LOD och best practices vid modellering – vad är viktigt att tänka på?

Fagdag om modellering hos Riksantikvaren

2014-03-07

Hannes Ebner

hannes@metasolutions.se

meta solutions

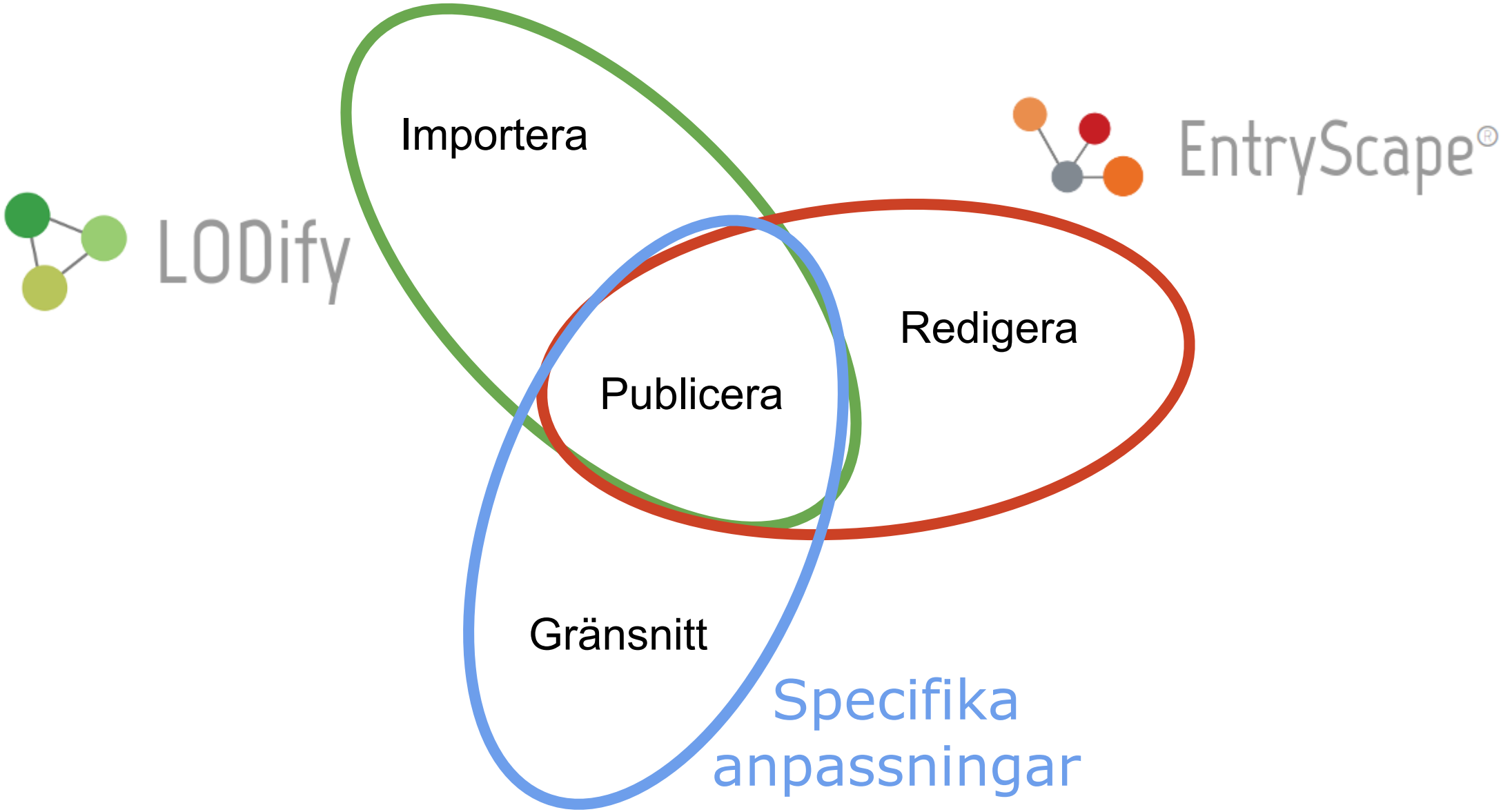
Lite om mig

Hannes Ebner

- Bakgrund: telekommunikation, informationssystem
- Doktorsavhandling med fokus på samarbetsmiljöer som är baserade på semantiska teknologier och länkade data

meta solutions

Om MetaSolutions



meta solutions

Innehåll

1. Stjärnor
2. Identifierare
3. Modellering
4. Datahantering
5. Övrigt

Länkade data-stjärnor

★ gör din information tillgänglig på Webben
(oberoende av format) under en öppen licens

★★ gör informationen tillgänglig som strukturerad data
(t. ex., Excel format istället för en bild av en tabell)

★★★ använd icke-proprietära format
(t. ex., CSV istället för Excel)

★★★★ använd URI:er för att identifiera ting,
och RDF för att uttrycka påståenden om dem

★★★★★ länka dina data till andras data, det ger sammanhang

Fördelar fyra stjärnor

Andra kan referera till dina data (URI:er)

Dataintegration förenklas
(Återanvändning av existerande utryck => andra har kännedom eller till och med utvecklat stöd för att förstå delar av datan)

Genomtänkta uttryck pga. "*stå på jättars axlar*"

RDF => existerande mjukvara och tjänster
(för att skapa, validera, lagra, maskinellt bearbeta, kombinera, redigera eller utforska datan med existerande frågespråk)

Fördelar fem stjärnor

Förtydliga dina data
(länka till väletablerade och väl uttänkta termer/begrepp istället för att skapa egna eller skriva fritext)

Omedelbar återanvändning av termer/begrepp
(inga tekniska aspekter av dataintegration som import, konvertering och drift/underhåll)

