WESTERN NORWAY RESEARCH INSTITUTE
VESTLANDSFORSKING
www.vestforsk.no

Vestlandsforsking-rapport nr. 3/2012

Semantisk samhandling i kulturformidlinga

Bruk av semantisk teknologi for å binda saman ressursar på tvers av organisasjonar

Svein Ølnes og Rajendra Akerkar

Vestlandsforsking, Pb 163, 6851 Sogndal • Tlf.: 906 33 600 • Faks: 947 63 727

Vestlandsforsking-rapport

Tittel Semantisk samhandling i kulturformidlinga - Bruk av semantisk teknologi for å binda saman ressursar på tvers av organisasjonar	**Rapportnummer** 3/2012 **Dato** Juni 2012 **Gradering** Open
Prosjekttittel Semantisk fotoprosjekt	**Tal sider** : 27 + 20 **Prosjektnr.** : 6224
Forskarar Svein Ølnes (intern prosjektleiar) Rajendra Akerkar	**Prosjektansvarleg** Ivar Petter Grøtte
Oppdragsgivar Fylkesarkivet ved Sogn og Fjordane fylkeskommune	**Emneord** opne data lenka data kultur foto semantisk teknologi samhandling interoperabilitet

Samandrag

ISBN: 978-83-428-0316-0	Pris:

Forord

Fylkesarkivet i Sogn og Fjordane søkte sommaren 2011 Kulturrådet (den gongen ABM Utvikling) om støtte til eit prosjekt for å prøva ut semantiske teknologiar i kultursektoren. Søknaden vart innvilga og Fylkesarkivet har saman med partnarane Preus Fotomuseum, Oslo Byarkiv, ESIS og Vestlands-forsking gjennomført prosjektet. Vestlandsforsking har hatt hovudansvaret for dokumentasjonen av prosjektet.

Rapporten har tre hovuddelar; problemstilling, korleis problemet er forsøkt løyst og kva vi har lært undervegs. I tillegg er det utarbeida ein ganske omfattande omtale av aktuelle vokabular. Denne er teken med som vedlegg og er skriven på engelsk.

Robert Engels, ESIS, har skrive kapitla 2.1 – 2.4 og 2.6 om semantisk løfting, vokabular-mapping m.m. Dokumentasjon av web-demonstrator er gjort av Eirik Stanghelle Morland (tidlegare Fylkesarkivet, no Ny Media AS). Denne delen av dokumentasjonen ligg på nettsidene for demonstratoren.

Takk til Fylkesarkivet og prosjektleiar Øystein Åsnes for eit godt samarbeid i eit spennande prosjekt.

Vestlandsforsking, juni 2012

Innhald

Samandrag

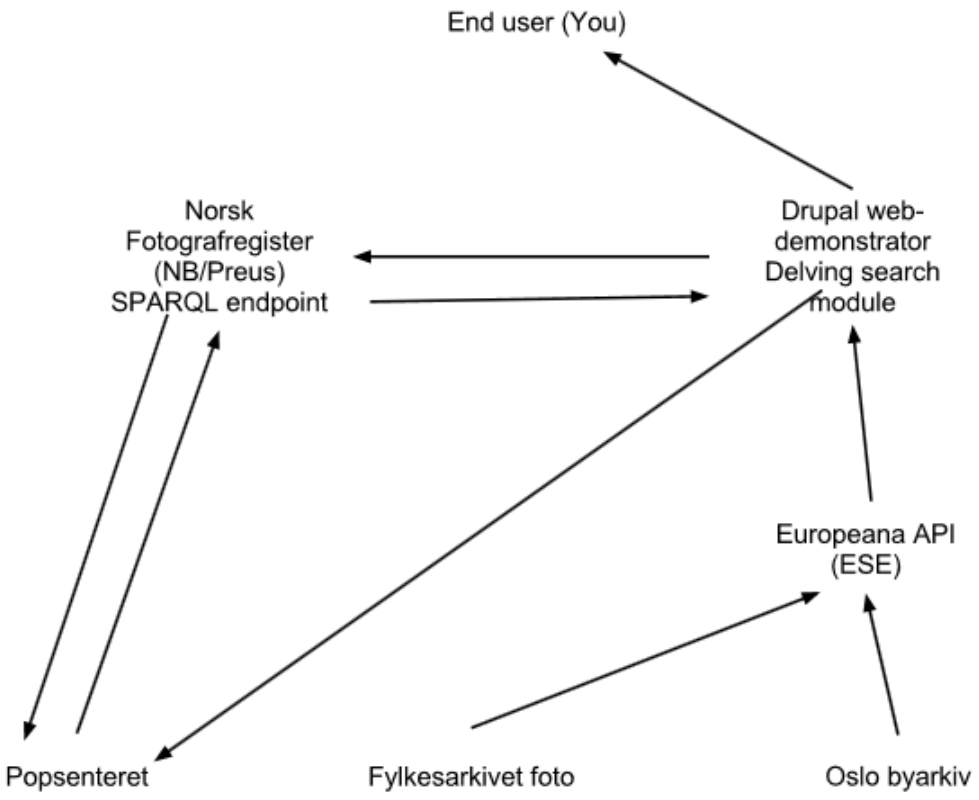
Prosjektet "**Semantisk samhandling i kulturformidlinga – Etablering av eit URI-register for fotografar**" skal:

1. Etablera eit URI-register for fotografar, basert på fotograf-basen frå Preus Fotomuseum

2. Gjennom to demonstratorar (ein på Pop-sentereet i Oslo og ein web-basert) visa korleis publisering av lenka data frå dei aktuelle datakjeldene (Preus' fotograf-base, Oslo Byarkiv sitt bilet-arkiv og Fylkesarkivet i Sogn og Fjordane sin foto-base) kan kopla samhøyrande element

3. Dokumentera problem og løysingsforslag som tilrådingar for vidare arbeid i kultursektoren

Prosjektet har teke utgangspunkt i fotografbasen ved Preus fotomuseum og for å visa korleis ei semantisk løfting og opning av data kan gjera det lettare å utnytta ressursar på tvers i sektoren. Den semantiske versjonen av fotografbasen skal så lenkast til fotobasar ved Oslo Byarkiv og Fylkesarkivet i Sogn og Fjordane, og eit URI autoritetsregister for fotografar skal etablerast med utgangspunkt i fotografdatabasen.



Figur 1: Oversikt over ulike delar av prosjektet

Tilrådingar for vidare arbeid

Trass i utfordringane omtalte i rapporten, er erfaringane våre gode frå arbeidet med semantisk samhandling basert på teknologien *lenka data*. Det er eit stort behov for å opna opp dei mange lukka databasane i sektoren, til liks med andre sektorar, og lenka data er ein lovande teknologi i så måte.

Arbeidet med tilpassing av informasjon til Europeana, via det norske knutepunktet Norvegiana, er eit viktig steg på vegen mot betre samhandling. Men metadata-informasjonen i Norvegiana/Europeana må styrkast og få ein tydelegare semantisk retning, dvs. baserast på tilrådde semantiske standardar. Dagens ESE-standard brukar t.d. berre streng-verdiar og ikkje URI-ar. Det føregår likevel mykje interessant arbeid for å ta Europeana-data vidare semantisk. Det er ein tung prosess, og førebels skjer mykje av utprøvinga "top-down" der ein tek utgangspunkt i ESE-basert Europeana-informasjon og konverterer til EDM. Eit relevant arbeid sett i høve vårt prosjekt, er Europeana LOD[1] der Europeana-informasjon blir konvertert til EDM og så vidare til LOD.

Ut frå erfaringane i prosjektet "Semantisk samhandling i kulturformidlinga" tilrår vi ei slik vidareføring:

1. Utarbeida generelle retningslinjer for etablering av **autoritative URI-register på ulike nivå**; nasjonalt og regionalt.

2. Føreta ein **gjennomgang av vokabulara i sektoren** der målet må vera å komma fram til vokabular som støttar samhandling. Vokabulararbeidet i kultursektoren (som i andre sektorar) har i altfor liten grad brydd seg med samhandling, men vore veldig oppteke av presisjon i katalogiseringsarbeidet. Dette er ein stor og tung prosess som dreier seg om å snu perspektivet frå ei avsendar- til ei mottakarorientering. Det handlar om å bli mindre intern og meir ekstern og i langt større grad tenkja på korleis informasjonen kan bli brukt av andre, jfr også prosjekttittelen "Semantisk samhandling i kulturformidlinga" (kultur*formidlinga*, ikkje kultur*sektoren*!)

3. Sy saman strategiane for URI-register og vokabular med vidare planar for Norvegiana/Europeana. Her må ein finna ut kor stort handlingsrom ein har nasjonalt utan at endringar i informasjongrunnlaget svekkjer leveransane til Europeana. Her er det også naturleg å sjå på den vidare utviklinga i Europeana og overgang frå dagens **Europeana Semantic Elements (ESE)** til bruk av den nye semantiske standarden **Europeana Data Model (EDM).** Det føregår det mykje relevant arbeid på Europeana-nivå, som omtalt over, og særleg interessant er diskusjonen om EDM introduserer for mykje kompleksitet på ein første veg mot lenka data (Haslhofer & Isaac, 2011). Problema er heller ikkje utelukkande på Europeana-nivå, men også på framleis manglande standardisering på W3C-nivå med eit

[1] http://pro.europeana.eu/linked-open-data

manglande overordna perspektiv[2] og grep på vokabular-utvikling og –bruk.

Organisatorisk samhandling må takast på alvor. Dette er kanskje den viktigaste og vanskelegaste delen av arbeidet. I første omgang er det viktig at sentrale institusjonar blir samla og at ein får ei felles forståing for utfordringane og hovudretninga vidare. Så må den enkelte institusjonen gjera ein innsats for at ein til saman skal kunna oppnå målet om ei betre samhandling for ei betre kulturformidling.

[2] EU har nyleg gjennomført ei høyring på forslaget om Core Vocabularies Specification – eit forslag om tre overordna vokabular for personar (Person), næringsliv (Business) og geografisk informasjon (Location). Sjå nærmare omtale på https://joinup.ec.europa.eu/sites/default/files/Core_Vocabularies-Business_Location_Person-Specification-v0.2_1.pdf

1. Kva er problemet?

1.1 Prosjektet

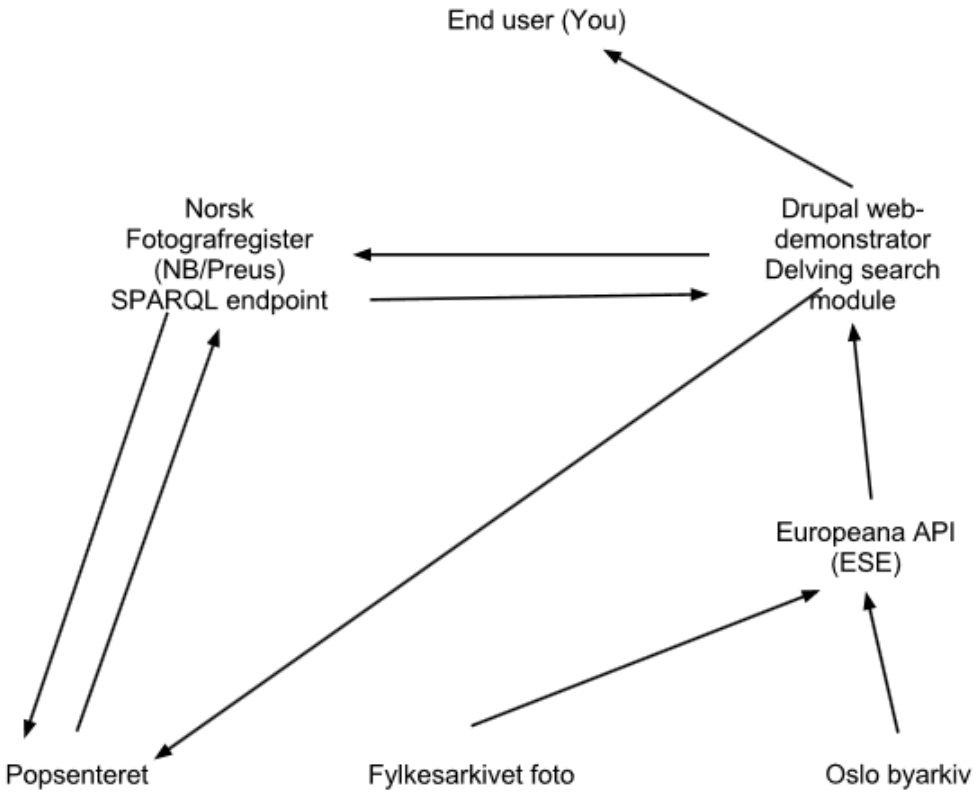
Prosjektet "**Semantisk samhandling i kulturformidlinga – Etablering av eit URI-register for fotografar**" skal:

1. Etablera eit URI-register for fotografar, basert på fotograf-basen frå Preus Fotomuseum

2. Gjennom to demonstratorar (ein på Pop-sentereet i Oslo og ein web-basert) visa korleis publisering av lenka data frå dei aktuelle datakjeldene (Preus' fotograf-base, Oslo Byarkiv sitt bilet-arkiv og Fylkesarkivet i Sogn og Fjordane sin foto-base) kan kopla samhøyrande element

3. Dokumentera problem og løysingsforslag som tilrådingar for vidare arbeid i kultursektoren

Prosjektet har teke utgangspunkt i fotografbasen ved Preus fotomuseum og for å visa korleis ei semantisk løfting og opning av data kan gjera det lettare å utnytta ressursar på tvers i sektoren. Den semantiske versjonen av fotografbasen skal så lenkast til fotobasar ved Oslo Byarkiv og Fylkesarkivet i Sogn og Fjordane, og eit URI autoritetsregister for fotografar skal etablerast med utgangspunkt i fotografdatabasen.



