettar*. Ein triplett er eit utsagn (statement) på forma *subjekt – predikat – objekt*. Alle data i *lenka data* blir beskrivne på denne måten. Både subjekt og objekt er URI-ar som representerer ressursar, dvs. adresser som kan nåast ved hjelp av HTTP-protokollen. Predikatet seier noko om korleis dei to ressursane er relaterte, det er her *typen* relasjon blir definert.

Subjekt	Predikat	Objekt
Den aktuelle ressursen	*er av typen*	*by*
http://dbpedia.org/page/oslo	*http://www.w3c.org/.../22-rdf-syntax-ns#type*	*http://dbpedia.org/ontology/city*

Subjekt	Predikat	Objekt
Den aktuelle ressursen	*har namnet*	*"Oslo"*
http://dbpedia.org/page/oslo	*http://www.w3c.org/.../rdf-schema#label*	

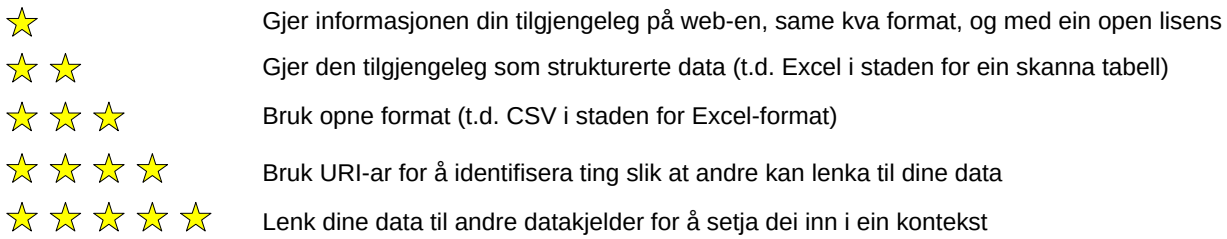
Figur 6: Eksempel[9] på to triplettar som seier noko om Oslo

[7] RDF = Resource Description Framework, ein W3C-standard for å beskriva semantisk informasjon. Sentralt i RDF er *triplettar* – subjekt, predikat og objekt. Alle utsagn om ein ting, blir gjort med triplettar.
[8] Tim Berners-Lee et al. (2009): International Journal On Semantic Web and Information Systems, vol. 5, nr. 3
[9] Eksempelet er henta frå http://datavisualization.ch/opinions/introduction-to-linked-data

Tim Berners-Lee[10] har formulert desse enkle reglane for å publisera data på web-en:

1. Bruk URI som namn på ting
2. Bruk HTTP-URIar slik at menneske kan slå opp namna
3. Når nokon hentar informasjon om ein URI, gi nyttig informasjon ved hjelp av standardane (RDF og SPARQL)
4. Inkluder lenker til andre URI-ar slik at fleire ting kan bli oppdaga

DERI ved University of Galway, Irland, har teke Tim Berners-Lee tilrådingar eitt steg lenger og lansert noko dei kallar *Linked Open Data Star Scheme* (til kvar av tilrådingane høyrer det til eksempel, sjekk den oppgitte URL-en i fotnoten):



Figur 7: DERI (Univ. of Galway) [11] enkle tilrådingar for å byggja ut lenka data

Men det er ein illusjon å tru at vi vil bruka dei same namna og referera til dei same definisjonane når vi lagar utsagn av typen over. Kva gjer vi når ulike namn blir brukte om den same tingen? Eksempelet under er eit tenkt eksempel som viser korleis ein kan seia at filmen ”The Lord of the Rings” er den same som ”Ringenes herre”:

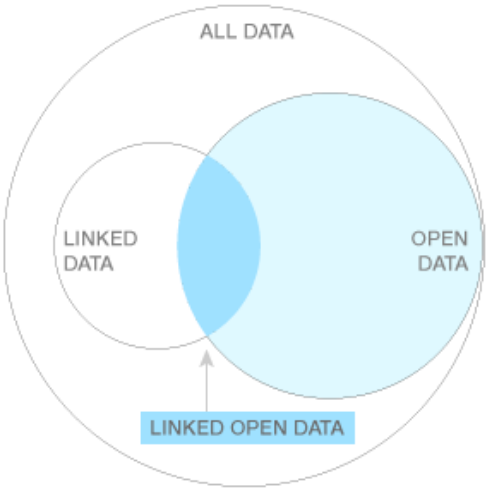
Subjekt:	http://dbpedia.org/resource/The_Lord_of_the_Rings
Predikat:	http://www.w3c.org/2002/07/owl#sameAs
Objekt:	http://dbpedia.org/resource/Ringenes_Herre

Her brukar vi OWL-uttrykket (statement) sameAs for å seia at filmen med den engelske tittel ”The Lord of the Rings” er den same som filmen med den norske tittelen ”Ringenes Herre”.
Lenka data byggjer på den generelle web-arkitekturen og kan såleis seiast å vera eit ekstra lag tett samankopla med den klassiske dokument-sentriske web-en. Dei delar mange av dei same eigenskapane:

- er generisk og kan omfatta alle type data
- alle kan publisera *lenka data*
- dei som publiserer data blir ikkje tvinga til å bruka spesielle vokabular
- objekta (entities) er kopla saman ved hjelp av RDF-triplettar som til saman skaper eit global nett som gjer det muleg å oppdaga nye data kjelder

[10] Tim Berners-Lee (2006): Linked data – Design Issues.
http://www.w3.org/DesignIssues/LinkedData.html
[11] University of Galaxy, DERI: http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/

Skilnaden på *lenka data* (Linked Data) og *opne lenka data* (Linked Open Data) er at lenka data som teknologi kan brukast både på private data og på opne data. Lenka data-teknologien brukt på opne data, blir *opne lenka data* som figuren under viser:



Figur 8: Lenka data og opne lenka data[12]

Fleire kjelder for *lenka data*:
http://linkeddata.org (nettstad med informasjon om Linked Data)