Möjliggör ökad dataspecialisering
(fokus på data som är unika för din organisation och förlita dig på information i andra datakällor)

Länkar till andra datakällor ökar förtroende och synlighet (jmf. referenser i artiklar som visar på att informationen är förankrad i ett större sammanhang)

meta solutions

Identifierare 1/3

Hierarkiska eller mönster-baserade

- Om det finns en given hierarki
  - `:collection/:item/:sub-collection/:item`
  - `http://example.org/collections/10/picture/2`
- Utan hierarki: `http://example.org/books/1`
- Bättre läsbarhet, förutsägbara

Naturliga

- Befintliga unika identifierare (t.ex. ISBN) -> URI
- `http://example.org/base/1234567890`

meta solutions

Identifierare 2/3

Slugs

- Mönster för att skapa URIer av t.ex. taggar
- `http://example.org/tags/linked-data`

Proxies

- När det saknas identifierare för externa resurser (3rd party)
  - ISO, IANA, ...
- Skapa en URI i sin egna namnrymd
  - `http://example.org/mime/application/pdf`

Identifierare 3/3

- Fragment URIer
  - Sista optionala delen i ett URI (#)
  - Identifierar vanligtvis (t.ex. HTML) en del av ett dokument (“fragment”)

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

- Identifierar konceptet “narrower” i SKOS Core
- “narrower” beskrivs (med andra koncept) i RDF filen `http://www.w3.org/2004/02/skos/core`
- Hela RDF filen måste laddas för att komma åt konceptet

Modellering - beteckningar

- Varje resurs bör ha en beteckning
  - rdfs:label
  - även rdfs:prefLabel om det finns flera

"Rule number 1 in building web software: never show the URI. If the URI does not have label, go and beat somebody up."

Sir Tim Berners-Lee på LDOW 2012

Modellering - multipla värden

- “Repeated properties”
- Samma tillvägagångssätt för flera språk

```
<http://example.org/bilder/semla1>
    dc:subject “grädde”@sv
    dc:subject “fløte”@no
    dc:subject “fløyte”@nn
    dc:subject “Sahne”@de
    dc:subject “Schlagobers”@de-at
    dc:subject “semla”@sv
    dc:subject “fastlagsbulle”@sv
    dc:subject “fastelavnsbolle”@dk
```