Figur 2: Oversikt over ulike delar av prosjektet

Figuren over viser samanhengen mellom dei ulike delane av prosjektet:

- Norsk Fotografregister: Database med oversikt over fotografar og fotosamlinga. Preus Fotomuseum har vedlikehaldsansvaret og Nasjonalbiblioteket har det tekniske ansvaret. Data frå basane blir utstyrte med ekstra metadata og eksporterte til rdf-format. Det blir oppretta eit SPARQL endepunkt (SPARQL Endpoint) slik at det kan gjerast spørjingar mot data.

- Fylkesarkivet sin fotobase: Data om utvalde fotografar frå basen er tilførte ekstra metadata og så eksporterte i rdf-format, saman med bileta frå desse fotografane.

- Web-demonstratoren er utvikla på Drupal-plattforma og les data frå xml-repositoriet (utgangspunktet for Europeana-straumen). Data blir så sendt til SPARQL-endepunktet for fotografbasen.

- På grunn av uføresette hendingar (jobbskifte) har det ikkje vorte levert data frå Oslo Byarkiv.

1.2 Demonstratorar og dokumentasjon

I tillegg til å visa dataintegrasjon ved bruk av semantiske teknologiar, skal prosjektet også visa korleis ein kan oppretta autoritative kjelder, også kalla autoritative URI-register.
Løysingane skal visast i to demonstratorar:

- ein demonstrator som viser integrasjonen av fotografar og fotobasar ved å projisera foto frå ulike kjelder på ein vegg i Popsenteret (Oslo kommune)

- ein web-basert demonstrator som viser integrasjonen av dei same fotokjeldene.

Denne rapporten er i tillegg eit viktig resultat av prosjektet. Demonstratorar er nyttige verktøy for å visa kva ein kan oppnå og ein fin måte å visa i praksis korleis noko fungerer. Demonstratorar er ikkje like gode til å visa hindera på vegen fram mot ei ferdig løysing eller ein ferdig demonstrator. Tvert om vil ein demonstrator ofte skjula problema ein har støytt på på vegen, i staden for å kasta lys over dei.

I eit prosjekt som dette, er hindera på vegen viktig å identifisera og dokumentera. Berre på den måten kan ein læra av prosjektet og utarbeida råd og retningslinjer for å hindra at andre kjem opp i dei same problema. Og om dei gjer det, må dokumentasjonen visa korleis dei ulike hindera kan forserast. Alt i alt blir dette ein dokumentasjon av beste praksis. I tillegg vil det vera nokre sentrale problemstillingar som prosjektet ikkje kan gi svar på, men løfta opp til dei som har ansvaret for å finna løysingane.

1.3 Problem med samhandling og kopling av data

Utgangspunktet for søknaden om prosjektmidlar til ” ”Semantisk samhandling i kultursektoren – Etablering av eit URI-register for fotografar” var problemet med mangel på elektronisk samhandling i sektoren. Det er eit problem kultursektoren deler med dei fleste andre sektorar, og det er ei stor utfording både i offentleg og privat sektor. Det finst mange isolerte datakjelder som vanskeleg let seg

kombinera, sjølv om dei har stort slektskap mellom seg. Vi har mange øyar av data, med svært få bruer eller annan kommunikasjon mellom.

Kultursektoren kan ha stor nytte av semantiske teknologiar sidan informasjonen her ofte er knytt til annan informasjon, også utanfor sektoren. Nytten kan vera stor både ved å binda saman informasjon innan sektoren, og ved å binda saman informasjon frå kultursektoren med informasjon frå andre sektorar, som t.d. reiselivet. I dette prosjektet er det bruk av semantisk teknologi for samanbinding av informasjon innan sektoren som er hovudpoenget.

Vi valde å ta utgangspunkt i fotografar og fotografi, og målet med prosjektet har vore å etablera eit URI-register for fotografar og å visa korleis ein ved semantiske teknologien *lenka data* (Linked Data) kan knyta saman ulike datakjelder på ein open og inkluderande måte. Open fordi det blir tilgjengeleg for alle, og inkluderande fordi teknologien nyttar seg av same grunnfilosofi som veven sjølv: hyperlenking . Skilnaden på lenka data og tradisjonell vev-teknologi er at lenka data lagar hyperlenker mellom data, som namnet seier, medan den tradisjonelle veven lagar hyperlenker mellom dokument.

Med prosjektet vil vi visa at teknologien lenka data gir gode mulegheiter for betre samankopling av datakjelder i kultursektoren, i tillegg til å visa kor viktig det er med autorisert namngjeving gjennom det som heiter eit URI-register.

1.4 Utfordringar knytt til betre samhandling

Innhald frå kultursektoren er tilgjengeleg i mange format og det er semantisk heterogent, publisert ved ulike institusjonar og likevel fullt av overlapp og slektskap. Sett frå synsstaden til ein brukar ville det vore nyttig om innhald knytt til eit emne, person eller lokalitet kunne vorte samla og integrert for å skapa ein rikare og meir saumlaus presentasjon. Dette er muleg berre viss samankoplinga av heterogent innhald kan oppnåast enten på eit syntaktisk eller eit semantisk nivå.

Det er to viktige måtar å møta samhandlingsproblem (problem med manglande interoperabilitet) på:

- prøva å unngå problema når innhaldet blir skapt
- prøva å løysa problemet i etterkant når innhaldet blir samla og applikasjonar utvikla

Utvikling av standardar og harmonisering av innhald, metadata, vokabular og rutinar for katalogisering er viktig for å unngå samhandlingsproblem. Innhaldsutvikling kan støttast av verktøy som delte ontologiar og databasar for ulike metadata-format. Harmonisering av innhald er likevel den beste strategien for å hindra manglande elektronisk samhandling. Men i praksis er dette berre muleg å oppnå til ein viss grad. Resultatet blir eit stort behov for etterarbeid (post-prosessering) for å skapa slektskap mellom isolerte innhaldssamlingar.

Syntaktisk og semantisk samhandling (interoperabilitet)

Syntaktisk samhandling betyr at data er representerte på same format eller struktur. Det krev at dei same felta blur brukte i metadata-strukturen og at verdiane er oppgitte på same format. Eit eksempel på syntaktisk samhandling er at eit personnamn må skrivast på forma ”førenamn” og ”etternamn”.

For semantisk samhandling der innhaldet blir representert ved å bruka ontologiar og system for metadata, er det ulike tilnærmingar:

1. *Skjema-samhandling (Schema interoperability)*: Korleis to ulike metadata-system kan samhandla. Til dømes kan ”fotograf” og ”målar” definerast som relaterte fordi begge er kunstnar-yrke.

2. *Samhandling ved hjelp av vokabular (Vocabulary interoperability)*: I dette tilfellet kan vi ha metadata-verdiar knytt til innhald frå ulike organisasjonar der verdiane er henta frå ulike vokabular. Samanhengen mellom vokabulara er kanskje ikkje uttrykt. Eit døme er begrepet ”foto” brukt i eitt vokabular, medan ”bilete” er brukt i eit anna. Vi har i tillegg fleire utfordringar som synonym og homonym, språk og målformer.

Kultursektoren omfattar område som sine eigne thesauri og vokabular (bibliotek, museum, arkiv). Sjølv innan eit bestemt emne kan ulike ikkje-samhandlande vokabular vera brukte. Det er minst tre ulike tilnæringar for å oppnå samhandling gjennom metadata-skjema:

1. Eit “kjerne”-skjema som inneheld felles delar av alle vokabular-skjema som er brukte. Seinare kan meir detaljerte skjema (= applikasjonar) leggjast til kjernen ved å introdusera nye felt og detaljera eksisterande. Denne tilnærminga har vorte brukt av Dublin Core Metadata Initiative (DCMI). Her er t.d. ”date” eit DC-element som kan spesifiserast vidare i “date published” eller “date last modified”. Detaljeringa vil behalda koplinga til det generelle elementet. Koplinga mellom ein kjerne-eigenskap og ei detaljering av denne, kan representerast i RDFS ved å bruka eigenskapen rdfs:subPropertyOf.

2. Det er muleg å definera ein harmoniserande ontologi eller skjema som kan representera alle metadata-skjema som skal integrerast. Semantisk samhandling på skjema-nivå blir då oppnådd ved å transformera metadata frå ulikt hald inn i den harmoniserande ontologien.

Meir om semantisk teknologi generelt i neste kapittel.

Semantiske teknologiar i kultursektoren

Semantiske portalar[3] har vore ein dominerande type semantisk teknologi i kultursektoren. Av mange eksempel kan nemnast MuseumFinland[4] som presenterer artefaktar frå ulike museum, MultimediaN E-Culture demonstrator[5] som presenterer kunst og kunstnarar frå ulike museum, CultureSampo[6] som presenterer alle slags kulturelle objekt (artefaktar, personar, kunst, kart, musikk m.m.), CHIP[7] for personifisert mobil tilgang til kunstsamlinga og Mobile DBpedia[8] for mobil tilgang til lenka data. System som Wikipedia (gjennom DBpedia) og Freebase inneheld store mengder semantisk kopla kulturelt innhald.

Eit vanleg mål for kulturbaserte portalar er å prøva å skapa ein global oversikt over kultursamlingar rundt om på veven, som om samlingane var ein sjølvstendig database. Denne idéen, utvikla i forskingsprosjekt, er også omfamna på internasjonalt hald i prosjekt/program som European Digital Library og Europeana. Europeana gjer det muleg for folk å utforska digitale, kulturbaserte ressursar frå alle europeiske museum, bibliotek og arkiv. Trass i forsøket på å binda saman likelydande informasjon, er Europeana likevel basert på tradisjonell teknologi og ikkje semantisk teknologi. Det finst likevel ein demonstrator som viser Europeana-innhald basert på semantisk søk. Det er basert på MultimediaN E-Culture-plattformen (http://eculture.cs.vu.nl/europeana/session/search).

1.5 Semantisk teknologi

Tittelen på prosjektet er ”Semantisk samhandling i kulturformidlinga”. Semantisk samhandling betyr at vi gjer bruk av semantisk teknologi. Det finst fleire standardar for semantisk teknologi, m.a. semantisk web (Semantic Web) og emnekart (Topic Maps). I dette prosjektet brukar vi semantisk web-teknologi, som er W3C sin tilrådde standard på området, og den spesielle retninga *lenka data* (*Linked Data*), sjå neste kapittel.

Semantisk web og lenka data baserer seg på to hovudelement:

- Resource Description Framework (RDF)
- URI-ar

Tim Berners-Lee, oppfinnaren av verdsveven, var også den første som formulerte visjonen for den semantiske veven. Han såg for seg at ulike objekt (*entities*) og relasjonar mellom dei, skulle beskrivast

[3] Hyvönen, E.: Semantic portals for cultural heritage. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, 2nd edn., pp. 757–778. Springer, Dordrecht (2009)

[4] Hyvönen, E., Ma¨kela, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., Kettula, S.: MuseumFinland – Finnish museums on the semantic web. J. Web Semant. 3(2), 224–241 (2005)

[5] Schreiber, G., Amin, A., Aroyo, L., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Omelayenko, B., van Ossenbruggen, J., Tordai, A., Wielemaker, J., Wielinga, B.J.: Semantic annotation and search of cultural heritage collections: The MultimediaN E-Culture demonstrator. J. Web Semant. 6(4), 243–249 (2008)

[6] Hyvönen, E., Mäkelä, E., Kauppinen, T., Alm, O., Kurki, J., Ruotsalo, T., Seppälä, K., Takala, J., Puputti, K., Kuittinen, H., Viljanen, K., Tuominen, J., Palonen, T., Frosterus, M., Sinkkilä, R., Paakkarinen, P., Laitio, J., Nyberg, K.: CultureSampo – Finnish culture on the semantic web 2.0. In: Proceedings of the Museums and the Web (MW 2009), Indianapolis (2009)

[7] van Hage, W.R., Stash, N., Wang, Y., Aroyo, L.: Finding your way through the Rijksmuseum with an adaptive mobile museum guide. In: The Semantic Web: Research and Applications, Seventh Extended SemanticWeb Conference (ESWC 2010), Proceedings, Part I, Heraklion. Lecture Notes in Computer Science, vol. 6088, pp. 46–59. Springer, Berlin (2010)

[8] Becker, C., Bizer, C.: DBpedia mobile: a locationenabled linked data browser. In: Proceedings of the FirstWorkshop about Linked Data on the Web (LDOW 2008), Beijing

gjennom bruk av RDF. RDF omformar den abstrakte modellen ved å bryta informasjonen ned i små bitar og bruka enkle reglar for semantikk for å seia noko om kvar av desse bitane. Målet er å komma fram til ein generell modell som er enkel og fleksibel nok til å uttrykkja alle slags fakta, likevel strukturert nok til at datamaskiner kan behandla den uttrykte kunnskapen.