Tim Berners-Lee (2006): Linked data – Design Issues. (artikkel)
http://www.w3.org/DesignIssues/LinkedData.html

Introduction to Linked Open Data for Visualization Creators: (web-artikkel)
http://datavisualization.ch/opinions/introduction-to-linked-data

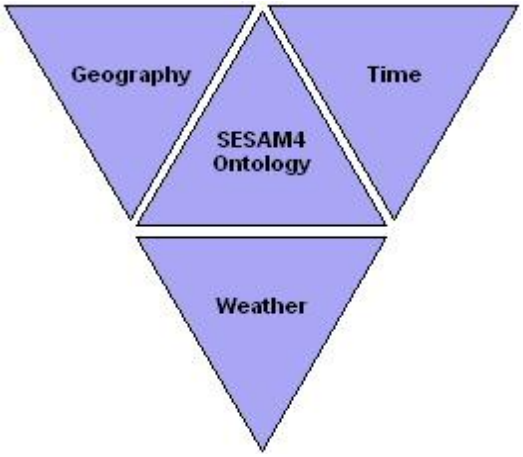
Tom Heath og Christian Bizer: Linked Data – Evolving the Web into a Global Data Space, (bok)
http://linkeddatabook.com/

[12] http://datavisualization.ch/opinions/introduction-to-linked-data

Vedlegg 2: Mapping mellom VisitNorway og tellUs

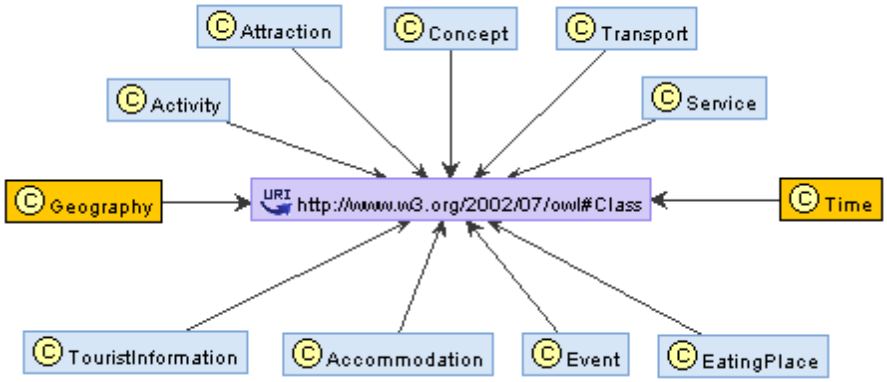
Visitnorway kat 1	Visitnorway underkategori 1	Visitnorway underkategori 2	tellUs ID	tellUs 1	tellUs 2	tellUs 3	K
What to do (Ting å gjøre)							
What to do (Ting å gjøre)	Sports & Activities (Sport og aktiviteter) **tellUs: aktiviteter**	Extreme sports (Ekstremsport)	36_1475 36_775 36_690_2349 36_690_2342* 36_690_2354* 36_690_738 36_690_880 36_690_881 36_690_1279	Aktiviteter Aktiviteter Aktiviteter Aktiviteter Aktiviteter Aktiviteter Aktiviteter Aktiviteter Aktiviteter	Fossepadling Grotting Sport Sport Sport Sport Sport Sport Sport	 Isklatring Skikiting Snøkiting Fallskjermhopping Hanggliding Paragliding Luftsport anlegg	*D a
What to do (Ting å gjøre)	Sports & Activities (Sport og aktiviteter) **tellUs: aktiviteter**	Hunting	**36_137** 36_137_994 36_137_995 36_137_1087 36_137_1088 36_137_1152 36_137_1153 36_137_1154 36_137_1155 36_137_1156 36_137_1157 36_137_1158 36_137_1159 36_137_1191 36_137_1399 36_137_1400	**Aktiviteter** Aktiviteter Aktiviteter Aktiviteter Aktiviteter Aktiviteter Aktiviteter Aktiviteter Aktiviteter Aktiviteter Aktiviteter Aktiviteter Aktiviteter Aktiviteter Aktiviteter Aktiviteter	**Jakt** Jakt Jakt Jakt Jakt Jakt Jakt Jakt Jakt Jakt Jakt Jakt Jakt Jakt Jakt Jakt	 Jakt kurser Jakt med guide Elg Guidet Elgjakt Hjort Guidet Hjortejakt Rådyr Guidet Rådyrjakt Guidet Reinsdyrjakt Rype/Skogsfugl Hare Annen jakt Gåsejakt Storvilt Småvilt	B h F in u

Vedlegg 3: Reiselivs-ontologi utvikla i prosjektet Sesam4



Figur 9: Modulær ontologi

I utviklinga av ontologi i Sesam4-prosjektet har vi lagt stor vekt på å ta i bruk eksisterande ontologiar så langt som muleg og berre laga eigne definisjonar der det ikkje eksisterer passande. Gjenbruk er viktig på dette området. Dersom alle lagar sine eigne modellar/ontologiar, blir det vanskeleg med utveksling av informasjon.Det blir også dyrare fordi ein må starta på nytt kvar gang.