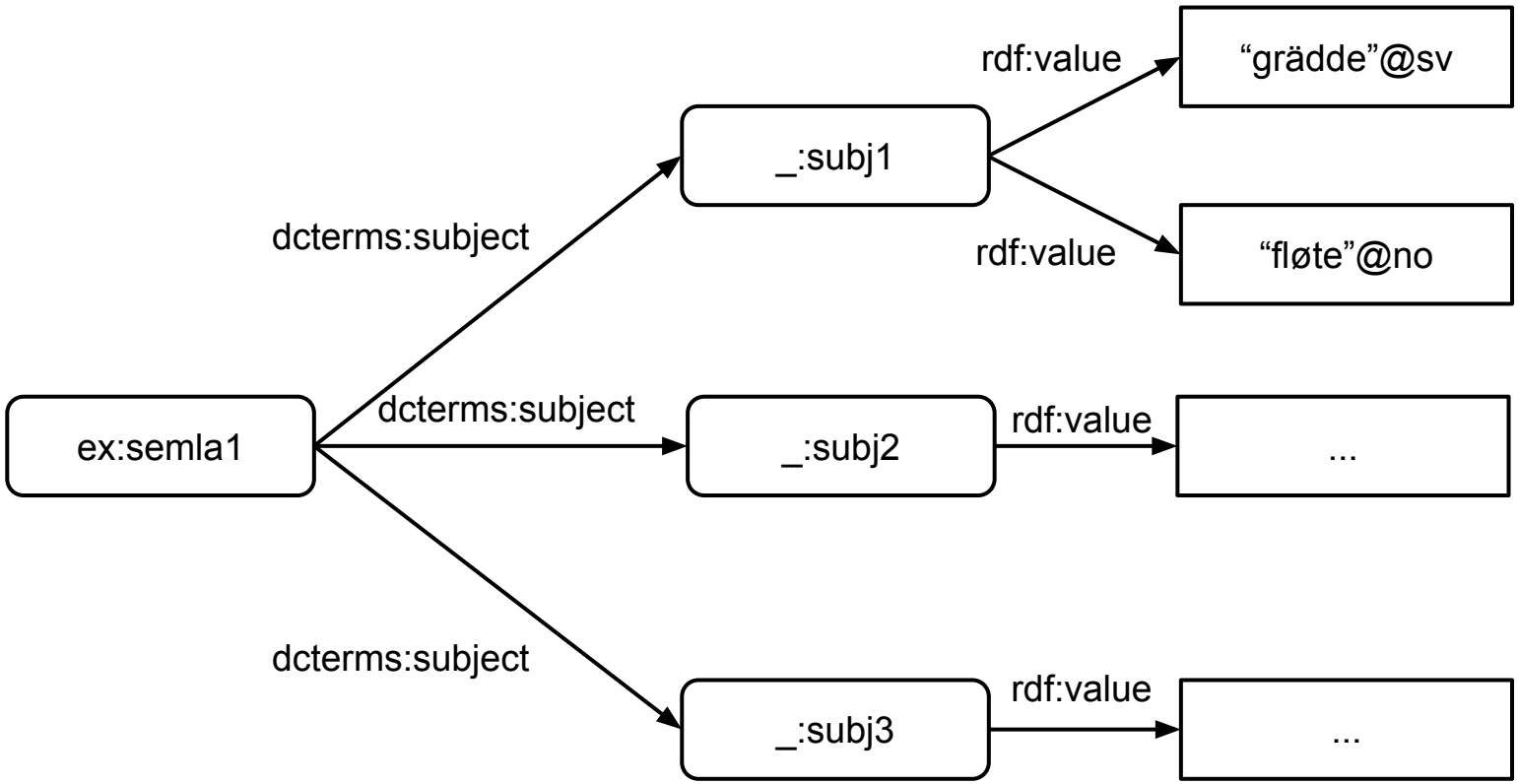
Blanka noder

- Resurs utan URI eller Literal: “anonymous resource”
- Representation av komplex data

```
<http://example.org/bilder/semla1>
    dcterms:subject _:subj1 ;
    dcterms:subject _:subj2 ;
    dcterms:subject _:subj3 .

_:subj1 rdf:value “grädde”@sv ;
    rdf:value “fløte”@no .

_:subj2 rdf:value “fastlagsbulle”@sv ;
    rdf:value “fastelavnsbolle”@dk .

_:subj3 rdf:value “någonting helt annat”^^ex:Encoding .
```

Syntax Encoding Scheme URI

Blanka noder



meta solutions

Modellering - indexresurser

- Repeated properties är inte ordnade
- Ordning behövs för indexering
- `rdf:List`
  - Om indexet är permanent
  - Innehållet är bekant vid publicering
- Annars: `rdf:Seq`
- Alternativt, utan indexresurs
  - naturlig ordning
  - t.ex. sortera inlägg efter publiceringsdatum

meta solutions

Datahantering - named graphs

- “Quadruple”
- Identifierare för en mängd tripplar
- Standardiserade sedan RDF 1.1

Exempel i TriG:

```
<http://example.org/graph/1> {
    <http://example.org/document/5> rdfs:label "Example"
;
        dct:format “application/pdf” ;
        dct:extent 1200000 .
}
```

meta solutions

Datahantering - grafannotering

- Information om named graphs
  - “Provenance”
  - Vem får läsa/skriva (ACL)
  - ...
- Relationer mellan named graphs

Exempel

```
<http://example.org/graphs/1> {
    <http://example.org/graph/1> dct:created "2014-03-02"^^xsd:date.
}
```

Vokabulärer

Vanliga vokabulärer

- DC, DC Terms
- FOAF
- Schema.org
- ...

Koncept-, vokabulär- och ontologispråk

- SKOS
- RDFS
- OWL

meta solutions

Koncept-, vokabulär- och ontologispråk SKOS, RDFS och OWL

SKOS - *"Simple Knowledge Organisation System"*

- Concepts in ConceptSchemes
- Namn via prefLabel, altLabel
- Hierarkier via narrow/broader
- relationer via related

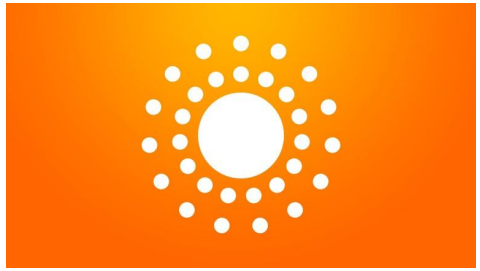
RDFS - *"RDF Vocabulary Description Language"*

- Definera klasser och properties i RDF
- subClassOf och subPropertyOf för att förfina

OWL - *"Web Ontology Language"*

- Kraftfullare än RDFS

DCMI Terms - Qualified Dublin Core



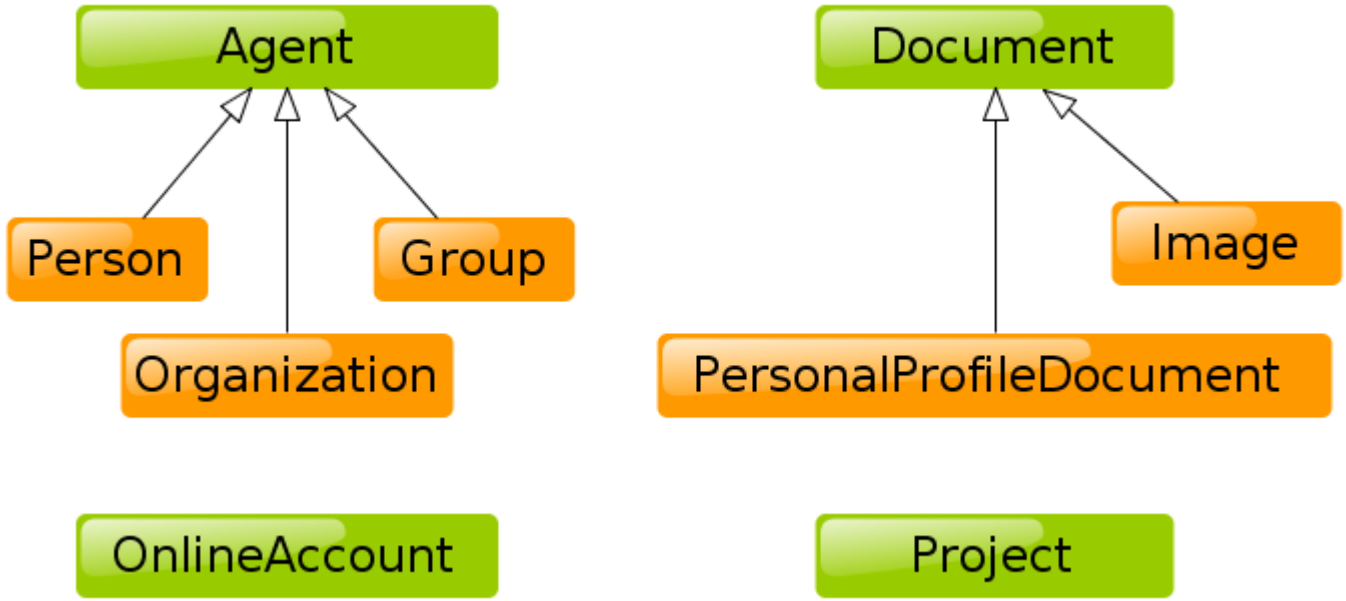
Properties:

abstract, accessRights, accrualMethod, accrualPeriodicity, accrualPolicy, alternative, audience, available, bibliographicCitation, conformsTo, **contributor**, **coverage**, created, **creator**, **date**, dateAccepted, dateCopyrighted, dateSubmitted, **description**, educationLevel, extent, **format**, hasFormat, hasPart, hasVersion, **identifier**, instructionalMethod, isFormatOf, isPartOf, isReferencedBy, isReplacedBy, isRequiredBy, issued, isVersionOf, **language**, license, mediator, medium, modified, provenance, **publisher**, references, **relation**, replaces, requires, **rights**, rightsHolder, **source**, spatial, **subject**, tableOfContents, temporal, **title**, **type**, valid

- Intoducerades 1995 på Workshop i Dublin Ohio av olika biblioteksorganisationer
- Fokus: beskriva resurser/verk av olika slag
- Underhålls av DCMI (Dublin Core Metadata Initiative)

meta solutions

FOAF - Friend Of A Friend



Properties: account | age | based_near | birthday | currentProject | depicts | dnaChecksum | gender | givenName | holdsAccount | img | interest | knows | lastName | mbox | member | nick | openid | page | phone | plan | status | surname | thumbnail | title | topic | weblog

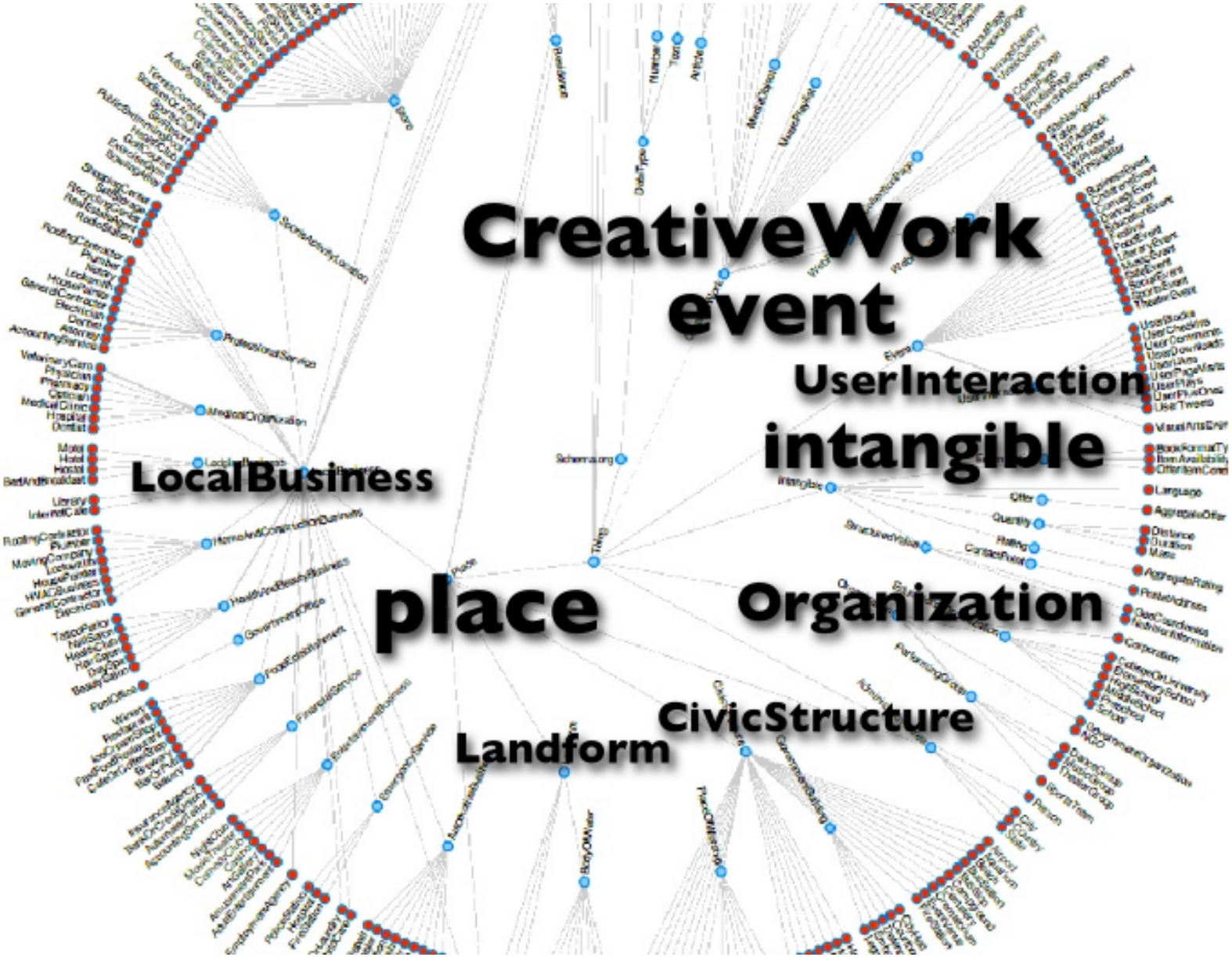
Specifikation: http://xmlns.com/foaf/spec

Introducerades 2000, senast uppdaterad 2010

Bygger på Dublin Core

meta solutions

Schema.org



Bing,
Google
och Yahoo

Introducerade schema.org 2011
Mer än 800 typer och 600 egenskaper
Används för att förbättra sökresultat