Denne abstrakte modellen har følgjande sentrale delar:

- påstand (*statement*), formelt kalla ein triplett (*triple*)
- ressursar som subjekt og objekt
- predikat

Frå norskundervisinga hugsar vi kanskje analysar av setningar med *subjekt*, *predikat* og *objekt*:

Tim Berners-Lee	fann opp	verdsveven
subjekt	*predikat*	*objekt*

Dette er kanskje svake (og ubehagelege?) minne for mange, men når vi arbeider med semantisk vev og RDF, må vi friska opp att denne kunnskapen. Det er akkurat det same vi snakkar om, sjølv om ein i norskundervisinga i dag brukar begrep som *verbal* i staden for predikat.

Subjekt og objekt er namn på "ting i verda" og predikatet er namnet på relasjonen mellom subjektet og objektet. Denne relasjonen er av og til også kalla eigenskap (*property*) og det skaper ein god del forvirring at det same begrepet kan ha fleire namn. Som modellen viser, er RDF ein enkel måte å uttrykkja påstandar (statements) om ting på veven på ein slik måte at maskinene kan behandla det automatisk.

RDF brukar Uniform Resource Identifiers (URIs) som namn for å skilja ressursar frå kvarandre. URI-ar er ei meir generell form for namngiving enn URL-ar, elles er dei like. Det er muleg å bruka ulike URI-ar for den same ressursen på same måten som det er muleg å bruka same URI-en for ulike ting. Dette skaper sjølvsagt problem for tolkinga. Ein måte å unngå slik forvirring er ved å bruka vel-definerte vokabular, dvs. strukturerte ordlister der orda er definerte og ein er einige i tydinga av dei (ofte blir *ontologi* også brukt om vokabular). Eit eksempel på eit omforeint vokabular er Dublin Core. Sjå elles vedlegget "Vocabularies – Basics and Guidelines" for meir generell informasjon om vokabular og eksempelt på mykje brukte vokabular.

1.6 Lenka data

I prosjektet er det sagt at teknologien *lenka data* (Linked Data) skal brukast. Lenka data er ei retning av den semantiske web-en som inneber ei forenkling i høve dei opphavelege tankane om semantisk web og semantisk teknologi. Lenka data byggjer på dei same grunnsteinane som semantisk web og hovudprinsippa, slik dei er formulerte av Tim Berners-Lee[9], er:

1. Bruk URI-ar som namn for ting
2. Bruk HTTP (HTTP URIs) slik at folk kan slå opp namna

[9] Tim Berners-Lee. Linked Data - Design Issues, 2006. http://www.w3.org/DesignIssues/LinkedData.html

3. Gi nyttig informasjon for dei som slår opp ein URI og bruk standardar (RDF, SPARQL)
4. Lenk til andre URI-ar slik at folk kan oppdaga nye ting

Der lenka data er ope tilgjengelege, brukar vi begrepet lenka opne data (Linked Open Data – LOD[10]).

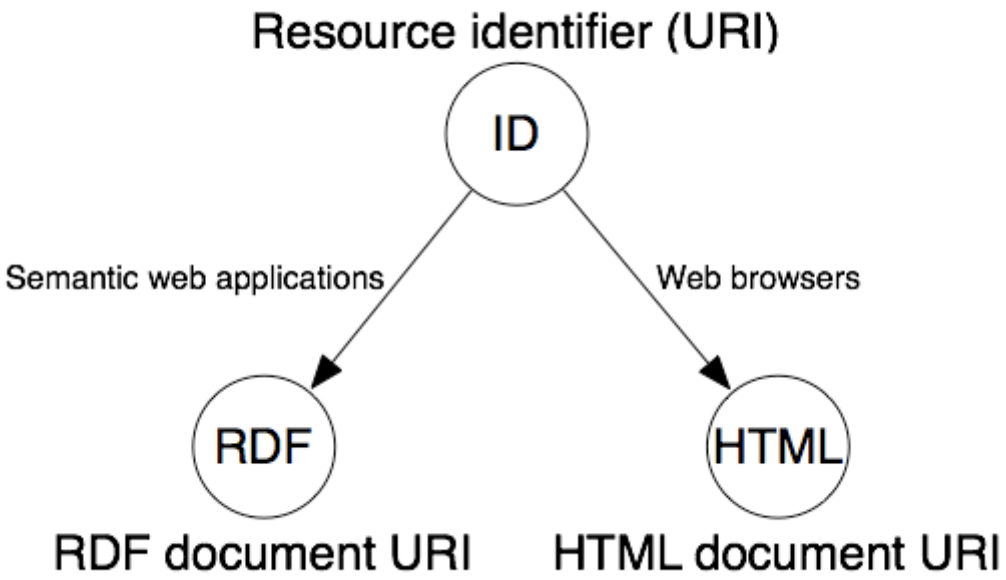
1.7 Kva er eit URI-register?

URI står for *Uniform Resource Identifiers* og er ein W3C-standard for å identifisera ressursar. I standarden RFC 2396 (”Uniform Resource Identifiers (URI): Generic syntax”) er URI definert slik:

> A Uniform Resource Identifier (URI) is a compact string of characters for identifying an abstract or physical resource.

URI-en blir adressa til ein ressurs som kan vera abstrakt eller fysisk. Dei mest vanlege adressene er av type http (URL-ar), men også andre skjema (scheme) kan vera aktuelle (ftp://, urn:// m.fl.).

Vi er mest vande med URI-ar til dokument på nettet (URL-ar). Med ei stadig utvikling av den semantiske veven må vi i aukande grad venja oss til å tenkja at URI-ar også kan referera til fysiske ting, som t.d. personar, organisasjonar, andre fysiske objekt. Vi må då skilja mellom det fysiske objektet og ressursen på veven som omtalar objektet.



Figur 3: Samanhengen mellom referanse til fysisk objekt og til dokument som omtalar objektet

Skiljet mellom ressurs og omtale (dokument) er svært viktig for forståinga av den semantiske veven og lenka data. I figuren over er ”resource identifier” det som blir kalla **ID-URI**, medan RDF- og HTML-

[10] http://linkeddata.org

ressursane er **dokument-URI-ar**. Ein ressurs kan vera kva som helst, og som det heiter om definisjonen av *subject* i emnekart-modellen (Topic Maps)[11]:

> "can be anything whatsoever, regardless of whether it exists or has any other specific characteristics, about which anything whatsoever may be asserted by any means whatsoever"

Stort opnare kan det ikkje bli, og slik er det med ressursar i semantisk vev-samanheng; det kan vera absolutt alt mellom himmel og jord. Det viktige er at alle ressursar får ein representasjon på nettet og kan refererast til.

I prosjektet vårt etablerer vi eit URI-register for fotografar. Fotografane kan sjølvsagt ikkje fullt ut representerast på nettet, men dei kan refererast til som fysiske personar ved hjelp av ein ID-URI. Og ein annan URI vil kunna gi oss ein omtale av ein enkelt fotograf. Det siste blir gjort ved å presentera ein ressurs for menneske (HTML) og ein for maskiner (RDF). Figur 1 viser samanhengen mellom ID-URI og dokument-URIar. Eit eksempel frå DBpedia for ressursen 'Sogndal':

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

Den første ressursen er den fysiske ressursen 'Sogndal' sin representasjon på nettet. Dei to andre URI-ane er omtalar av denne ressursen, både i menneskeleg (html) og maskinell (rdf) form. Merk at URI-tilrådingane (sjå under) tilseier at ID-URI blir merka med 'id' i staden for 'resource'.

1.8 Korleis bør URI-ar utformast?

Sjølv om vi står relativt fritt til å utforma URI-ar, gitt at vi har kontroll over det domenet som inngår i adressa, er det ikkje likegyldig korleis vi namngir dei. Det finst retningslinjer for god URI-utforming, og tilrådingane under byggjer på W3C si tilråding ("Cool URI's for the Semantic Web") og prosjektet SemiColon II sine retningslinjer for namngiving av URI-ar i Norge[12].

1. **Eigarskap og opphav**
    a. URI-ar bør vera varige (*persistent*) jf. Tim Berners-Lee: "Cool URIs don't change"
    b. Bruk berre namnerom du kontrollerer (dine eigne domene)
    c. Sektor bør visast i strukturen, men ikkje etatsnamn (heller http://kultur.data.norge.no enn http://kulturradet.data.norge.no)
    d. Unngå implementeringsdetaljar (t.d. filtype, portnummer o.l.) FoaF-vokabularet kan tena som eit døme på korleis det ikkje må gjerast: http://xmlns.com/foaf/0.1 (som dei sjølve seier: " We are left with the digits "0.1" in our URI. This stands as a warning

[11] ISO 13250 (Topic Maps): ISO/IEC 13250-2 5.3.1
[12] "URI-er for begreper og data i norsk offentlig sektor" - Rapport under utarbeiding av Semicolon II-prosjektet

to all those who might embed metadata in their vocabulary identifiers."

2. **Informasjonsberande URI-ar**
    a. Mal for ID-URI (instans):
    http://{domene}/id/{term}/{referanse}
    http://{domene}/page/{term}/{referanse} (leseleg for menneske)
    http://{domene}/data/{term}/{referanse} (leseleg for maskiner)
    b. Mal for ID-URI (begrep):
    http://{domene}/vocab/{vokabular}/{term}
    http://{domene}/page/{vokabular}/{term}
    http://{domene}/data/{vokabular}/{term}

Det er ulike teknikkar for oppslag mot denne typen URI-ar, såkalla "303 redirect" og "hash". Vi går ikkje nærmare inn på det, men interesserte kan lesa "*Cool URI's for the Semantic Web*"[13] eller "*Linked Data – Evolving the Web into a Global Data Space*"[14] for nærmare omtale av teknologien.

1.9 Semantisk løfting: Tilføring av metadata

I dei fleste tilfella, også i dette prosjektet, tek ein utgangspunkt i eksisterande data. Det er ofte strukturerte data i form av ein database. Då har ein alt mykje tilleggsinformasjon (= metadata) som det er viktig å få med seg og få uttrykt eksplisitt. Problemet med datakjelder basert på databasar er at tilleggsinformasjonen går tapt på vegen. Det same ser vi når XML er utgangspunktet for HTML-presentasjonar: Strukturinformasjonen blir strippa vekk i omforminga til HTML-kode.

Når vi løfter data semantisk er det den omvendte prosessen; vi trekkjer ut eksisterande struktur-informasjon om synleggjer den. I tillegg er det ofte nødvendig å tilføra meir metadata, noko som blir kalla annotering.

Kva vokabular?

Det finst rikeleg litteratur om klassifisering av metadata. Eit viktig skilje går mellom beskrivelse av eigenskapane til objektet, la oss seia eit fotografi (som fotograf, tid, stad osv.) og eigenskapane til det avbilda objektet (tingen, personen, konseptet som er avbilda). I Standard for fotokatalogisering[15] blir dette omtalt som skiljet mellom motiv og eksemplar.

Den typiske annoteringa gir tilleggsinformasjon (metadata) av typen tittel, opphavsmann (*creator*), format, opphavsrett, dato for publisering m.m. Det er vanleg praksis å nytta eit felles, førehands-definert og relativt avgrensa vokabular for å beskriva slike eigenskapar. Eksempel på slike vokabular er Dublin Core[16] og VRA Core[17]. Men vokabulara som blir nytta kan variera frå å vera svært domene-

[13] http://www.w3.org/TR/cooluris/
[14] http://linkeddatabook.com/editions/1.0/
[15] Standard for fotokatalogisering, ABM-skrift #44, http://www.abm-utvikling.no/publisert/abm-skrift/abm-skrift-44-fotokatalogisering
[16] http://www.w3.org/2005/Incubator/mmsem/XGR-image-annotation/#DublinCore

spesifikke (t.d. særskilte vokabular for kulturbaserte foto, eller sportsfoto) til domene-uavhengige og generelle vokabular som kan beskriva alle typar foto, eller alle typar kulturelle ting. Vokabular kjem såleis i alle storleikar, granularitetar og bruksomfang.

Vokabular og meta-vokabular

Det er heller ikkje uvanleg at eit vokabular berre definerer eigenskapane og overlet definisjonane av verdiane til eigenskapane til eit anna vokabular. Eit typisk eksempel er Dublin Core. Opprinneleg vart Dublin Core definert med 15 (grunn-)element. Eitt av desse er *subject*. Her er det vanleg å leggja til ord frå spesielle vokabular, for Dublin Core tilbyr ikkje noko spesielt vokabular for innhaldet. På denne måten fungerer Dublin Core som eit meta-vokabular; det kan hjelpa oss til å seia ein del om objektet (tittel, beskrivelse, identifikator, subjekt ++), men kan ikkje hjelpa oss til å leggja til informasjon om sjølve innhaldet. Dersom objektet er eit fotografi av Knut Hamsun, må vi ty til eit anna vokabular, t.d. Friend of a Friend (FOAF) for å beskriva personen Knut Hamsun.