Figur 10: Sesam4-ontologi med sentrale klassar[13]

I den Sesam4-spesifikke delen av ontologien er følgjande klassar definerte

Classes

Accommodation
Attraction
Event
EatingPlace
Transport
TouristInformation
Service

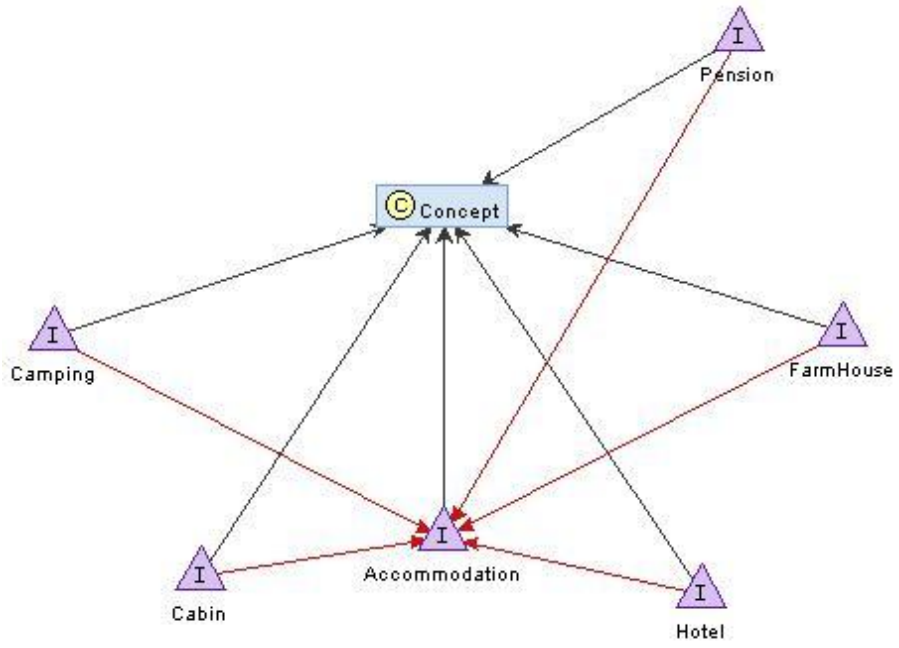
[13] Klassar slik dei er definerte i W3C-standarden OWL (Web Ontology Language)

Activity
Concept

Properties

locatedIn
hasDate
hasLocation
hasOpeningHours
isLocatedIn
hasLocation
hasContact
hasRoom
hasPostalAddress

Klassen *Concept* er sentral i ontologien. Den inneheld alle emnorda (keywords) og er såleis den som kjem nærmast dagens tellUs-kategoriar. Den inneheld dei same hovudkategoriane som tellUs, men har langt færre emneord. Vedlegg 3 syner alle konsepta definerte i Sesam4.



Figur 11: Modell av *Concept* med eksempelet 'Accomodaton'

Klassen *Concept* har desse begrepa:

Accommodation
Hotel
Pension
Cabin
Bed & breakfast
Farm house
Camping

Eating place

- Restaurant
- Café
- Snack bar
- Farm kitchen
- Catering
- Bar and Pub

Activity

- Skiing
- Rafting
- Fishing
- Walking
- Climbing
- Cycling
- Kayaking
- Fitness
- Trekking
- DogSledding
- Sledding
- SnowScooter
- SnowBoating
- Sailing
- Horseriding
- Diving

Excursion

- Guided tour
- Fjord tour
- Railway
- Canyon tour
- Glacier tour

Events

- Cultural Events
- Sport Events
- Concert

Attraction

- Museum
- Visitor centre
- Art gallery
- Architecture
- Memorial
- Monument
- Viewpoint
- Picnic area
- Fjord

- Waterfall
- Mountain pass
- National park
- Zoo and Aquarium

Service

- Bank
- Gas station
- Car Parking
- Car Rentals
- Bicycle Rentals
- Boat Rentals
- Kayak Rentals
- Ski Rentals
- Shopping Centre
- Food Store
- Clothing Store
- Sport Store
- Antiquity Store
- Book Store
- Medical centre
- Dentist
- Hospital

Tourist Information

- Tourist Organization
- Contact Person

Transport

- Airline
- Boat
- Bus
- Express Bus
- Railway

Ontology Construction: Background and Practices

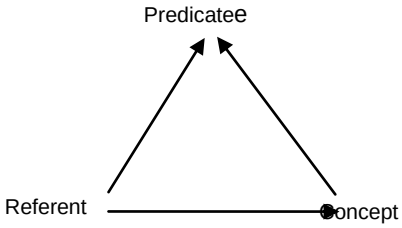
Terje Aaberge

Vestlandsforsking
taa@vestforsk.no

Abstract. The paper discusses shortly what an ontology is, what purpose it has and how to proceed to construct an ontology for the object language of a domain. An ontology does not concern the semantics *per se,* but the construction of an ontology has to be made in an interpreted language and is evaluated accordingly. I thus start to introduce notions from formal semantics. Ontology construction cannot be seen in isolation, but must be considered in relation to its use. I therefore add an account on the storage and retrieval of information. In particular, I discuss this with respect to the paradigm of Internet of Things/Linked Data which presently appears as a promising realisation of the vision of the Semantic Web.

1. Language and semantics

A language has a vocabulary and it is characterised by properties that are referred to as syntax, semantics and pragmatics. The syntax decides what are to be accepted as well formed sentences, i.e. valid juxtapositions of words from the vocabulary. The semantics is a theory on how meaning of words is tied to external objects and activities, and the pragmatics the study on how the situations or contexts of human communication contribute to meaning. It is worthwhile to notice that these properties are not independent. Thus, the meaning of a sentence is determined by the meaning of the words composing it provided it is well formed, i.e. an interpretation of a language is an interpretation of its vocabulary. Sentences that are not well formed are meaningless.
A natural point of departure for a discussion of semantics is the semiotic triangle:



It expresses that our *idea* of object is represented by a word (predicate). This word also represents the set of objects (referent) that is the origin of the idea (concept). The concept is a cognitive entity expressing the meaning of the predicate. The semiotic triangle therefore pictures what meaning is and in part how it is grounded. Descriptive or referring words are thus not isolated entities they are related to something outside language. In particular, when speaking about a language we cannot isolate it form the domain of discourse, i.e. the set of objects the language is to discourse about. At least, when one is discussing formal language which I will do in the following it is presupposed that it is done in relation to a strictly delimited domain.