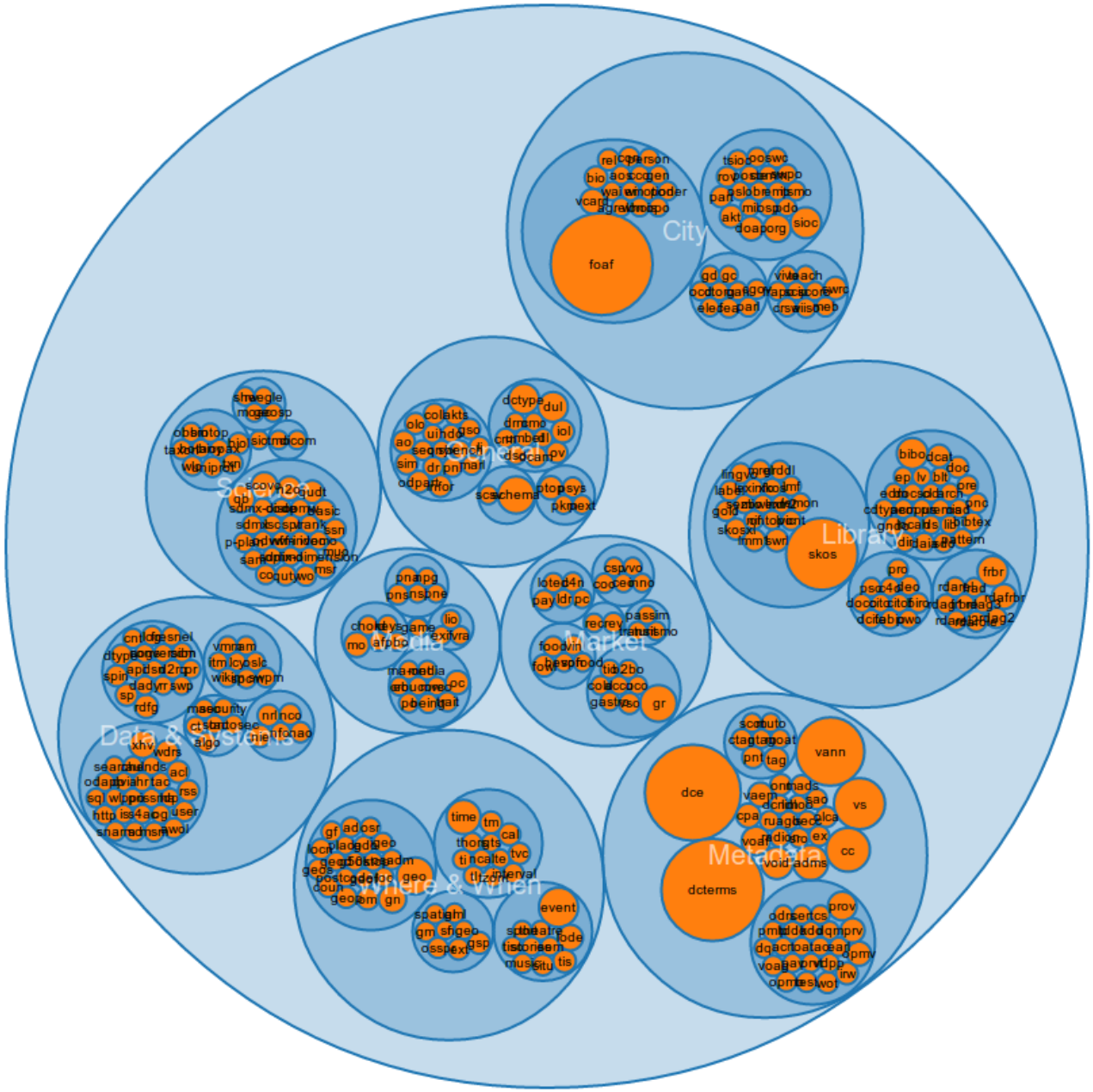
meta solutions

Linked Open Vocabularies

http://lov.okfn.org

Vokabulär

- RDFS
- OWL



Förstådd och rätt använd

Hitta/anpassa den **bästa** standarden

- Hur avgränsa ett område
- Hur komma överens, legitimitet

Använd **många** vokabulärer och Länkade data

- Kombinera existerande vokabulärer + egna
- Best practise växer fram, dubblera där så saknas

Interoperabilitet vs. Harmonisering

En enskild standard ger interoperabilitet

- Maskiner kan utbyta data efter noggrann programmering
- Oftast punkt till punkt

Länkade data ger harmonisering mellan standarder/vokabulärer

- Olika data kan blandas och samexistera
- Maskiner förstår de delar de programmerats för
- Ibland genom att förgrova och dra slutsatser enligt förberedda regler

Dokumentation av vokabulärer

Även återanvändning av existerande vokabulärer

- OWL property restrictions
- Återdefiniera mha RDFS eller OWL
- Description Set Profiles (DCMI)
- Shape expressions (W3C, standardisering pågår)
- Beskriv i text på ett strukturerat sätt
  - välj denna!!!