Standardar ikkje utforma med tanke på samhandling

Også Standard for fotokatalogisering kan oppfattast som eit meta-vokabular fordi det ikkje gir støtte for annotering av sjølve innhaldet/motivet. Felt 13 i standarden (obligatorisk) er 'Emneord'. Her heiter det i omtalen:

> "Her beskrives elementer av eller helheten i motivinnholdet i enkeltbildet, serien eller samlingen/arkivet. Emneord hentet fra kontrollerte emneordlister eller fritt tildelte stikkord/nøkkelord kan brukes. Dersom bevaringsinstitusjonen eller registreringssystemet bruker/utvikler en kontrollert emneordliste, skal emneord hentes fra denne."

Innhaldsmessig er ein altså like langt, og resultatet blir gjerne eit tilfeldig eigeprodusert emneord som ikkje gir kopling til andre annoteringar med same standarden. Dermed kan utveksling og samhandling berre skje på bakgrunn av eksemplar-felta og ikkje på bakgrunn av motivet, som ofte er det viktigaste. Standard for fotokatalogisering viser kor vanskeleg samhandling blir i praksis fordi utforming av slike standardar ikkje har utveksling som det primære målet, men mest muleg presis beskrivelse av objektet (i dette tilfellet eit fotografi eller ei samling). I omtalen av standarden er det rett nok teke med forslag til mapping mot meir kjende standardar som Dublin Core og CIDOC CRM, men forslaga viser at det ikkje er heilt enkelt å få til i praksis.

Valet av vokabular for å beskriva foto er eit avgjerande val i prosjektet. Som vist ovanfor trengst det vanlegvis meir enn eitt vokabular for å dekkja alle relevante aspekt ved eit foto (eller ein annan kulturell ting for den saks skuld). Vedlegget "Vocabularies – Basics and Guidelines" gir ei nærmare innføring i aktuelle vokabular for annotering av fotografar og foto.

Mange av dei relevante vokabulara vart utvikla før den semantiske web-en. Ein viktig internasjonal standard som the Multimedia Content Description Standard (betre kjent som MPEG-7) vart til dømes uttrykt ved hjelp av XML Schema.

[17] http://www.w3.org/2005/Incubator/mmsem/XGR-image-annotation/#VraCore

Eit aktuelt vokabular er VRA Core (sjå vedlegg 1). Der Dublin Core (DC) spesifiserer eit avgrensa og vanleg brukt vokabular for online-ressursar generelt, definerer VRA eit liknande sett spesielt retta mot visuelle ressursar. Både DC og VRA Core kallar termane i vokabularet for *elements* og brukar *qualifiers* for å raffinera (gi ytterlegare presisjon) elementa. Alle elementa i VRA Core har enten direkte mapping til samanliknbare element i DC, eller er definerte som spesialiseringar av eitt eller fleire DC-element. I tillegg er begge vokabulara definerte på ein måte som abstraherer frå aktuell implementering og underliggjande uttrykte språk (annoteringane kan altså gjerast i ulike språk og format). Ein viktig skilnad er at for DC finst det ein allment akseptert mapping til RDF saman med eit assosiert skjema (*associated schema*).

Mange annoteringar på veven er for heile ressursar. Til dømes angir <dc:title> heile dokumentet, ikkje delar av det. For foto er det ofte trong for å annotera delar av ressursen, til dømes ein person. Det er viktig at eit vokabular støttar denne slike behov og på den måten kan ha mange annoteringar som refererer til det same innhaldet.

2. Korleis har vi prøvt å løysa det?

I dette kapitlet går vi gjennom prosessen med semantisk løfting og publisering av URI-ar steg for steg.

2.1 Semantisk løfting: Konvertering til RDF

Projektet har tatt utgangspunkt i flere forkjellige datakilder fra flere aktører. Nasjonalbibliotekets fotografbase (PREUS museum) dannet hovedbasen, mens Oslo Byarkivet bidro med en eksport av sitt arkiv[18] (22 316 bilder) og Popsenteret tilføyde en base relatert til Norsk populærmusikk. Den siste databasen inneholder metadatabeskrivelser av mennesker, hendelser og steder, samt bilder, filmer, spor som er knyttet til disse.

De tilgjengelige datakildene hadde en vidt forskjellig karakter. PREUS-databasen foreligger som en relasjonelll database (RDBMS) som ble eksportert til XML (en fil per tabell, brutt opp i flere filer pga størrelsen). Byarkivets data stammer fra Apples bildehåndterings-programvare som tillater eksport til XML i et meget rudimentært format. Dataene til Popsenteret var i utgangspunkt lagret som RDF/S-basert graf, og trengte dermed ikke noen konvertering i motsetning til de andre datakildene. I resten av dette kapitlet henviser vi hovedsaklig til de to databasene som krevde en konvertering. Referanser til «basene» er derfor en referanse til Preus' fotografbase og Byarkivets Oslobilder-database.

2.2 Semantisk løfting: mappe «lokal» definert semantikk til «global» semantikk

For å kunne åpne opp data på internett og relatere informasjon til andre informasjonskilder og ontologier (begrepsdefinisjoner) er det viktig å forsikre seg om to ting; data må ha relasjoner og «ting» må kunne representeres med en unik identifikator (URI).

Dataene må inneholde relasjoner med en klar og spesifikk semantikk som kan beskrives.
Siden relasjoner (med en bestemt mening/semantikk) mellom «ting» danner grunnlaget for den semantiske veven, betyr dette i praksis at man sliter når man får data som IKKE har relasjoner. Relasjoner ligger naturligvis i RDBMS-basert data som navnet "relasjonsdatabaser" indikerer, men også i et regneark og eventuelt også i en CSV-fil (komma-separerte verdier). I slike tilfeller kan man ofte definere gode løftingsregler. Det er altså ikke formatet som avgjør om en konvertering blir vanskelig, men mer betydningen/semantikken i innholdet. Er dette kjent og kan gjøres eksplisitt av et menneske eller gjennom regler, så kan man ofte gjøre en god jobb på en slik konvertering.

En eksempel hvor man har hatt en utfordring er typiske SIFT-baser (Søk i Fritekst). Disse databasene inneholder ingen relasjoner og er ren tekstsøk-basert. Ofte er det definert et antall felter som har en «string»-verdi. I tilfelle NRK's musikkdatabase innholder SIFT 10 felt som beskriver et musikkspor, albumet, gruppen/artisten osv. Relasjoner finnes ikke, og siden løsningen ble brukt i over 30 år har det vært en del endringer i måten man legger inn data. Dette, og det faktum at det var et begrenset

[18] Som tidlegare nemnt vart ikkje arbeidet ved Oslo byarkiv fullført pga. jobbskifte.

antall felter, medførte at man la inn mange flere dimensjoner i et felt etter hvert. Det er ikke noen enkel jobb å «skrelle ut» data fra slike felter med inkonsistent bruk av syntaktiske midler.

I tillegg er det en utfordring å sammenligne felter for å kunne slå sammen alle forekomster av samme «ting», siden alt er«bare tekststrenger». Man kan ikke garantere at en forekomst av «Peer Gynt» er det samme som en annen forekomst av samme streng et annet sted i en base. Det finnes ikke noen unike identifikatorer som tillater den konklusjonen. Dette innebærer at man må bruke skjønn og eventuelt programmere noen regler som gjør at slik sammenslåing kan utføres. Slike utfordringer har man ikke i samme grad i databaser som er bygget på prinsippene til semantisk web fra start.

Identifikatorer for objekter (ting)

I en semantisk web sammenheng trenger man en identifikator (URI). I mange tilfeller kan man gjenbruke interne nøkler (f.eks RECID=23151) og lage en URI av det som er unik. Ved korte tall kan det være nyttig å bruke en generator av tilfeldige tall e.l. Da kan man generere URI-er som f.eks ser slikt ut: «http://data.sfj.no/nb/person456af158_1 ».

Løfting av eksisterende databaseskjemaer

En annen mulighet (som vi ikke har hatt behov for i dette prosjektet) er løfting av eksisterende databaser gjennom selve datamodellen. Det finnes flere verktøy (bl.a. fra Oracle) som tillater en mapping fra Entity-Relationship diagrammet til en RDF/S & OWL-modell som publiseres (f.eks gjennom et SPARQL-endepunkt). På denne måten kan man publisere innholdet fra eksisterende baser mens man tar vare på muligheten for å beholde etablerte arbeidsprosesser og redigerings-verktøy for basen. I noen tilfeller kan dette være et fornuftig valg.

Vokabular og gjenbruk

Når man publiserer sine data ved hjelp av semantisk teknologi er det anbefalt å ikke definere vokabularer og terminologi selv. Det finnes mange ontologier og vokabularer som er mye brukt og som dekker behovet. Ved å (gjen-)bruke disse blir det mulig å automatisk hente inn/slå samme data fra eksterne kilder, siden resonneringsalgoritmene kan anta at dine egne objekter og eksterne objekter bruker samme begrep/definisjoner og har samme betydning. Det kan bety at du f.eks. finner mer informasjon om et sted du har i databasen din (f.eks geo-location, eller værdata) automatisk.

Hva gjorde vi?

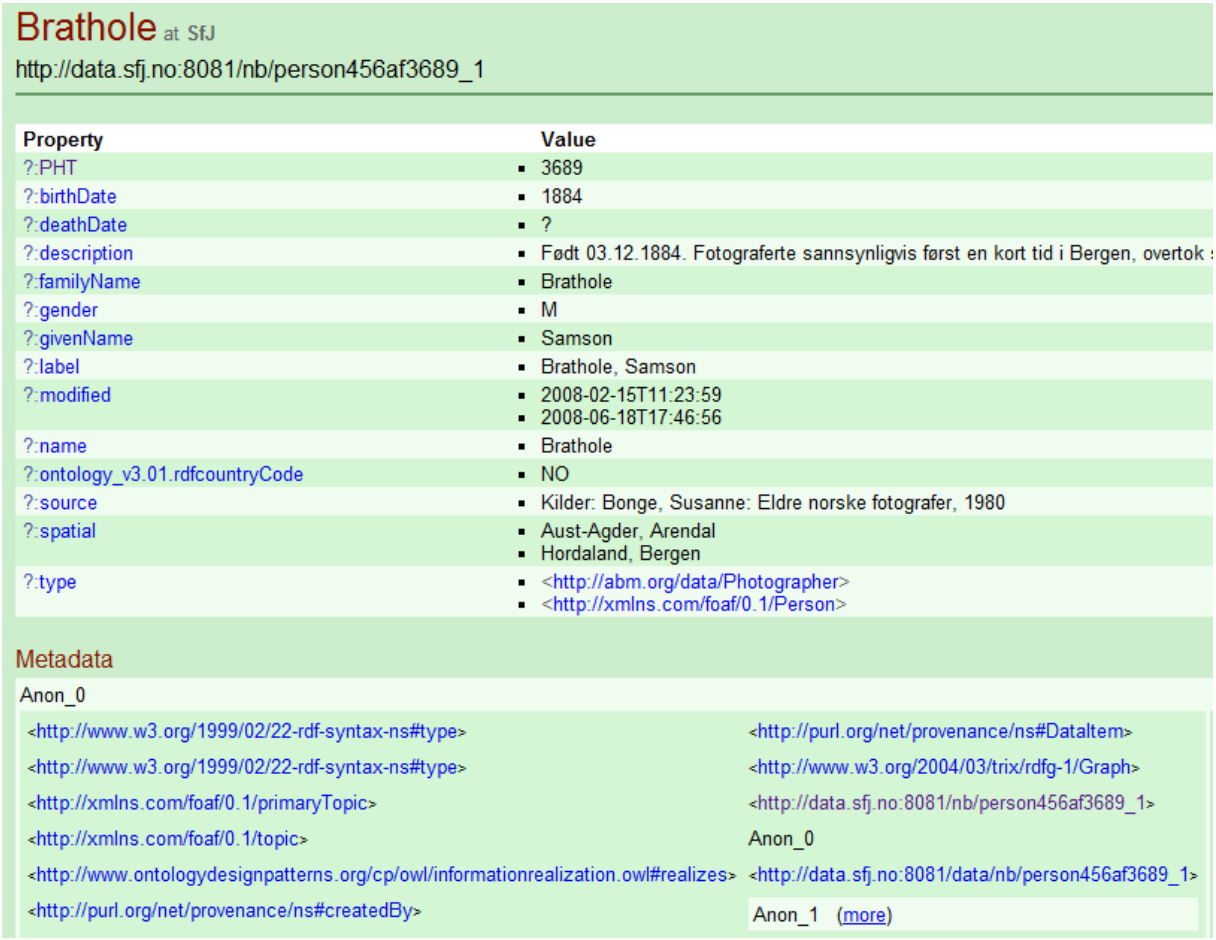
I vår tilfelle fikk vi data på XML-format og et naturlig valg var å bruke XSLT for transformering til andre formater. Så lenge det var et en-til-en-forhold mellom XML-felter og en eller annen RDF-egenskap (property), gikk dette fint. Dessverre var det også behov for en del datavask, pluss at det var flere felter som til sammen hadde en bestemt betydning. For disse tilfellene utviklet vi et eget lite rammeverk basert på PERL og XSLT som tok seg av pre-prosessering av data, selve konverteringen og en post-prosessering (ta ut duplikater, ta bort noen rare karakterer osv).

2.3 Publisering av semantisk løfta data

Ferdig annotert RDF-data vart så overførte til SPARQL-endepunktet http://data.sfj.no:2020/query.html og er tilgjengelig derfra. I dette prosjektet har vi valgt å bruke virtuelle servere med fast IP og en LAMP-stack (Linux,/Apache/MySQL/PHP).