1.1 Formal semantics

There are two ways of representing meaning formally, referred to as *extensional* and *intensional*. The extensional meaning of a predicate can formally be expressed by the set of objects (referents) that satisfies the predicate. This set is called the *extension* of the predicate. Thus, for example, the meaning of the predicate ”red” that stands for a property of objects is the category of all red objects and the meaning of the predicate “car” that stands for a

class, the class of all cars[14]. The meaning of a predicate that stands for a kind of relation is the class of all ordered pairs that satisfy the relation etc. Extensional (model theoretic) semantic conceives the structure of the domain to be imposed by the structure of language. Thus, an extensional interpretation is represented by a map from the vocabulary to a conceptual model of the domain pictured as the set consisting of the individuals of the domain, subsets of individuals, sets of ordered pairs of individuals etc. The interpretation map being an isomorphism (one-to-one), maps a name to an individual and a predicate to its extension.

Intensional semantics is based on a conceptual model that conceives the domain as consisting of objects with properties and relations, pictured as a directed graph. The objects are represented as nodes and the relations by arrows (edges). An intensional interpretation is then represented by maps from the domain to the vocabulary of the language: an isomorphism that maps the individuals (or relations) to names and maps (called *observables*) that maps individuals (or relations) to predicates. In this case it is thus the structure of the domain that determines the structure of language.

Observables are identified by mutual exclusion of properties. Two properties that cannot simultaneously be possessed by an individual are represented by predicates belonging to the range of the same observable; an individual cannot at the same time be red and green, colour is therefore an observable. It maps an individual to the predicate representing its colour. Other observables are weight, position in space, temperature etc. An observable represents a kind of measurements and is associated with an operational definition exhibiting a standard of measure, laws on which the measuring device is based and rules of application of the measuring device. For example, the measurement of the colour of an individual consists in holding a colour chart representing the standard of measure for the colours against the individual. If the mental pictures that the observer gets of the colour of the individual and the colour marked red on the colour chart coincide, then red is taken to denote the result of the measurement.

1.2 Formal languages

The vocabulary of the (formal) object language for a given domain consists of names representing the individuals of the domain, predicates standing for properties, and relations and logical constants. Moreover, when we write or utter a sentence informally we introduce some help words like “is” or “has”. That an individual possesses a property or that two individuals are related is then expressed by atomic sentences, e.g. “S-2003 is Black” or “Edward is the FatherOf John”.

The content of a predicate can be stated by *extensional* or *intensional definitions*. An extensional definition of a predicate is simply (directly or indirectly) the list of the names of the individuals that constitute its extension. When the names are denoting identifiable individuals of the domain, the extension of the predicate representing its meaning is given. An intensional definition of a predicate (*definiendum*) is the conjunction (with the connective “and”) of atomic sentences (*definientia*) stating which properties that an individual must possess for the predicate to apply. When the meaning of the definientia is given the definition explains the meaning of definiendum. From an intensional definition of a predicate an extensional one can be derived; the extension of the predicate is the class of individuals that satisfies definientia in the intensional definition. The contrary is not possible. Intensional definitions cannot be derived from extensional definitions. The reason is that intensional definitions contain more information than extensional definitions. Finally, we can express restrictions on the

[14] ”Category” will stand for the extension of a predicate that represent one property, something that in principle can be observed/measured or an extensionally defined predicate. “Class” will stand for the extension of an intentionally defined predicate and “set” will cover both meanings. These notions will be clarified later.

possible meaning of predicates by means of *axioms*. An axiom is an implicit definition that relates the *primary terms* of the vocabulary.
An *ontology* for an object language is a set of

- axioms
- intensional definitions
- extensional definitions

The axioms picture structural properties of the domain and limit the possible interpretation of the primary terms. The intensional and extensional definitions are terminological. They define new predicates from the primary terms[15] that serve to facilitate the discourse, e.g. instead of having to repeat the properties that a individual must possess to be of a certain kind an terminological definition will introduce a predicate to denote the kind. Accordingly, the interpretation of the vocabulary and thus the language is determined by the interpretation of the primary vocabulary.

The notion of meaning considered in this paragraph is *formal*. It does not fully capture the notion of meaning attributed to humans, i.e. that meaning is a cognitive entity or concept. A computer cannot handle *real* meaning, but it can possess a formal semantic, and moreover, manipulate sentences mechanically in accordance with the rules of syntax. The Internet of Things (that objects of the world are represented by a URI on the Web) is a way of explicitly realising a formal semantic. An extensional interpretation is then realised by letting the names of individuals map to their URI and by taking the extension of a predicate to be the set of URIs representing the individuals of its (physical) extension. Or we can establish an intensional interpretation by attaching an index card to the URI representing an individual. The index card then lists the sentences (RDF-triples) attributing properties to the individual. The completion of the representation of the domain the relations between the individuals must be added. The kind of relations is in this case represented by URIs that point to the binary predicates implicitly defined by the axioms. The set of URIs of individuals and URIs representing kinds of relations then constitutes a representation of the conceptual (directed graph) model of the domain. For a computer to possess such a formal semantic its representation must be incorporated. This can be done by RDF-triples formulated in OWL (or any ontology language) that express the semantic relations. It is the possession of a formal semantic that justifies the qualification semantic of web or technology.

2. Principles of ontology construction

2.1 Methodology

The definition indicates that there is an ideal method for the construction of an ontology that consists of the following tasks,

1. delimit the domain of discourse
2. identify a primary vocabulary
3. establish the axioms
4. introduce secondary terms by intensional definitions
5. introduce further secondary terms by extensional definitions

Task number one is preliminary but important because if we do not delimit the domain properly we cannot establish a language of description with a well defined vocabulary. It is the nature of the individuals of the domain that determines the predicates needed for their descriptions. The primary vocabulary consists of the names of the individuals and predicates that represent properties and relations needed to describe the individuals and the structural properties of the domain. The axioms describe structural properties of the domain in terms of

[15] Or rather, all terminological definitions can be expressed by means of the primary term, but one will normally define new terms by terms already defined.

the primary vocabulary. The formulation of axioms will in general not use the name, but a variable that is representing an individual of a certain kind as in the following example,

"if x is the SonOf y and y is the BrotherOf z then z is the UncleOf x"

which is an axiom relating the predicates "SonOf", "BrotherOf" and "UncleOf". x, y and z are here understood to represent unidentified persons. If ... then is a logical connective. Notice that this axiom is picturing a structural property of family relations for a domain of persons. At the same time it expresses implicit dependencies of the possible meaning of the predicates.