meta solutions

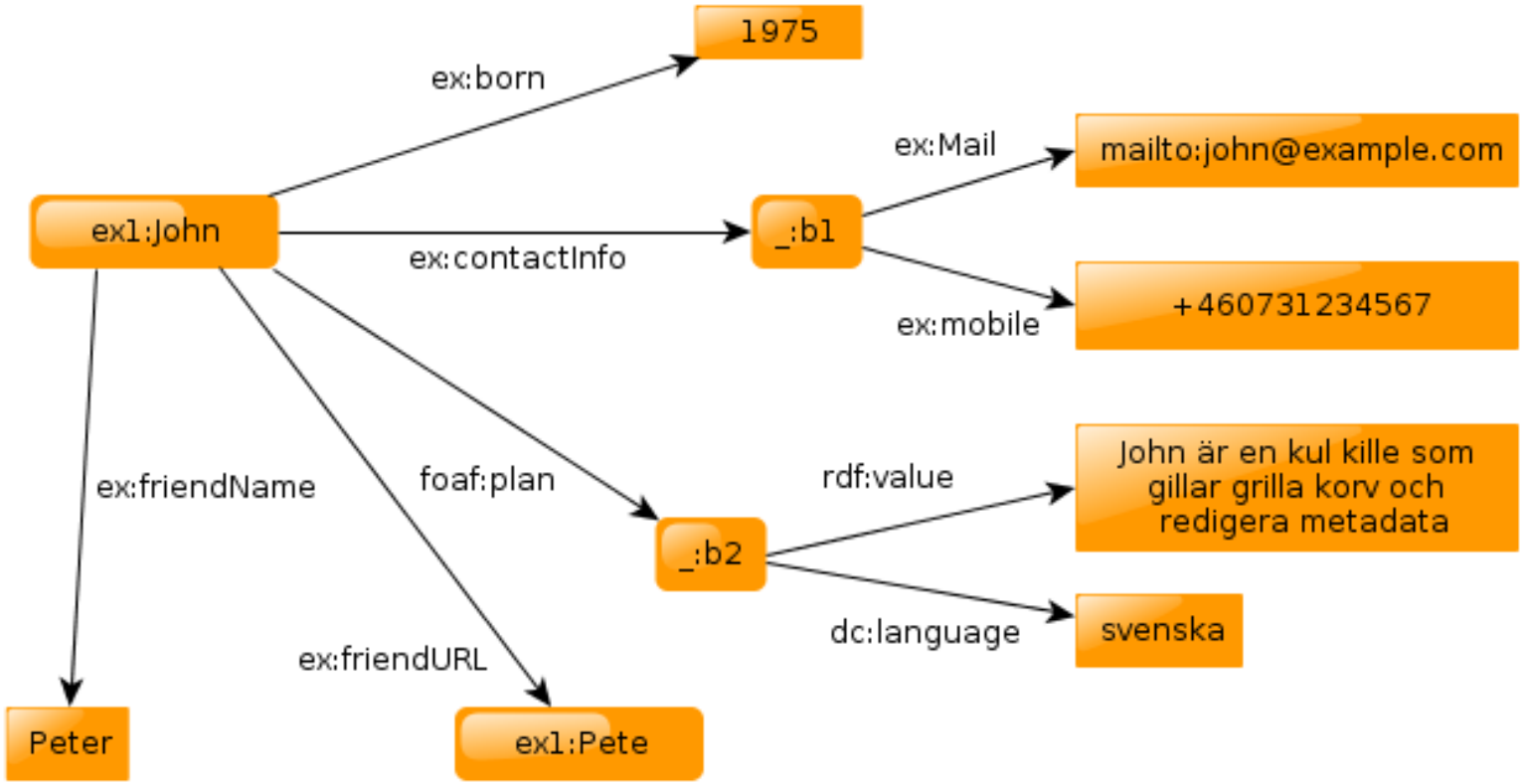
“Follow your nose”

- Hitta ytterligare relevant information på webben
- Det ska gå att följa länkar
  - URIer måste vara “dereferencable”

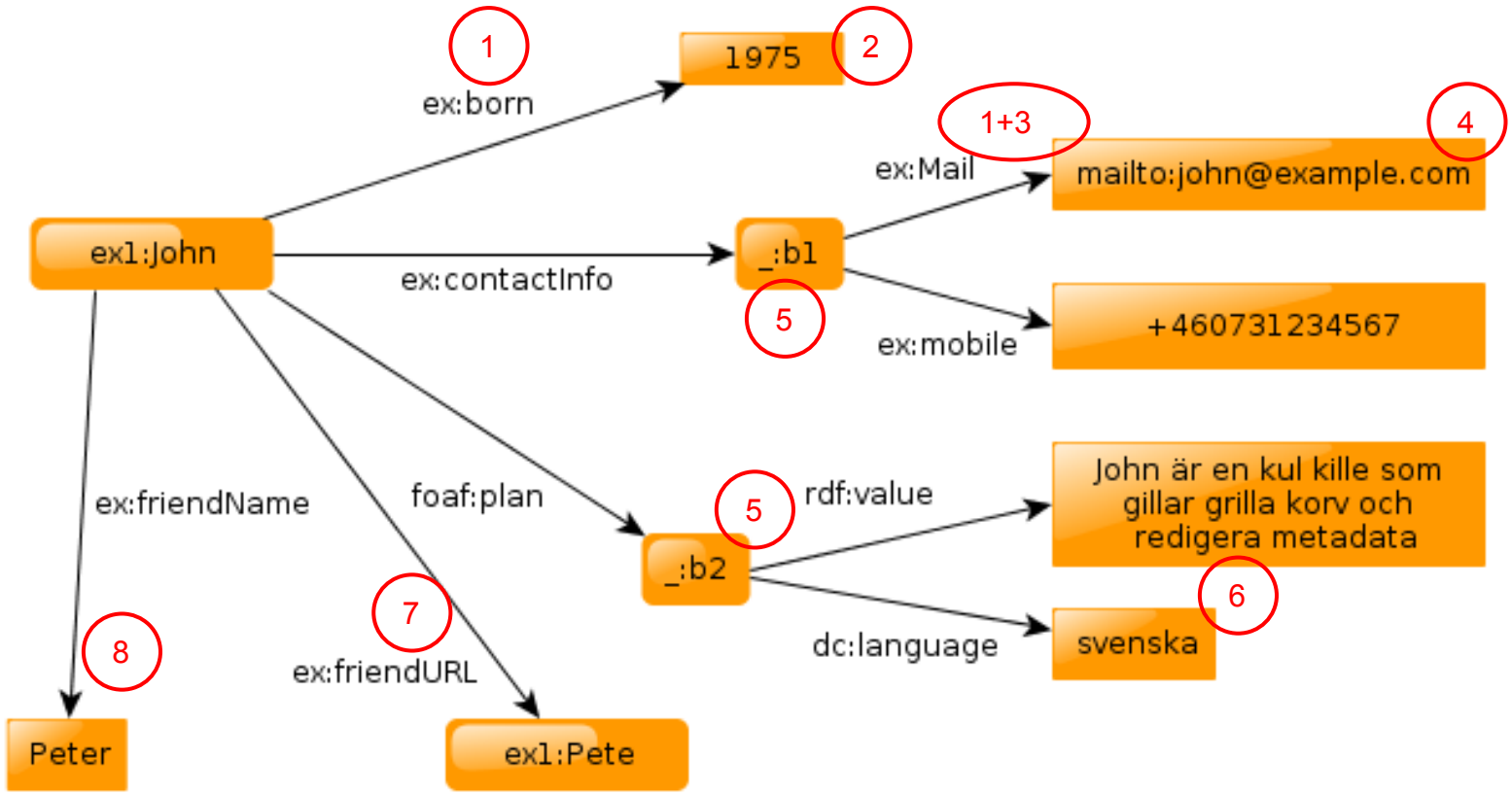
Länkade data som de ska vara!