Vi har brukt SPARQL-endepunktet til Apache prosjektet JENA. Oppsettet er godt dokumentert på hjemmesidene til JENA-prosjektet og burde kunne gjennomføres uten alt for store vanskeligheter: http://jena.apache.org/

For å kunne tilby et enkelt «inspeksjons»grensesnitt for demonstrasjon og som hjelp for programmering valgte vi Pubby (http://www4.wiwiss.fu-berlin.de/pubby/)
Her er et eksempel på data-presentasjon i Pubby:



Figur 4: Eksempel på presentasjon av data om ein person i Pubby (leseleg versjon av informasjon frå eit SPARQL-endepunkt)

2.4 Utviding av originaldata med URI-ar og publisering

Som utgangspunkt for URI-ene har originalnøklene fra RDBMS-en blitt brukt. På denne måten er det mulig å ta hver tabell fra basen for seg og konvertere den til RDF. Selve sammenslåingen skjer automatisk når alle tabeller har blitt konvertert og tilføyd RDF-basen (RDF repository). Datasettet framstår da som en graf med unike identifikatorer for objekter og relasjoner.

2.5 Utvikling av web-demonstrator for semantisk samhandling

Web-demonstratoren er dokumentert på nettsida http://sparql.devcloud.acquia-sites.com/

2.6 Utvikling av demonstrator på Pop-senteret

På Popsenteret har det blitt utviklet en demonstrator som bruker data fra forskjellige grafer og viser de som en helhet til publikumet. Det blir hentet kalenderdata (konserter osv), informasjon om artisten, informasjon om utestedet/sted for konserten, og eventuelt filmer/bilder fra de forskjellige basene. Disse blir satt sammen automatisk (resonnering) og visualisert (egenutvikling) slik at det kan vises på ulike måter i senteret (programtavler osv) og på senterets ytrevegg. På sikt er det også ønskelig med en sammenslåing med sosiale media, slikt at publikum kan kommentere/tilføye gjennom SMS og sosiale medier. En egen app er også under utvikling.

3. Kva har vi lært?

3.1 Teknologien er (relativt) enkel

Teknologien *lenka data*, som er ein slags lettvektsvariant av *semantisk web*, har modnast mykje på dei 5-6 åra som har gått sidan Tim Berners-Lee lanserte den. Det finst i dag operative tenester, både offentlege og private, basert på lenka data. Eit eksempel frå kultursektoren er Pop-senteret i Oslo sitt informasjonssystem. Eit eksempel frå privat sektor er det arbeidet Bouvet har gjort for Hafslund[19] med samordning av databasar. Den viktige delen *semantisk løfting* av data er også relativt grei, med unnatak av utfordringar knytt til vokabular (sjå eige kapittel).

Likevel skal ein ikkje underslå at det på verktøysida enno er ein del uferdig, og at særleg skalering er ei utfordring. Teknologien fungerer godt på små, avgrensa datamengder, typisk i demo-situasjonar, men enkelte verktøy har problem med å handtera større mengder. Difor er eksempel av typen Hafslund viktig sidan det der er snakk om store mengder data, om enn avgrensa til Hafslund sine eigne forretningsdata.

I dette prosjektet har utfordringane lege på andre område enn bruken av teknologien lenka data.

3.2 Organisatorisk samhandling er vanskeleg

Så langt følgjer erfaringane i godt opptrakka spor. Så må det også fort leggjast til at organisatorisk samhandling ikkje har vore ein viktig del av prosjektet. Men for å få dette til å fungera i varige løysingar, må dei ulike institusjonane trekkjast med og kjenna eit ansvar for utviklinga. I dette tilfellet må Preus Fotomuseum, Nasjonalbiblioteket og Sogn og Fjordane fylkesarkiv utvikla ei felles forståing for måla med den teknologiske endringa og vera villege til å innretta organisasjonane på det nye. Det betyr at Preus Fotomuseum må ta omsyn til den nye presentasjonen av informasjonen dei sit på, og utfordringar til informasjonskvalitet den auka merksemda fører med seg. Nasjonalbiblioteket som driftsorganisasjon av fotografbasen må på si side delta aktivt i utforminga av URI-register, ikkje minst sitt eige, og samspelet mellom URI-registra.

Som eit eksempel kan vi ta Fotobasen ved Fylkesarkivet i Sogn og Fjordane, der gjennomgang av data viste det typiske problemet med dårleg datakvalitet: Ein og same fotografen stod oppført med mange variantar av namnet. Fotografen .... stod til dømes oppført med 15 ulike skrivemåtar! Dette er eit generelt problem som ikkje har direkte samanheng med teknologien *lenka data*, men den semantiske løftinga av data som er nødvendig, synleggjer problemet med ein gong. Slik sett vil ei semantisering av data i kultursektoren ”tvinga” fram løysingar på dårleg datakvalitet. Men det er ei utvikling som kjem av at ein må sjå på data på nytt, uavhengig av kva teknologi som skal brukast.

[19] Presentert på fleire arrangement, m.a. Software 2012

3.3 Utfordringar kring vokabular er undervurderte

Det klart største problemet prosjektet har registrert, er bruken av vokabular i samband med tilføring av metadata. Når informasjon skal løftast frå ein database og over i ei semantisk løysing, må det tilførast metadata. Vi må forklara kva slags data det er snakk om. Vokabular er orda vi brukar for å beskriva data. Vokabular kan vera kontrollerte, dvs. dei systematiserte og har eit system for kontroll, eller dei kan vera heilt frie, som t.d. brukargenererte vokabular (”folksonomies”).

Det store problemet er å finna eit vokabular som både er presist nok, dvs. har dei ”rette” orda for å beskriva våre data, og som er utbreidd nok. Bruk av vokabular inneber eit kompromiss mellom utbreiing (som betyr mer generelt) og presisjon (som betyr mindre bruk). Ofte må vi ty til fleire vokabular fordi det ikkje finst eitt som dekkjer alle behova. Denne blandinga av fleire vokabular er ei utfordring i seg sjølv, sjå t.d. Guus Schreibers presentasjon ”Issues in publishing and aligning vocabularies” frå seminaret UDC 2011[20].

For kultursektoren er det viktig å identifisera kva vokabular som finst; både lokalt og nasjonalt innan sektoren, og internasjonalt på området. I vedlegget ”Vocabularies – Basics and Guidelines” er det ein gjennomgang av vokabular generelt og ein omtale av dei mest relevante for kultursektoren.

Kultursektoren må føreta ein gjennomgang av kva vokabular som eksisterer på domenet, og ut frå tilgjengeleg informasjon utarbeida retningslinjer for bruk av desse og meir generelle vokabular.

3.4 Tungt å starta frå eit null-punkt

I dette prosjektet har vi starta heilt frå botnen av med å få eksport av databasar, semantisk løfting, publisering av URI-ar og lenka data. Vi har også måtta gjera Nasjonalbiblioteket til eit ”underbruk” av Fylkesarkivet i Sogn og Fjordane gjennom URI-ane http://data.sfj.no/nb/... Det har vore nødvendig for å få realisera demonstratorane og visa potensialet i teknologien, men det er sjølvsagt ikkje ei varig løysing. På same tida som det er viktig med prosjekt som set søkjelyset på område det må gjerast eit felles løft på, er det viktig at erfaringar blir samla og omsette i ein nasjonal strategi.

Tida er inne for å formalisera ein del av dei rutinane vi har omtalt i denne rapporten. Det må utarbeidast retningslinjer for etablering av autoritative URI-register.

3.5 Samhandling er framleis akilles-hælen

Vi har i prosjektet demonstrert nytten av å opna opp isolerte databasar for å kunna knyta saman informasjon på tvers. Likevel ser vi at samhandlingsdelen (interoperabilitet) er vanskeleg, trass i gjennomført semantisering og URI-fisering. Dels er dette fordi samhandlinga bør skje på applikasjons-nivå, det må liggja ein intensjon bak samankoplinga, og dels på grunn av dei før omtalte

[20] http://www.cs.vu.nl/~guus/talks/11-udc.pdf

utfordringane med vokabular. At informasjon er publisert som lenka data på nettet, betyr ikkje at den utan vidare kan koplast med annan informasjon.

Vi er tilbake til hovudutfordringa med å setja namn på ting og at vi brukar ulike namn om (nesten) det same. Maskiner skjønnar ikkje utan vidare at "ein ting" definert med FOAF-vokabularet som `foaf:person` er det same som "ein ting" definert med Schema.org som `../schema.org/Person`. Nokon må leggja til informasjonen om at ein person definert med FOAF-vokabularet og ein person definert med Schema.org (eller eit vilkårleg anna vokabular) er same tingen; altså ikkje same personen, men at begge er *personar*

3.6 Tilrådingar for vidare arbeid

Trass i utfordringane omtalte over, er erfaringane våre gode frå arbeidet med semantisk samhandling basert på teknologien *lenka data*. Det er eit stort behov for å opna opp dei mange lukka databasane i sektoren, til liks med andre sektorar, og lenka data er ein lovande teknologi i så måte.

Arbeidet med tilpassing av informasjon til Europeana, via det norske knutepunktet Norvegiana, er eit viktig steg på vegen mot betre samhandling. Men metadata-informasjonen i Norvegiana/Europeana må styrkast og få ein tydelegare semantisk retning, dvs. baserast på tilrådde semantiske standardar. Dagens ESE-standard brukar t.d. berre streng-verdiar og ikkje URI-ar. Det føregår likevel mykje interessant arbeid for å ta Europeana-data vidare semantisk. Det er ein tung prosess, og førebels skjer mykje av utprøvinga "top-down" der ein tek utgangspunkt i ESE-basert Europeana-informasjon og konverterer til EDM. Eit relevant arbeid sett i høve vårt prosjekt, er Europeana LOD[21] der Europeana-informasjon blir konvertert til EDM og så vidare til LOD.

Ut frå erfaringane i prosjektet "Semantisk samhandling i kulturformidlinga" tilrår vi ei slik vidareføring:

1. Utarbeida generelle retningslinjer for etablering av **autoritative URI-register på ulike nivå**; nasjonalt og regionalt.

2. Føreta ein **gjennomgang av vokabulara i sektoren** der målet må vera å komma fram til vokabular som støttar samhandling. Vokabulararbeidet i kultursektoren (som i andre sektorar) har i altfor liten grad brydd seg med samhandling, men vore veldig oppteke av presisjon i katalogiseringsarbeidet. Dette er ein stor og tung prosess som dreier seg om å snu perspektivet frå ei avsendarorientering til ei mottakarorientering. Det handlar om å bli mindre intern og meir ekstern og i langt større grad tenkja på korleis informasjonen kan bli brukt av andre, jfr også prosjekttittelen "Semantisk samhandling i kulturformidlinga" (kultur*formidlinga*, ikkje kultur*sektoren*!)

[21] http://pro.europeana.eu/linked-open-data

3. Sy saman strategiane for URI-register og vokabular med vidare planar for Norvegiana/Europeana. Her må ein finna ut kor stort handlingsrom ein har nasjonalt utan at endringar i informasjongrunnlaget svekkjer leveransane til Europeana. Her er det også naturleg å sjå på den vidare utviklinga i Europeana og overgang frå dagens **Europeana Semantic Elements (ESE)** til bruk av den nye semantiske standarden **Europeana Data Model (EDM).** Her føregår det mykje relevant arbeid på Europeana-nivå, som omtalt over, og særleg interessant er diskusjonen om EDM introduserer for mykje kompleksitet på ein første veg mot lenka data (Haslhofer & Isaac, 2011). Problema her er heller ikkje utelukkande på Europeana, nivå, men også på framleis manglande standardisering på W3C-nivå.

4. Organisatorisk samhandling må takast på alvor. Dette er kanskje den viktigaste og vanskelegaste delen av arbeidet. I første omgang er det viktig at sentrale institusjonar blir samla og at ein får ei felles forståing for utfordringane og hovudretninga vidare. Så må den enkelte institusjonen gjera ein innsats for at ein til saman skal kunna oppnå målet om ei betre samhandling for ei betre kulturformidling.

Vedlegg 1: Vocabularies – Basics and Guidelines

Abstract

Many applications that process image/photograph assets make use of some form of metadata that describe the cultural heritage content. The goal of this document is to explain the basics of image metadata. In addition, it provides guidelines for Semantic Web-based image annotation. Relevant vocabularies are discussed.

Content

1. Introduction

We need a standard format for contributing content (photograph, article etc) to cultural heritage repository, because:

- increasing relevance of internet presence for museums and other collections
- increasing necessity to integrate your data in online services: facilitate resource discovery in a cross collection and even cross sectoral (archives, libraries, museums) manner
- Need for a convenient instrument to provide core data on heritage objects
  - from different collections / object classes
  - from different data structures
  - from different software systems

Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage heritage information resource. Metadata is often called data about data or information about information. The term metadata is used differently in different communities. Some use it to refer to machine understandable information, while others use it only for records that describe electronic resources. In the library environment, metadata is commonly used for any formal scheme of resource description, applying to any type of object, digital or non-digital.

There are three main types of metadata:

- Descriptive metadata describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.
- Structural metadata indicates how compound objects are put together, for example, how pages are ordered to form chapters.
- Administrative metadata provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it.