Task number four consists in formulating intensional definitions to introduce secondary predicates. These are predicates whose extensions are classes. Examples could be "Hotel", "Pension", "Bed&Breakfast", "Fjord" etc. And finally, (task number five) establish further secondary terms by extensional definitions. An example of such a definition in informal language is "an accommodation is a Hotel or a Pension or a Bed&Breakfast" that defines Accommodation. The (non logical) vocabulary is constructed from the primary terms.

The numbering of the tasks does not refer to the ordering of their execution. One better keeps in mind the different tasks to be accomplished and work iteratively. Moreover, it might be difficult or not practical to try to fully attain the ideal suggested by the work description. For different reasons one may be forced to compromise. In any case, however, I think it is important to keep this ideal in mind and use the work description as a guide. A difficulty one encounters is to identify sufficiently many properties of the individuals necessary to establish intensional definitions that distinguish between the wanted kinds of individuals. This difficulty rapidly appears with predicates borrowed from natural language and used to establish the vocabulary of an object language for a heterogeneous domain. The predicate "Fjord" in the context of tourism could be an example of this.

While an object language for the geological domain will have the primary vocabulary necessary to formulate an intensional definition of "Fjord", it would be an overkill to introduce this vocabulary in an object language for the tourism domain. Instead such predicates must be introduced as primary predicates. This said it is in general preferable to use intensional definitions because the meaning of the predicate stays more strictly delimited. Take the example of Hotel. If this is left as a primary term, then the information providers will use it as a descriptive element for any individual they consider to be a hotel. First of all there might be no common understanding of the predicate and secondly what individual calls itself Hotel is partly accidental. Of two similar individuals one may be registered as hotel and the other not. The registration of individuals as hotels will therefore be inaccurate and what turns out to be the extension of Hotel will be a very badly conceived category. If the creators are filling out a form that lists properties/attributes of the individuals of the domain and thus establishes a description of the individual that is measured against intensional definitions one gains in coherence.

These remarks indicate that there are a number of choices to be made throughout the construction of an ontology. Some of the choices are more or less imposed by the nature of the individuals, with respect to others it is a matter of convenience or of what one wants the ontology to do. This concern in particular the amount of reasoning that is desirable. There are several levels of reasoning. The foremost one is deductions that are based on the axioms. Using the axioms as premises and applying the rules of deduction one might be able to make interesting conclusions called theorems. Depending on the axioms it might, however, not be possible to do this mechanically. The axiom system is then said to be non decidable. In semantic technology one chooses to use a weak language based on a description logic instead of full predicate logic to avoid formulating non decidable axiom system. However, weakening the axioms means that they do not fully picture the structural properties of the domain. In any case, there is a trade off between expressibility and decidability to be decided.

Another and weaker kind of reasoning is base on satisfaction with respect to the intensional definitions. Thus, if the ontology contains an intensional definition of say Hotel, a search on hotel will retrieve all the individuals whose description satisfy the definition. The third kind of reasoning is syllogistic. This can be illustrated by the

following example. Assume that the description of a hotel does not include the information that it have a restaurant, but that there is an axiom that does, then the following syllogism concludes that a particular individual has restaurant,

Sogndal Hotel is a Hotel (follows from intensional definition of Hotel)
All hotels have a Restaurant (axiom)
Sogndal Hotel has a Restaurant (conclusion)

A search on restaurant will therefore retrieve Sogndal Hotel. Accordingly, considerations on what the ontology should be able to do are important for the formulation of the ontology.

2.2 Linguistic analysis

Language is a medium for storage and communication of information and knowledge. To fulfil this function it must have a certain degree of syntactic precision and common semantic for the users; a language is not private. Natural languages satisfy these conditions for large groups of users. Though the predicates might have personal connotations and thus meaning for the users which inhibit precision, human communication works quite well. However, natural language lacks in syntactic precision to include computers among the users. A compromise is then to construct formal languages with a precise syntax whose semantic, in order not to exclude humans, as far as possible is taken over from natural language. In the following I shall give some examples of linguistic analysis that is useful when establishing vocabulary based on natural language semantic and in particular indicate some pitfalls to avoid.
Set theory have two relations, ”element in” and “contained in”[16], that both are mirrored in the object language. The sentence “S-2003” is a Car” mirrors the “element in” by the fact that it means that the individual S-2003 is a specific element in the extension of the predicate “Car”. Furthermore, the sentence “all cars are Vehicles”[17] should be interpreted as saying that the extension of the predicate “Car” is *contained in* the extension of the predicate “Vehicle”. “Vehicle” has therefore a *broader* meaning than the predicate “Car”. As shown by the sentences “car is a Vehicle” and “vehicle is a Car” there is a direction in “element in”. It is only the first sentences that correspond to Car having a broader meaning than Vehicle and which thus corresponds to common usage. The partitative relation represented by the predicate “PartOf” is often confused with the set theoretic “contained in”. We can say that “Sogn is PartOf WesternNorway” but this does not mean that WesternNorway is broader than Sogn. It simply means that the geographical area of Sogn is covered by the geographical area of WesternNorway which is quite different. Notice also that both Sogn and WesternNorway are names.
The meaning of a predicate is narrowed by adding a qualification. Examples are StaveChurch and FjordHotel. In the first case the qualification is Stave that refers to the construction and in the second Fjord that refers to the location. In both cases the narrowing can be understood semantically as an extra condition in the intensional definition. In the first case the condition, informally formulated, is “constructed with staves as supporting elements” and in the second, “localised on the shore of a fjord”. Both of these conditions reduce the extensions of the predicates “Church” and “Hotel”. A more indirect method to narrow the meaning of a predicate is by extensional definitions. One may thus declare that “foss is an Attraction”. The extension of “Foss” then only consists of the fosses that may be considered to be attractions.
The meaning of a predicate is also determined by qualification and context in another way. For example, the extension of the predicate “Fjord” is easily guessed to be contained in that of “Attraction”, if the alternatives are “Activity”, “Accommodation” or “Event”. On the other hand, the FjordCruise falls under Activity because Cruise is to be considered an activity and Fjord a qualification.