meta solutions

Finn åtta fel (dålig RDF design)



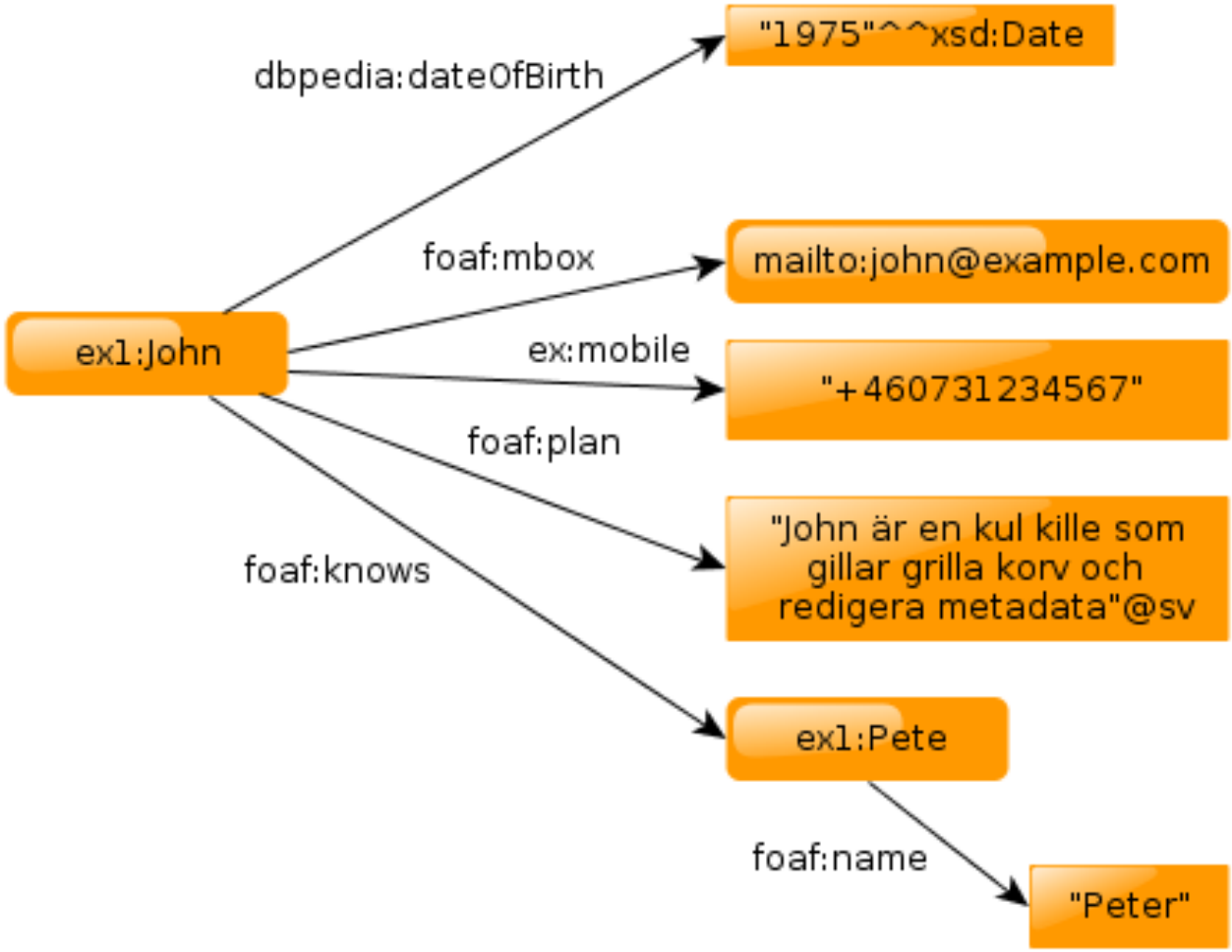
Finn åtta fel (dålig RDF design)



1. Återanvänd properties
2. Datatyp saknas
3. Lowercase på properties
4. Använd resurser för URIer
5. Artificiell gruppering
6. Ange språk på literal
7. Du pekar på resurser, inte URL:er, tänk konceptuellt
8. Properties ska utgå från rätt resurs

meta solutions

Bra RDF design



Frågor?