Metadata can be embedded in a digital object or it can be stored separately. Meta data is often embedded in HTML documents and in the headers of image files. Storing metadata with the object it describes ensures the metadata will not be lost, obviates problems of linking between data and metadata, and helps ensure that the metadata and object will be updated together. However, it is impossible to embed metadata in some types of objects (for example, artifacts). Also, storing metadata separately can simplify the management of the metadata itself and facilitate search and retrieval. Therefore, metadata is commonly stored in a database system and linked to the objects described.

2. Types of vocabularies

2.1 Folksonomy

A *folksonomy* is a system of classification derived from the practice and method of collaboratively creating and managing tags to annotate and categorize content. The main property of a folksonomy is that it captures the active language use of a community. Folksonomies do not typically have hierarchical structure or preferred terms for concepts, and they may not even cluster synonyms. They address the well-known problem of indexing data with content-descriptive

metadata. Folksonomies have become well-known through social software such as the photo-sharing platform Flickr or the video-community YouTube.

2.2 Vocabulary

A **vocabulary** is a list of words that have been enumerated explicitly. All terms in a vocabulary should have an unambiguous, non-redundant definition. A vocabulary may have no meaning specified, or it may have very detailed definitions for each term.

Expressivity of vocabulary is limited to RDF Schema plus selected OWL features, e.g. inverse functional properties and class disjointness. Their value is in providing common terminology for exchanging information between programs. The actual information is in the RDF instance data that is expressed with the vocabulary's terms. A vocabulary is created by publishing a description of its terms in natural using HTML or formal using RDFS/OWL language. Since classes and properties are identified by URIs, it is considered a good practice to make these URIs resolvable. This enables clients to look up definitions of the vocabulary terms, with the following benefits: Information publishers can refer to a specification. This is important to create interoperability around a vocabulary.

2.3 Knowledge Organisation Structure

A knowledge organization system (KOS) consists of vocabularies, thesauri, classification, terminologies, etc. The structure of these vocabularies changes from time to time. Thus, it creates vocabularies heterogeneity problem.

2.3.1 Classification and Taxonomy

Based on its Greek roots, taxonomy is the science of classification. Originally, it referred only to the classifying of organisms. Now, it is often used in a more general setting, referring to the classification of things or concepts, as well the schemes underlying such a classification. In addition, taxonomy normally has some hierarchical relationships embedded in its classifications.

Typically, these are organized by subtype-supertype relationships, also called parent-child relationships. For example “Photograph" is subtype of “Creative Work".

A small example of taxonomy of tourist attractions is given below,

```
Attraction
      Human artefact
            Museum
            Visitor centre
            Art gallery
      Architecture
            Stave church
            Stone church
```

The leaf terms in this taxonomy are the primary terms. The other terms are secondary terms that can be introduced by terminological definitions.

Taxonomy is a type of classification or directory that is used by a library for cataloging books or information, by company for presenting products for sale, or by the web for indexing information for easy navigation, e.g., Google, Yahoo, DMOZ, and LCC etc. These classifications are hierarchies of folders identified by labels. The semantics of these folders is given by the items they ultimately contain. Obviously, each independent entity tends to develop its own directory based on its own needs and tastes.

2.3.2 Thesauri

Thesauri are a kind of KOS, where specialists group terms together by judging their similar meaning. Terms are the basic unit for building thesauri. They are categorized into descriptor (also called preferred term) and non-descriptor (non-preferred term). A descriptor term is a term that is used for con- trolling the indexing in the thesauri and the rest of terms are considered non-descriptor terms. In thesauri, terms are associated with each other by relationships. These relationships can be divided into three types:

1. Hierarchical relationships include broader terms (BTs) and narrower terms (NTs). BTs or hyperonym are more general terms, e.g. “Creative work" is a broader term then “Creative product". Similarly, a narrow Term (NT) or hyponym is a more specific term, e.g. “Creative product" is narrow term then “Creative Work". Both of them are associated with class type relationships, as well as “IS-A" relationships.
2. Equivalency relationships are used primarily to connect synonyms and near-synonyms.
3. Associative relationships are used to connect two related terms whose relationship is neither hierarchical nor equivalent. This relationship is described by the indicator “Related Term" (RT). This relationship should be applied with caution, since excessive use of RT will reduce specificity in searches. The main usage of thesauri is for information retrieval. They are kinds of controlled vocabularies so they are used in indexing, tagging, subject cataloging, etc. We found these thesauri in TEXT, XML, RDF and OWL format. For example, the AGROVOC thesaurus from FAO is represented in OWL, TXT, SKOS, RDF, and SQL format.

2.3.3 Databases

In databases, data is stored in predefined tables. A database specifies the names of the tables as well as their types: the names and types of the columns of each table. A database also includes a key for each table: a subset of the columns that uniquely identifies each row. Finally, a column in a table may be specified as a foreign key pointing to a column in another table. This is used to keep referential constraints among various entities. Finally, it is worth mentioning widely used languages for specifying relational schemas, such as Structured Query Language (SQL). These support many modeling capabilities, such as user-defined types, aggregation, generalization, etc. Furthermore, RDF, SKOS and OWL have stored data in triple storage where data is stored as subject, predicate and object. To manipulate this kind of data, the SPARQL query language is used. There are some existing tool which are used for creating databases, for example, PostgreSql, Sql, and Oracle etc. for relational databases, and Jena, and Sesame, etc. for RDF storage.

2.3.4 Terminology

Terminology is a *specified vocabulary*. It consists of words and compound words that work in specific contexts. It should not be confused with “terms" in colloquial usages, the shortened form of technical terms (or term of art), which are defined within a discipline or specialty field. For example, Terms of ‘Creative Work’ mean all terms from the domain of ‘Creative Work’ are included with its labels and their definition so that people can understand the terms or concepts. It does not have any kinds of relationships like thesauri (BT, NT, RT, and UF. This is mainly used for documentation and promoting correct usage. It is not limited to a single language, it does not have any particular structures. It mainly consists of a text file with term description.

3. Challenges of Matching

The manipulation of vocabularies is a very difficult task due to different factors involved that create *heterogeneity* problems.

3.1 Factors of heterogeneity problem

- Time: vocabulary changes with times
- Place: vocabularies change with place
- Structure: vocabularies are not written in specific formats or there are no universal formats
- culture diversity: vocabularies change with culture
- different vocabulary specialists: written vocabularies can be different for different specialists with different views

3.2 Different heterogeneity

The main purpose of matching vocabularies is overcoming the *heterogeneity* problem. The problem does not lie solely in the difference of ultimate goals of the applications according to which they have been designed, or in the expression formalisms in which the vocabularies have been encoded. Defining factors creates many heterogeneity problems. We describe here some typical heterogeneity problems.

Syntactic heterogeneity occurs when two vocabularies are not ex- pressed in the same syntax as the vocabulary language. This generally happens when two vocabularies compare, for instance, a classification with a conceptual model. This also happens when two vocabularies are modeled by using different knowledge representation formalisms, for instance, RDF, OWL, or SKOS. This kind of mismatch is generally tackled at the theoretical level by establishing equivalences between constructs of different languages. Thus, it is sometimes possible to translate vocabularies between different vocabulary languages while preserving the meaning.

Lexical heterogeneity occurs due to variations in label names (descriptor terms) when referring to the same entities in different vocabularies. This can be caused by the use of different natural languages, e.g., “photo" vs. “snap".

Semantic heterogeneity occurs due to structure factors. In general, it occurs due to the use of different expressions for defining concepts and their related relationships, e.g. “Paris" is a city in

“France" or “Paris" is a name of person. This conceptualization mismatch depends on modeled concepts.

Pragmatic heterogeneity is concerned with how entities are designed by vocabularies specialists. Indeed, entities which have exactly the same interpretation are often interpreted by specialists with regard to the context. One example is how they are ultimately used. This kind of heterogeneity is difficult for the computer to detect and even more difficult to solve, because it is out of its reach. The intended use of entities has a great impact on their interpretation, therefore, matching entities which are not meant to be used in the same context is often error-prone.

Metadata heterogeneity is concerned with how data is presented in the metadata registry, with entities which have the same names but different expressions. This kind of heterogeneity problem occurs in bibliographic data expression. For example, in photographer names are expressed in different styles: “J.Heimdal" or “Heimdal, Jal". There is no specific format for this. There are many other heterogeneity problems.

4. Meta-Vocabularies

4.1 SKOS

SKOS, short for simple knowledge organization systems, is an RDF vocabulary for representing KOSs, such as taxonomies, thesauri, classification schemes, and subject heading lists. It is used to port existing KOSs into the shared space of the Semantic Web; therefore they can be published on the Web and they can be machine readable and exchanged between software applications.

SKOS is developed by W3C Semantic Web Development Working Group (SWDWG) and has an official Web site4 which contains all the information related to SKOS. It has become a W3C standard on 18 August 2009. This standard includes the following specifications:

- SKOS Reference W3C Recommendation;
- SKOS Primer W3C Working Group Note;
- SKOS Use Cases and Requirements W3C Working Group Note; and
- SKOS RDF files.

Whenever we would like to use RDF statements to describe a document, we should use the terms from Dublin Core vocabulary. SKOS is the vocabulary we should use when we try to publish a given KOS into the shared space of the Semantic Web. Note that the URIs in SKOS vocabulary have the following lead strings:

http://www.w3.org/2004/02/skos/core#

By convention, this URI prefix string is associated with namespace prefix skos: and is typically used in different serialization formats with the prefix skos.

4.1.1 SKOS Core Constructs

In this section, we will discuss the core constructs of SKOS, which will include the following:

• Conceptual resources should be identified by URIs and can be explicated as concepts.

• Concepts can be labeled with lexical strings in one or more natural languages.
• Concepts can be documented with different types of notes.
• Concepts can be semantically related to each other in informal hierarchies.
• Concepts can be aggregated into concept schemes.

These SKOS features are not all that are offered by the SKOS model, but will be enough for representing most KOSs on the Semantic Web. For the rest of this section, we will use Turtle for our examples, and the following namespaces will be needed. We now list these namespaces here so they will not be included in every single example:

```
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dc: <http://purl.org/dc/elements/1.1/>.
@prefix ex: <http://www.example.com/> .
@prefix ex1: <http://www.example.com/1/> .
@prefix ex2: <http://www.example.com/2/> .
```

Concept is a fundamental element in any given KOS. SKOS introduces the class skos:Concept, so that we can use it to state the fact that a given resource is a concept. To do so, we first create (or reuse) a URI to uniquely identify the concept; we then use one RDF statement to assert that the resource, identified by this URI, is of type skos:Concept.

For example, the following RDF statement says photograph is a skos:Concept:

<http://dbpedia.org/resource/Photograph> rdf:type skos:Concept.

Clearly, using SKOS to publish concept schemes makes it easy to reference the concepts in resource descriptions on the Semantic Web. In this particular example, for the resource http://dbpedia.org/resource/Photograph, besides everything that has been said about it, we know it is also a skos:Concept.

The first thing to know about skos:Concept is that SKOS allows us to use labels on a given concept. Three label properties are provided: skos:prefLabel, skos:altLabel and skos:hiddenLabel. They are all sub-properties of the rdfs:label property, and they are all used to link a skos:Concept to an RDF plain literal, which is formally defined as a character string combined with an optional language tag. More specifically,

- skos:prefLabel property is used to assign a preferred lexical label to a concept.

This preferred lexical label should contain terms used as descriptors in indexing systems and is normally used in a KOS to unambiguously represent the underlying concept. Therefore it is recommended that no two concepts in the same KOS be given the same preferred lexical label for any given language tag.

- skos:altLabel property is used when synonyms, near-synonyms, or abbreviations need to be represented.

- skos:hiddenLabel property is used mainly for indexing and/or searching capabilities.

For example, the character string as the value of this property will be accessible to applications performing text-based indexing and searching operations, but will not be visible otherwise.

4.2 Dublin Core

Dublin Core[22] is a set of pre-defined URIs representing different properties of a given document. Since they are widely used in RDF documents, they can also be understood as another set of pre-defined RDF vocabulary. Dublin Core was developed in the *March 1995 Metadata Workshop* sponsored by the Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA). The workshop itself was held in Dublin, Ohio, hence the name Dublin Core. Currently, it is maintained by the Dublin Core metadata Initiative project.

Dublin Core has 15 elements called the Dublin Core metadata element set (DCMES). It is proposed as the minimum number of metadata elements required to facilitate the discovery of document-like objects in a networked environment such as the Internet. Table 1 shows some of these terms.

Generally speaking, if we are using RDF to describe a document, or maybe part of our RDF document is to describe a document, we should use Dublin Core predicates as much as we can. For example, Title predicate and Creator predicate are all good choices. Note that the URIs in Dublin Core vocabulary have the following lead strings:

http://www.purl.org/metadata/dublin-core#

By convention, this URI prefix string is associated with namespace prefix dc: and is typically used in XML with the prefix dc. For example, following example is a simple RDF description about my personal Web page. The two statements use Dublin Core terms to indicate the creator of this Web site and the date this site was created (lines 8 and 9).