[16] ”element in” and “contained in” are not predicates in the object language.
[17] I use the convention of applying upper case first letter in predicates and lower case to when the same word stands for a variable over the extension of the predicate. “car” stands for an individual but unspecified object of the type Car.

Relational sentences can be used to determine whether two predicates have a broader/narrower meaning. "excursion StartsIn Flåm" is an acceptable sentence, but "accommodation StartsIn Flåm" is not. The extension of Excursion is thus not contained in the extension of Accomodation and vice versa.
Non-hierarchical associative relations may be applied to test whether a binary predicate applies to a pair of individuals. Examples are

Cause/Effect	a CauseOf b
Producer/Product	a ProducedBy b
Activity/Actor	a PerformedBy b
Activity/Place	a DoneAt b
Purpose/Activity	a ObtainedThrough b
Tool/Function	a UsedTo b
Material/Product	a MadeOf b
Supplier/Service	a OfferedBy b

The test is to verify that the first term in the relational sentence is in the extension of the first unary predicate in the associative relation and the second term belongs to the extension of the second unary predicate. The unary predicates might not belong to the vocabulary of the language but introduced as means for semantic analysis. Notice that for such sentences to be well formed, both a and b must be names of individuals of the domain.
As the discussion indicates, unary predicates are as a rule expressed by nouns in singular. Binary predicates are often expressed by the juxtapositions of a verb and a preposition.
The few and relatively simple examples in this paragraph have been given to show that linguistic analysis is referred to the semantic and how the choices made can be justified by semantic theory.

3. An Example of ontology

As already mentioned in 2.1, the ontology construction is preceded by a delimitation of the domain. Next one may start to identify the natural language vocabulary that is used to describe the elements or classes/categories of the domain. Putting the name of the domain on top one may then try to arrange a subset of the vocabulary hierarchically as a taxonomy. It is then useful to think extensionally with respect to the individuals of the domain. No individual should *naturally* belong to extensions of predicates in different branches of the taxonomy. However, if companies are considered to be the individuals of the domain, a company providing different kinds of services will do so. On the other hand, if it is the services offered that are the individuals of the domain, this might not be the case.

The following example of such a taxonomy was constructed for the Sognefjord tourism domain[18]. Destination Sognefjord is thus the top of the hierarchy:

DestinationSognefjord
- Attraction
  - Artefact
    - VisitorCentre
    - ArtGallery
    - Architecture
      - *StaveChurch*
      - *StoneChurch*
    - Memorial

[18] Sognefjord cultural heritage domain would have a different ontology, but part of it will be contained in the Sognefjord tourism domain. The two ontologies can thus be aligned.

- Monument
- Museum
- Viewpoint
  - *PicnicArea*
- NaturalSite
  - Fjord
  - Waterfall
  - NationalTouristRoad
  - NationalPark
- Activity
  - SkiCentre
  - Rafting
  - Fishing
  - Walking
  - Climbing
  - Cycling
  - Kayaking
- Excursion
  - GuidedTour
    - *CanyonTour*
    - *GlacierTour*
  - FjordTour
  - Railway
- Event
  - Culture
  - Sport
- Accommodation
  - Hotel
  - Pension
  - Cabin
  - Bed&Breakfast
  - FarmHouse
  - Camping
- EatingPlace
  - Restaurant
  - Café
  - SnackBar
  - FarmKitchen
- Service
  - Bank
  - GasStation
  - Garage
  - Rentals
    - *CarRentals*
    - *BicycleRentals*
    - *BoatRentals*
    - *KayakRentals*
    - *SkiRentals*
  - Shopping

ShoppingCentre
FoodStore
ClothingStore
SportStore
AntiquityStore
BookStore
MedicalCentre
Dentist
Hospital
Community

The taxonomy lists predicates for individuals in the Sognefjord area that a priori are judged interesting for tourists. As will be seen there is an extensive use of qualifications which at this stage should be interpreted extensionally. Chosen as an ontology, the leaf predicates are the primary terms of the description language. The other predicates are extensionally defined. At this stage one should thus reflect on how deep to make the ontology. The predicates stand for categories/classes of individuals. This categorisation/classification is inherited by the set of information documents on the individuals. The documents belonging to a given category/class is annotated by the appropriate leaf predicate. The granularity of the leaf level thus decides the preciseness of search results on the document set using the predicates as query terms. This can however, be compensated by other means to be discussed.

The next step to consider is the properties/attributes[19] of the individuals of the domain identify the predicates and define the observables. All the individuals in the tourism domain can be assigned a geolocation. Geolocation is thus an observable. Other observables are PostalAddress, TelephoneNumber, EMailAddress etc. Clearly, individuals of the domain have value 0 for some of the observables. For example, a natural site will normally have value 0 for all of the above observables. The object language description of an individual is a designation of a value for each of the observables. A more complete list of observables established mostly with respect to accommodations is

GivenName
Owner
Address
PostalNumber
Telephon
Telefax
MobilPhone
eMail
WebSite
GeoLocation
AlcoholLicence
ConferenceFacility
Breakfast
Lunch
Dinner

[19] It is natural to distinguish between properties and attributes of individuals. Thus a property is something that an individual possesses and that can be measured (e.g. geolocation) while an attribute is associated to the individual by some human decision (e.g. address). In the following I will extend the meaning of observable also to cover kinds of attributes.