Hannes Ebner

hannes@metasolutions.se

Kontakta mig gärna, t.ex. om ni:

- har frågor om länkade data
- behöver en plattform för länkade data

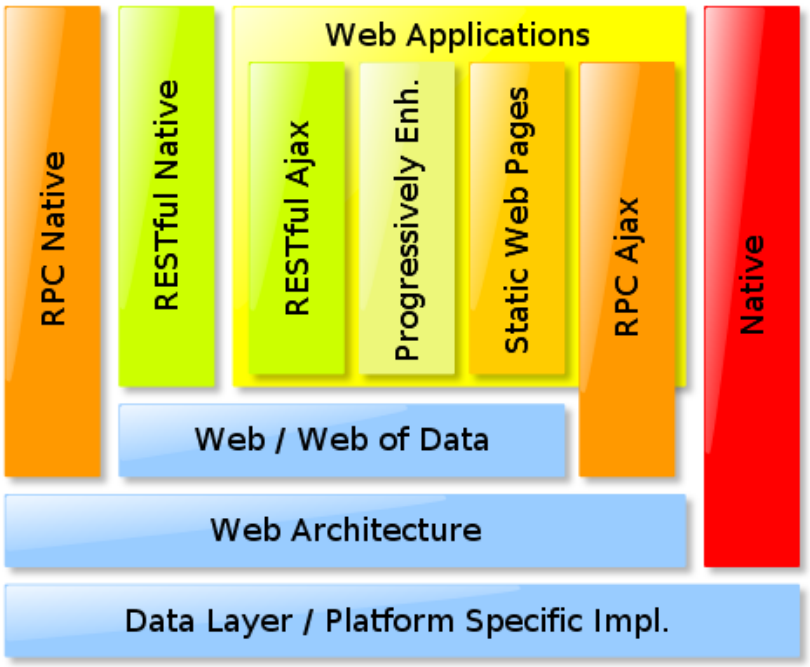
MetaSolutions AB

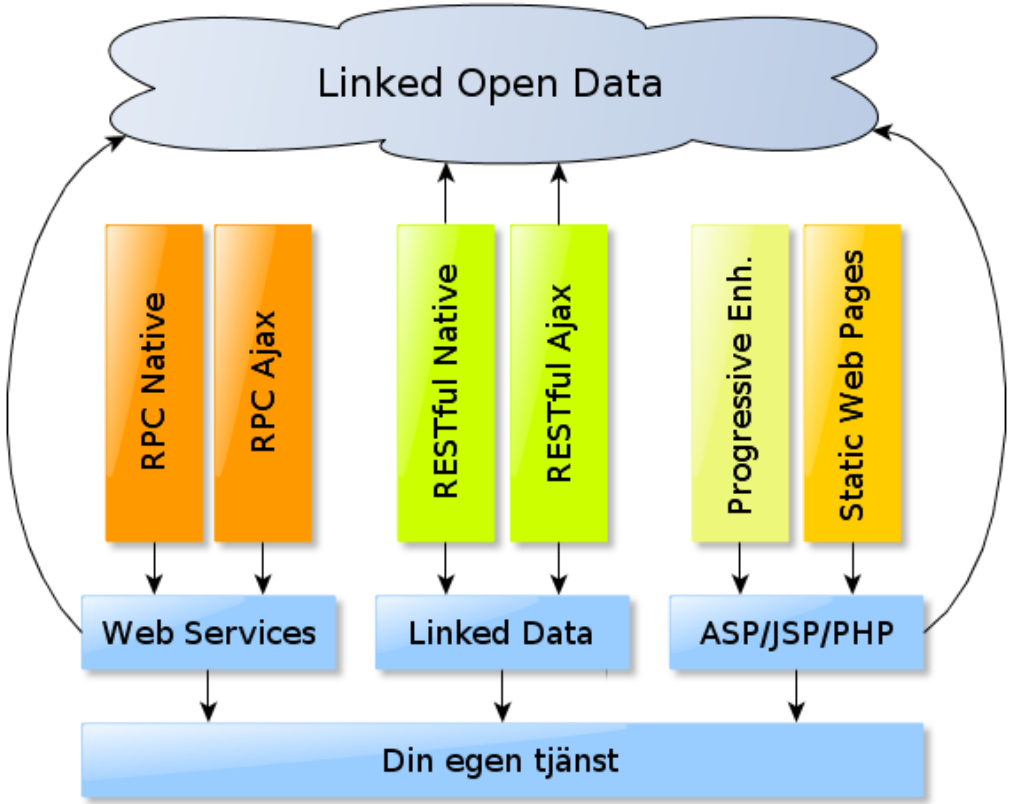
www.metasolutions.se

meta solutions

1. Länkade data Proxy / Cache

- Komma runt browserbegränsningar
- Erbjuda stabil/snabb åtkomst
- Partitionera så delar av grafen kan laddas asynkront
- Vissa delar kan cachas on-demand
- Enhetligt format, tex JSON-LD
för minimalt footprint



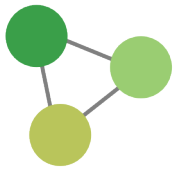


meta solutions

2. Ladda in några centrala dataset

- Alla datakällor finns inte som LOD
- Ha koll på stabilitet i åtkomst
- Vid behov harmonisera identifierare mellan dataset
- Vid behov skapa aggregerad vy (tex mha EDM)
- Använd named graphs för att hålla koll på ursprung

Man kan bygga ovanpå existerande lösningar, tex



lodify.com

meta solutions

3. Länkade data virtuell vy av relevanta ting (*)

- Lista av rekommenderade ting (givet ett ting)
- Resultat som ett RDF uttryck (återstår att tänka ut)
- Troligtvis använder man LDP containers
- Genereras via en SPARQL construct fråga
- Frågor kan skapas allteftersom
- Ingen nyutveckling krävs för nya frågor

(*) Överkurs