Example of using Dublin Core terms

```
1: <?xml version="1.0"?>
2: <!DOCTYPE rdf:RDF
2a:       [<!ENTITY xsd "http://www.w3.org/2001/XMLSchema#">]>
3:
4: <rdf:RDF
4a:   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
5:    xmlns:dc="http://www.purl.org/metadata/dublin-core#">
6:
7:  <rdf:Description rdf:about="http://www.vestforsk.no">
8:    <dc:creator>J. Heimdal</dc:creator>
9:    <dc:date rdf:datatype="&xsd;date">2011-25-11</dc:date>
10: </rdf:Description>
11:
12: </rdf:RDF>
```

[22] Dublin Core Metadata Element Set

Table 1: Element examples in Dublin Core Metadata Scheme

Element name	Element description
`Creator`	This element represents the person or organization responsible for creating the content of the resource, e.g., authors in the case of written documents
`Publisher`	This element represents the entity responsible for making the resource available in its present form. It can be a publishing house, a university department, etc
`Contributor`	This element represents the person or organization not specified in a `Creator` element who has made significant intellectual contributions to the resource but whose contribution is secondary to any person or organization specified in a `Creator` element, e.g., editor, transcriber, illustrator
`Title`	This element represents the name given to the resource, usually by the `Creator`
`Subject`	This element represents the topic of the resource. Normally this will be expressed as keywords or phrases that describe the subject or content of the resource
`Date`	This element represents the date associated with the creation or availability of the resource
`Identifier`	This element is a string or number that uniquely identifies the resource. Examples include URLs, Purls, and ISBN, or other formal names
`Description`	This element is a free text description of the content of the resource. It can be in flexible format, including abstracts or other content descriptions
`Language`	This element represents the language used by the document
`Format`	This element identifies the data format of the document. This information can be used to identify the software that might be needed to display or operate the resource, e.g., postscript, HTML, text, jpeg, XML

The second level of DC called *Qualified Dublin Core* includes three additional elements namely - `Audience`, `Provenance`, and `RightsHolder` along with qualifiers to refine the semantics of the elements. Each element in both the category is optional and may be repeated. Most elements may use qualifiers, adjectives or refinements that clarify meaning of the element.

Each metadata defined by Dublin Core emphasis on unique identification of entity. This principle is known as *One-to-One Principle*. Further, the qualification of Dublin Core properties is guided by a rule known as the *Dumb-Down Principle*. This deals with the fact that, though qualifier is present and defined by the metadata, user may not use it. In this situation, the entity is treated as unqualified entity, provided it is useful and can be tracked through discovery. For this purpose, generally qualifiers are the adjectives adding values to the base/correct meaning of the entity. Here, qualifier can be used as refining adjective to the entity. The third principle state the appropriateness of the qualifiers. One thing one must keep in mind is that not only machines, but humans also use and interpret the metadata. Hence, the qualifier of an entity must be of appropriate values and exhibits correct and same values to machines as well as to human beings.

4.3 Friend of a Friend

The *Friend of a Friend* (*FOAF*) project[23], one of the largest projects in the semantic web, is a descriptive vocabulary built based on RDF and OWL, for creating a Web of machine-readable pages for describing people, the links between them and the things they create and do. It is accepted as standard vocabulary for representing social networks, and many large social networking websites use it to produce Semantic Web profiles for their users.

FOAF has the potential to become an important tool in managing communities, and can be very useful to provide assistance to new entrants in a community, to find people with similar interests or to gather in a single place, people's information from several different resources, decentralizing the use of a single social network service for example.

The things described in the web are connect by people. People attend meetings, create documents, are depicted in photos, have friends, and so on. Consequently, there are a lot of information that might be said about people and the relations between them and objects (documents, photos, meeting, etc). FOAF describes the most common information we usually want to know about a person and because it is built upon RDF, it also uses some vocabulary from other resources, such as the Dublin Core (DC).

In the Figure 1 we can see a summary of what FOAF stands for. The base class described in FOAF is the `foaf:Agent` class. The Agent class describes "the things that do stuff" and have `foaf:Group`, `foaf:Person` and `foaf:Organization` as sub-classes. FOAF describes resources such as `foaf:Document`, `foaf:Image` or `foaf:OnlineAccount` and people properties like `foaf:name`, `foaf:title` or `foaf:mbox`.

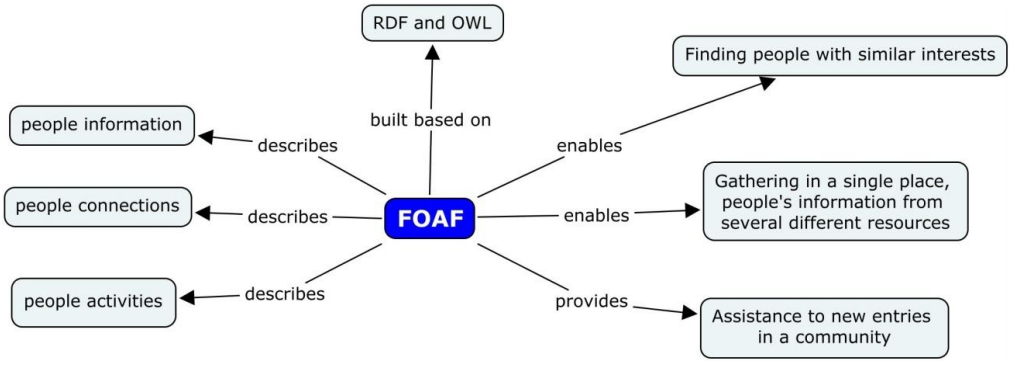


Figure 1: Description of the FOAF importance

The following figure 2 shows a list of the FOAF classes and properties. One important property that should be mentioned is the `foaf:knows` property. It can be used to link two people together. FOAF identifies other people by stating their properties.

[23] http://www.foaf-project.org/

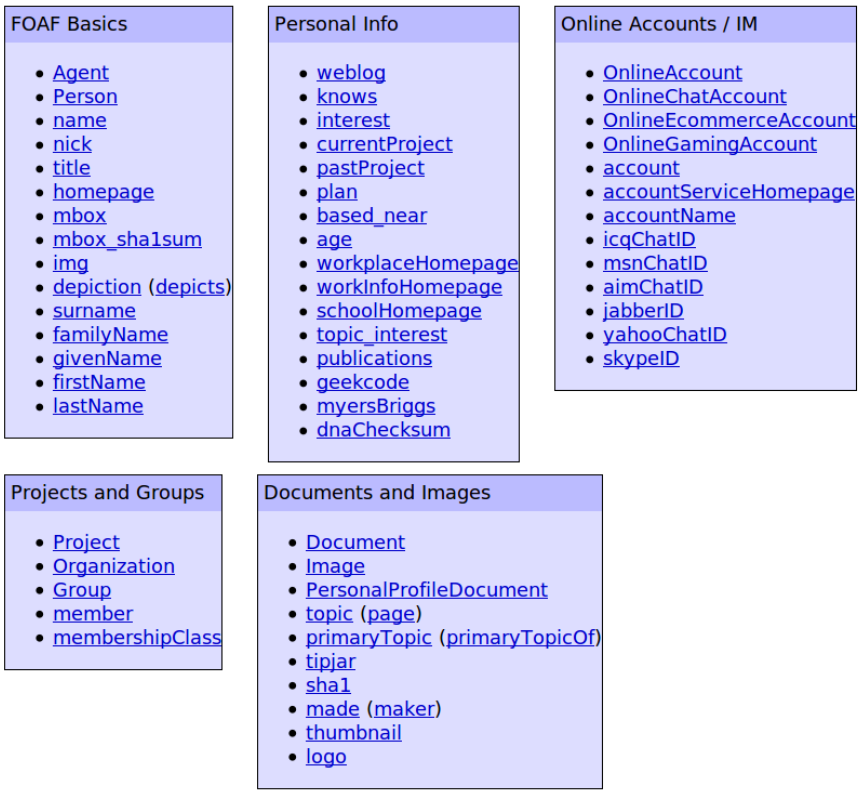


Figure 2: List of FOAF terms

An example of a FOAF statement, with the `knows` property included, can be:

```
<foaf:Person>
<foaf:name> Øystein Åsnes</foaf:name>
<foaf:mbox rdf:resource="mailto:Oystein.Asnes@sfj.no " />
...
<foaf:knows>
<foaf:Person>
<foaf:mbox rdf:resource="mailto: Elin.Østevik@sfj.no" />
<foaf:name> Elin Østevik  </foaf:name>
</foaf:Person>
</foaf:knows>
</foaf:Person>
```

The example above describes the Person with name "Øystein Åsnes" and email "Oystein.Asnes@sfj.no ", that knows the Person with name "Elin Østevik", referenced by his email, " Elin.Østevik@sfj.no ".

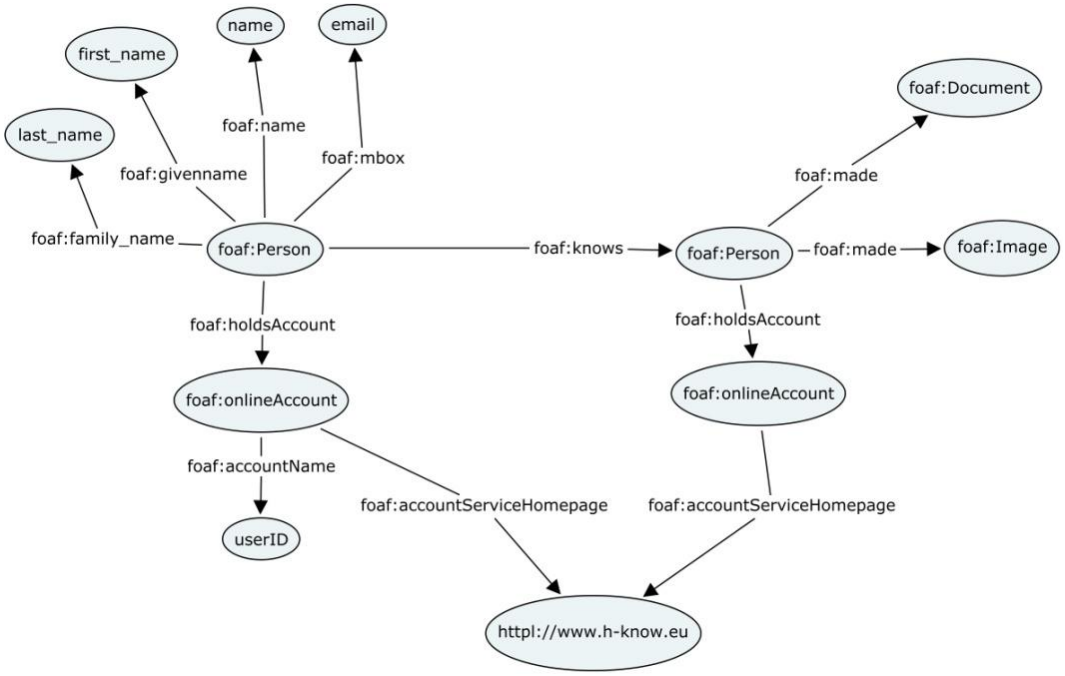


Figure 3: Example of FOAF mapping

The Figure 3 shows a general use of FOAF to describe users and their interactions in a social community. There are some concerns related to the use of FOAF, such as trust issues, since you can say you know whatever person without any verification and you can create whatever FOAF profile you want.

4.4 SIOC

The **SIOC** project[24] (Semantically-Interlinked Online Communities), is a meta-vocabulary for representing rich metadata from the Social Web in RDF/OWL, accepted by W3C. It aims to enable the integration of online community information (wikis, message boards, weblogs, etc).

SIOC aims to meet the needs of communities and users on the evolving Web, as community-centric content sites become more prevalent and finding relevant items from these communities is now more important than ever.

The figure 4 illustrates the characteristics of SIOC.

As SIOC can't incorporate on it everything that might be important to know about communities, about their users and about the contents that users create, otherwise it would be too large. Being built over RDF, we can take advantage of other specific description vocabularies, to complement the domain we want to specify. Being built in a modular design, we can create additional modules for specializing and further extending classes and properties contained within the SIOC core.

[24] http://sioc-project.org/

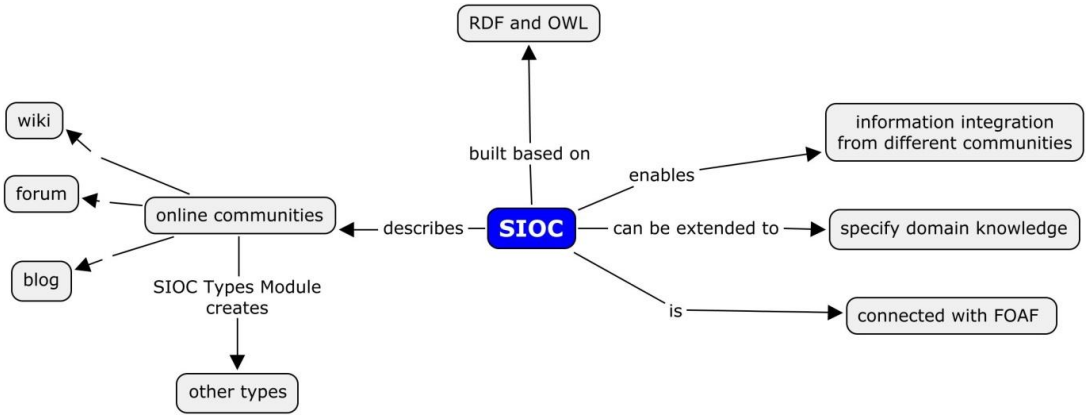


Figure 4: Overview of SIOC

To make the link between SIOC and specific domain vocabulary, SIOC Types module uses an **rdfs:seeAlso** property to point SIOC Types objects to the related vocabularies and classes.