Room
Reception
TrainingFacility
SwimmingPool
TVinRoom
Parlour

The list must be sufficiently furnished to allow distinctive descriptions of all items belonging to the domain. This list might not be satisfying that condition even for all kinds of accommodations listed in the taxonomy.
Once the observables are defined one may proceed to establish intensional definitions for the leaf predicates. It is a tedious and difficult work to define observables and even more so to establish intensional definitions. These must as well as possible distinguish between individuals belonging to the extensions of the different leaf predicates. The gain is however considerable, even more so for big domains. For example, descriptions of individuals might contain qualifications that are not part of the intensional definition determining the extension of the leaf predicate to which they belong. Information on only the restricted class of individuals can still be retrieved by adding the qualification as a conjugate search term. Thus, if the value Fjord of the observable Location is part of the description of a hotel, then the search on "Fjord and Hotel" will retrieve information on all fjordhotels but not other hotels. The search would then determine the individuals that satisfy the intensional definition of Hotel and restrict to those with the qualification Fjord. The same qualification could be used for individuals that are not hotels without retrieving information on these. Examples of intensional definitions of the different kinds of accommodations are:

Hotel:
(hotel has GeoLocation) and (hotel has AcoholLicence) and (hotel has ConferenceFacility) and (hotel serves Breakfast) and (hotel serves Lunch) and
(hotel serves Dinner) and (hotel offer Room) and (hotel has Reception) and
((hotel has TrainingFacility) or (hotel has SwimmingPool)) and (hotel has TVinRoom) and (hotel has Parlour)

Pension:
(pension has GeoLocation)and (pension serves Breakfast) and (pension serves Dinner) and (pension offer Room) and (pension has Reception) and (pension has Parlour) and (pension hasnot AcoholLicence)

Bed&Breakfast:
(bed&reakfast has GeoLocation) and (bed&breakfast serves Breakfast) and
(bed&breakfast offer Room) and (bed&breakfast hasnot AcoholLicence) and
(bed&breakfast hasnot Reception)

The words "has", "serves", and "offer" are here help words, not binary predicates. These examples, formulated in the object language are given to illustrate the principle; that they should be distinctive. There is no claim that these definitions represent the final answer. Any choice must be tested on a sample of descriptions of real objects and the description scheme and intensional definitions adjusted according to the results of the tests.

It is even more difficult to discover structural properties of the domain that may be put down as axioms. However, some simple axioms can be decided on. For example, if it is true that all hotels have a restaurant and that this is not a distinguishing feature, then "all hotels have a Restaurant" can be stated as an axiom. The axiom depends on the definition of Restaurant. Another example would be the case that all providers of kayaking (tours) also offer kayak renting. The axiom would be "Kayaking is the same as KayakRental". Such a proposition might be true in one tourism destination, but false in another.

4. Organisation of the work

As already mentioned, the development of a ontology is arduous work, especially if the domain is heterogeneous. It also depends on what purpose the ontology should serve, if its only uses is to organise the information resources about the individuals of the domain or if it is the core element in the language of discourse and therefore how deep and complete it should be. The ontology construction is a process involving two kinds of participants, domain experts and experts on formal languages and semantic. The last ones should also possess a certain amount of domain knowledge before starting the work. The domain experts will know the informal domain language and assure that the incorporated in the formal language. The language experts will on their side be able to analyse terms in relation to syntactic role formal meaning, formulate definitions and give informed advices.

It is advisable to start with a proposal possible made by the language experts. The proposal that does not have to be complete is then put to the scrutiny of the domain experts following an explanation of the structure by the domain experts. The discussion might then lead to an extension of the ontology and revision of definitions. Experience shows this to be an efficient work method. It works particularly well if the domain experts also possess domain knowledge.

The work process can be as follows:

1. determine
    - domain (subject area)
    - use
    - scope
    - user group
2. choose a vocabulary
3. order the terms according to type
4. structure the predicates with respect to meaning
5. formulate axioms and terminological definitions

It is important to clarify the four point under 1 because they to a big degree determines the further process. Without a precise delimitation of the domain it is not possible to establish a semantically well founded language. What use the ontology is intended to serve should also be decided at the beginning as it will determine the extent of work needed. Is the ontology only to be used to thematically order the information resources or is it to be used in a language of discourse for the domain. An ontology that is made only to order information resources thematically is much simpler than an ontology for a description language that also must contain axioms picturing structural properties of the domain. It only contains predicates that stand for kinds of objects. A thematic ontology kan be used as the core of a description langage, but must then be supplied by predicates that stand for properties. How one formulates terminological definitions depends on the user group. If the users are persons whose knowledge about the domain is lacking but are to catalogue the information resources it is important to express definitions/explanations is expressed without to heavy use of specialised terms but without simplifying to the extent that they are wrong or imprecise.

An ontology is not made for eternity. New thing appear that one must take into account and the language changes with the evolution of the society. An ontology thus has to be reviewed and revised. This can be done within the framework already discussed. One might also need to align ontologies of unified domains. His is more challenging, especially if the domains are composed of similar objects.

5. Information Storage and the IoT/LD paradigm

The language of discourse to be employed when including computers among the users has two levels, a linguistic level and a metalinguistic level. It includes an object language that can formulate descriptions of the individual objects of the domain and an object metalanguage that can formulate descriptions of the descriptions of the individual objects. For example, a database storing information about companies will register each item by an Id and otherwise describe a company by giving its name, address, telephone number etc. The information is organised in a table whose abscise is the Ids and the ordinate the field names Name, Address, TelephoneNumber etc. There are thus three levels of abstraction involved, the domain level represented by the Ids serving as unique names of the companies, the object predicate level represented by the set of given names, addresses, telephone numbers and the meta level represented by the field names that are predicates in the metalanguage. An example of an object language sentence attaching the name to an Id xy is "xy is Sogndal Hotel". Since Sogndal Hotel is the only with this name in our domain the extension of Sogndal Hotel is the singleton set informally denoted {xy}, i.e. the set that only contains Sogndal Hotel. The extension of the predicate "Name" on the other hand is the set of all names and thus a set of sets of individuals.

In the metalanguage we can make statements like

Sogndal Hotel NameOf xy
Sogndal AddressOf Sogndal Hotel

because both xy and Sogndal Hotel can be employed as names in the metalanguage, at least when one refer to an intensional semantic. NameOf and AddressOf are binary predicates in the metalanguage. This shows how field names in a database are related to predicates in the metalanguage.

Ontology construction in computer science and Semantic Web context includes considerations on the meta level. It is in this respect important to be consciously aware of the distinctions between the object and meta levels. In particular, metalanguage sentences must not be mixed with object language sentences. This might lead to incoherence and in the worst case inconsistencies.
A database representation of information does not possess a formal semantic. A way of introducing one is suggested by the idea of Internet of Things (IoT). In the context of the Semantic Web the IoT stands for the idea of referring to "things of the world" on the Web by Universal Resource Identifiers (URI). A URI comes as a URL (locator, dereferenceable URI) or URN (name). Technically they both function as Ids, though one denote the Web-location or Web-identifier of a resource and the other the name.

There are two kinds of resources; those that can have a digital representation on the Web and those that can only be referred to. For example, a digital representation of a book may be stored on the Web and located by a URL and it may have a unique name (URN), its isbn-number. On the other hand, "things of the world", i.e. physical objects, persons, towns, services etc. can only be referred to by URIs.

The distinction between objects and information documents in the IoT suggests a modelling methodology. Given an object domain, the set of individual objects and information documents has a conceptual model pictured as a directed graph

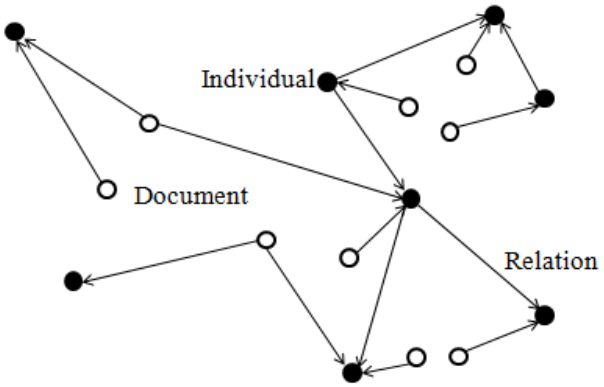


Figure: conceptual model of domain

where nodes represents objects and documents. The relations between documents and individuals are of one kind, About. The other relations are internal to the object domain. A formal description of the domain needs two, *a priori* distinct, languages; one to describe the objects and one to describe the documents. Each of them is endowed with an ontology. The ontology of the document language interferes only with the information architecture, i.e. how one can possibly present the information contained in the documents. The description language for the total domain is the union of the two adding the predicate "About". URLs are then representing objects and documents and kinds of relations.

The conceptual graph model can be realised as a Linked Data (LD) set whose tenets are to

1. use the RDF data model to publish structured data on the web
2. use RDF links to interlink data from different sources

The Linked Data paradigm has evolved as a powerful enabler for the transition of the current document-oriented Web into a Web of interlinked Data [1,2]. The term Linked Data here refers to a set of best practices for publishing and connecting structured data on the Web. These best practices have been adopted by an increasing number of data providers over the past few years, leading to the creation of a global data space that contains many billions of assertions – the Web of Linked Open Data (LOD).
The methodology then consists in attaching a machine readable "index card" to each thing-URI. An example is the URL http://www.w3.org/People/Berners-Lee/card locating the index card that is attached to the URI http://www.w3.org/People/Berners-Lee/card#i that represent the person Berners-Lee [1]). Attached means here if one enters the URI for Berners-Lee into the browser one is automatically redirected to the index card. By means of this method an agent will however, "know" that this is a thing in the world and by reading the index card it will also "know that it is a question of a person etc. In other words, an index card lists the properties of its Thing as RDF-triples. The Thing is thus represented as a bundle of properties (attributes). The description framework can be considered to be constituted by four components,

- a domain model (linguistically represented as linked data)
- ontology (for the object domain)
- an index (constituted by the index cards for the Things providing the formal semantic)
- a knowledge base (the collection of documents)

of which the domain model is contained in the index. The directed graph being a model of the domain serves as a basis for the formal semantic, i.e. we are by this method able to represent not only the description of objects but also semantic relations.

Whether this framework is represented in a database or as an owl/xml-file the retrieval mechanism is based on search through the ontology. This means that the search term is compared with the ontology. If it is a primary (non-defined) term it will pass through and then be compared with the content of the index cards. This

determines a set of URLs representing individuals that satisfy the search criteria. Next one uses the domain model to pick the documents (or sub parts of documents) that are about these individuals, i.e. have the relation About to these individuals. If it is a secondary predicate defined by an intensional definition, i.e. the conjunction of atomic sentences stating which properties that an individual must possess for the predicate to apply, then the search engine is comparing the definition of the predicate with the descriptions of the individuals on the index cards. It picks the individuals that satisfy the definitions. The next step is identical to the former case. Other mechanisms will also be studied. The semantically assisted queries in this framework are thus fundamentally different from the keyword or statistically weighted text-based searches in a document base.

The paradigm of IoT/LD opens up for many possible applications, technically quite feasible at present. In fact, to build an IoT/LD based application one may follow the given methodology. This means to constitute a domain by choosing URLs representing individuals and documents from several published repositories. One then establishes a domain model and constructs an ontology. The domain model and the ontology will use elements from the sources chosen, but may be adapted to the tasks to be performed by the application. The ontology and domain model will be implemented in the application, i.e. encoded therein. It will retrieve information from the published knowledge bases by means of the mechanisms already described. How smart the applications can be made will depend on the ontology and the amount of semantic information available in the index.

[1] Christian Bizer, Tom Heath, Tim Berners-Lee: Linked Data - The Story So Far, http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf

[2] Bizer, C., Cyganiak, R. and Heath, T.:How to Publish Linked Data on the Web, http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/