In the figure 5 we have an overview of the classes that compose the SIOC and the relations between them. There are SIOC exporter tools that can be used to export RDF information about the contents and structure of Web 2.0 platforms (wikis, forums, blogs, message boards, etc). This allows information from every page of a site to be represented in RDF, making all the information contained there available in a machine readable form and so, ready for reuse. Some examples of those exporters are the Wordpress exporter or the vBulletin exporter.

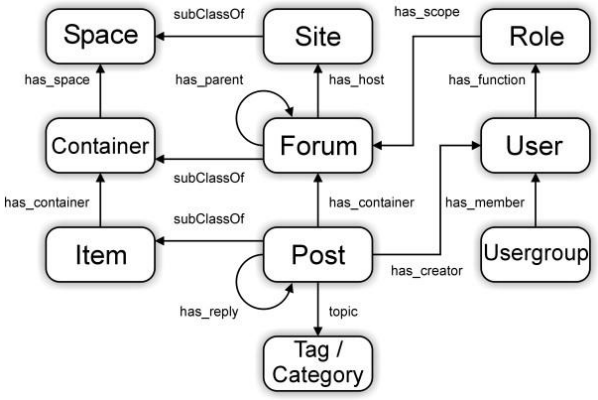


Figure 5: SIOC classes diagram

These classes allow structuring the information on online community sites and distinguishing between different kinds of objects. sioc:has creator and foaf:maker properties links all the user-created content to more information about its authors. One of the problems with combining social media data is in knowing which accounts users hold on different social media sites. SIOC attempts to solve this by re-using the FOAF (Friend of a Friend) vocabulary which can describe links between a person and accounts it holds in a distributed manner. By combining SIOC with FOAF data we can also re-use the information from personal FOAF profiles, e.g., the foaf:knows relationships.

4.5 Schema.org

Schema.org[25] provides a collection of shared vocabularies. Schema.org describes a variety of other item types, each of which has its own set of properties that can be used to describe the item. The broadest item type is *Thing*, which has four properties: name, description, `url`, and image. More specific types share properties with broader types. For example, a *Place* is a more specific type of Thing, and a *LocalBusiness* is a more specific type of Place. More specific items inherit the properties of their parent.

Following namespaces are used:

```
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix dct: <http://purl.org/dc/terms/>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix : <http://vocab.sindice.net/>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix xml: <http://www.w3.org/XML/1998/namespace>.
@prefix schema: <http://schema.org/>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix owl: <http://www.w3.org/2002/07/owl#>.
@base <http://schema.rdfs.org/all>.
```

One important item type is 'photograph'. Its properties are given in the following table.

Property	Expected Type	Description
description	Text	A short description of the item.
image	URL	URL of an image of the item.
name	Text	The name of the item.
url	URL	URL of the item.
about	Thing	The subject matter of the content.
accountablePerson	Person	Specifies the Person that is legally accountable for the CreativeWork.
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.
alternativeHeadline	Text	A secondary title of the CreativeWork.
associatedMedia	MediaObject	The media objects that encode this creative work. This property is a synonym for encodings.
audio	AudioObject	An embedded audio object.
author	Person orOrganization	The author of this content.

[25] http://schema.org/Photograph

Property	Expected Type	Description
comment	UserComments	Comments, typically from users, on this CreativeWork.
contentLocation	Place	The location of the content.
contentRating	Text	Official rating of a piece of content—for example,'MPAA PG-13'.
contributor	Person orOrganization	A secondary contributor to the CreativeWork.
copyrightHolder	Person orOrganization	The party holding the legal copyright to the CreativeWork.
copyrightYear	Number	The year during which the claimed copyright for the CreativeWork was first asserted.
creator	Person orOrganization	The creator/author of this CreativeWork or UserComments. This is the same as the Author property for CreativeWork.
dateCreated	Date	The date on which the CreativeWork was created.
dateModified	Date	The date on which the CreativeWork was most recently modified.
datePublished	Date	Date of first broadcast/publication.
discussionUrl	URL	A link to the page containing the comments of the CreativeWork.
editor	Person	Specifies the Person who edited the CreativeWork.
encodings	MediaObject	The media objects that encode this creative work
genre	Text	Genre of the creative work
headline	Text	Headline of the article
inLanguage	Text	The language of the content. please use one of the language codes from the IETF BCP 47 standard.

Property	Expected Type	Description
interactionCount	Text	A count of a specific user interactions with this item—for example, 20 UserLikes, 5 UserComments, or 300 UserDownloads. The user interaction type should be one of the sub types of UserInteraction.
keywords	Text	The keywords/tags used to describe this content.
mentions	Thing	Indicates that the CreativeWork contains a reference to, but is not necessarily about a concept.
provider	Person orOrganization	Specifies the Person or Organization that distributed the CreativeWork.
publisher	Organization	The publisher of the creative work.
reviews	Review	Review of the item.
sourceOrganization	Organization	The Organization on whose behalf the creator was working.
thumbnailUrl	URL	A thumbnail image relevant to the Thing.
version	Number	The version of the CreativeWork embodied by a specified resource.

Schema.org supports a vast collection of vocabularies spanning media and entertainment content to local business data. There are standards, such as RDFa, that are more expressive and extensible than schema.org – however their sheer complexity has led to slow adoption.

Interestingly, schema.org offers the ability to specify additional properties or sub-types to existing types. The schema.org has created an extension mechanism that lets webmasters and developers extend our existing schemas. When you extend schemas and use these extensions to mark up your data, search applications can at least partially understand your markup and use the data appropriately.

4.6 CONA

A new structured vocabulary, the Cultural Objects Name Authority[26] (CONA), contains titles, current location, and other core information for cultural works. CONA includes architecture and movable

[26] http://www.getty.edu/research/tools/vocabularies/cona/index.html

works such as paintings, sculptures, prints, drawings, manuscripts, photographs, ceramics, textiles, furniture, and archaeological artefacts. The CONA online search module is under development. It is scheduled to go live in early 2012.

4.7 *Bio*

Bio[27] is a vocabulary for describing biographical information about people, both living and dead. The BIO vocabulary contains terms useful for finding out more about people and their backgrounds and has some cross-over into genealogical information.

The URI for this vocabulary is,

http://purl.org/vocab/bio/0.1/Death

When abbreviating terms the suggested prefix is bio

Each class or property in the vocabulary has a URI constructed by appending a term name to the vocabulary URI. For example:

http://purl.org/vocab/bio/0.1/

Example: The death of Peder Balke can be described as,

```
:e a bio:Death
    ; dc:date "1887-02-05"
    ; bio:principal <http://dbpedia.org/page/Peder_Balke >
    ; bio:place <http://dbpedia.org/page/Helg%C3%B8ya,_Hedmark
>
    .
```

This vocabulary is useful in specifying biographical information (including date of birth, date of death) of an artist or photographer.

3. Guidelines

Choosing which vocabularies to use for photo is a key decision in any cultural heritage project. Typically, one needs more than a single vocabulary to cover the different relevant aspects of the photos. Many of the relevant vocabularies have been developed prior to the Semantic Web. We have seen some of them in this report. The major International Standard in this area, the Multimedia Content Description standard[28], widely known as MPEG-7, is defined using XML Schema.

SIOC reuses and extends existing vocabularies such as Dublin Core and FOAF in order to be compatible with RDF data modeled using other existing vocabularies. Thus, SIOC is often used in combination with the FOAF vocabulary for describing people and their friends, and the Simple Knowledge Organization System (SKOS) model for organising thesaurus-like data, SIOC lets developers link content items to other related items, to people, and to topics (using specific "tags" or hierarchical categories).

[27] http://vocab.org/bio/0.1/.html
[28] http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm

SIOC recommends the reuse of Dublin Core elements (`dc`) and refinements (`dcterms`) for describing properties such as creation date (`dcterms:created`), modification date (`dcterms:modified`), parts (`dcterms:hasPart` / `dcterms:isPartOf`), title (`dc:title`), and subject keywords (`dc:subject`), thereby deprecating proprietary properties in earlier versions of SIOC. Also, `sioc:avatar` can have resources of type `dctype:Image`.

Another relevant vocabulary is the *VRA Core*[29]. Where the Dublin Core (DC) specifies a small and commonly used vocabulary for on-line resources in general, VRA Core defines a similar set targeted especially at visual resources, specialising the DC elements. Dublin Core and VRA Core both refer to terms in their vocabularies as *elements*, and both use *qualifiers* to refine elements in similar way. All the elements of VRA Core have either direct mappings to comparable fields in Dublin Core or are defined as specializations of one or more DC elements. Furthermore, both vocabularies are defined in a way that abstracts from implementation issues and underlying serialization languages. A key difference, however, is that for Dublin Core, there exists a commonly accepted mapping to RDF, along with the associated schema.

It is also worth trying to use http://schema.org/Photograph (features are discussed in the Section 6.5). It will be a first attempt to use such vocabulary in practice.

References (*Basic Vocabulary namespaces*)

Prefix	XML Namespace	Specification
dc	http://purl.org/dc/elements/1.1/	The Dublin Core Metadata Element Set
dcterms	http://purl.org/dc/terms/	Other Dublin Core Elements and Element Refinements
foaf	http://xmlns.com/foaf/0.1/	Friend of a Friend (FOAF) Vocabulary
sioc	http://rdfs.org/sioc/ns#	SIOC Core Ontology
skos	http://www.w3.org/2004/02/skos/core#	SKOS Core Vocabulary

[29] http://www.vraweb.org/projects/vracore4/

Vedlegg 2: Eksempel på vokabular-bruk i prosjektet

Først eit skjermbilete frå Pubby-grensesnittet som viser data for fotografen Olav Reppen:

Olav (Olaf) Reppen at SfJ

http://data.sfj.no:8081/nb/person456af2352_1

Property	Value
?:PHT	▪ 2352
?:birthDate	▪ ?
?:deathDate	▪ 1921
?:description	▪ Født på gården Reppen i Sogndal, men som 12-åring flytte han til Austlandet.
?:familyName	▪ Reppen
?:gender	▪ M
?:givenName	▪ Olav (Olaf)
?:label	▪ Reppen, Olav (Olaf)
?:modified	▪ 2008-02-01T13:15:08 ▪ 2008-07-29T16:41:47
?:name	▪ Olav (Olaf) Reppen
?:ontology_v3.01.rdfcountryCode	▪ NO
?:source	▪ Kilder: Bonge, Susanne: Eldre norske fotografer, 1980 Sogn og Fjordane 10.03
?:spatial	▪ Hordaland, Bergen ▪ Sogn og Fjordane, Sogndal
?:type	▪ <http://abm.org/data/Photographer> ▪ <http://xmlns.com/foaf/0.1/Person>

Metadata

Anon_0

<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://purl.org/net/provenance/ns#DataItem>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://www.w3.org/2004/03/trix/rdfg-1/Graph>
<http://xmlns.com/foaf/0.1/primaryTopic>	<http://data.sfj.no:8081/nb/person456af2352_1>
<http://xmlns.com/foaf/0.1/topic>	Anon_0
<http://www.ontologydesignpatterns.org/cp/owl/informationrealization.owl#realizes>	<http://data.sfj.no:8081/data/nb/person456af2352_1>
<http://purl.org/net/provenance/ns#createdBy>	Anon_1 (more)

expand all

DB-felt	Mappa til	Eksempel	Ytterlegare info
Tittel	dcterms.title	?	Fotografen sitt namn (dcterms.title, abm:aboutPerson?)
Tittel	dcterms.title	?	Fotograf-verksemda sitt namn, same som over (sjå CorporateName)
Virksted	dcterms.spatial	Sogn og Fjordane Hordaland	Geografisk virkestad/-område
Register-ID	marcrel.PHT	2352	http://id.loc.gov/vocabulary/relators/pht info:lc/vocabulary/relators/pht
Biography	dcterms.decription	Født på gården Reppen i Sogndal....	Tekst om fotografen eller verksemda
Dates	schema.org/Person.birthDate	?	<Dates> må delast opp i fødselsdato og døds-dato
Dates	schema.org/Person.deathDate	1921	Sjå over
gender	foaf:gender	M	
PreferredName	rdfs:label	Reppen, Olav (Olaf)	
PersonName			?
clientFormatName	foaf:name		
corporateName			Feltet skil mellom personar og organisasjonar: 'N' = Personar, 'Y' = Organisasjonar
firstName	foaf:givenName	Olav (Olaf)	Førenamn
name	foaf:familyName	Reppen	Etternamn
nationalities	gn:countryCode	NO	
synonyms			?
relatedParties	dcterms:relation		?
source	dcterms:source	Kilder: Bonge, Susanne: Eldre norske fotografer..	
-	dcterms:modified	2008-02-01T13:15:08 2008-07-29T16:41:47	Lagt til <modified date> for framtidig bruk

Eksempel på mapping av data for fotograf Olav Reppen (http://data.sfj.no:8081/nb/person456af2352_1)