

# Länkade öppna data i Sverige – Portal och nationell statistik

## Linked open data in Sweden – Portal and national statistics

### Sammanfattning

Denna ansökan syftar till att främja den form av öppna data som kallas *länkade öppna data* genom två komplementära målsättningar: (1) att etablera en central datakälla i Sverige vilken kan agera som katalysator för andra initiativ kring länkade data i Sverige, och (2) att etablera en svensk portal om länkade öppna data med information om vilka länkade data som finns i Sverige i kombination med lättillgänglig information på svenska om vad länkade öppna data är. Datakällan som avses i (1) är SCBs (Statistiska centralbyrån) klassificeringsdatabas som innehåller en stor mängd klassificeringar som används i den statistik som publiceras.

Det finns flera möjliga sätt att publicera öppna data, exempelvis som Excel-filer, CVS-filer eller som ett API (programmeringsgränssnitt). Representationsformat och frågespråk från den semantiska webben (exempelvis RDF, OWL för modellering av domänkunskap och SPARQL för att hämta ut information) har stora fördelar gällande återanvändning och kombination av data<sup>1</sup>. För att illustrera hur dessa alternativ förhåller sig till varandra finns följande modell<sup>2</sup>:

- ★ Data tillgänglig på webben (oavsett format), med en öppen licens.
- ★★ Tillgänglig som maskinläsbar strukturerad data (t ex Excel i stället för scannad tabell).
- ★★★ Alla ovan, samt ickeproprietärt format (t ex CSV i stället för Excel).
- ★★★★ Alla ovan, samt att använda W3C-standarder (RDF, SPARQL) för beskriva saker.
- ★★★★★ Alla ovan, samt att länka ens data till andras data.

Detta projekt syftar till att svenska öppna data i större utsträckning ska publiceras på 5-stjärnenivån, ofta kallade *länkade data*<sup>2</sup>. Genom att följa principerna för länkade data kommer vi att kunna länka viktiga svenska datamängder till existerande datamängder i det så kallade *Linked Open Data cloud*<sup>3</sup> (LOD). Inom LOD kan noteras att det finns ett antal datakällor (DBpedia<sup>4</sup>, Freebase,<sup>5</sup> GeoNames<sup>6</sup> m.fl.) som länkas till oftare än andra. Dessa datakällor ger ett stort mervärde då de medverkar till att vitt skilda datakällor som inte refererar till varandra explicit ändå är indirekt relaterade genom att de refererar till samma centrala datakällor. På samma sätt ser vi att SCBs klassificeringsdatabas har potential att bli en sådan knutpunkt för svenska länkade data. Notera att de länkar som etableras till en sådan svensk länkad data-knutpunkt är att betrakta som komplement snarare än som ersättning för de länkar som kan göras till internationella datakällor, vilka fortfarande är oerhört värdefulla.

Genom att även tillhandahålla en *portal för länkade öppna data i Sverige*, inklusive instruktioner, "tutorials" och länkar till relaterade verktyg, kommer vi att skapa en ökad kompetens kring publicering och användning av länkade data i Sverige. Till den samlade erfarenhet som finns tillgänglig utomlands kommer vi att kunna lägga våra egna erfarenheter

<sup>1</sup> <http://eprints.soton.ac.uk/271285/1/bizer-heath-berners-lee-ijswis-linked-data.pdf>

<sup>2</sup> <http://www.w3.org/DesignIssues/LinkedData.html>.

<sup>3</sup> <http://richard.cyganiak.de/2007/10/lod/>

<sup>4</sup> <http://dbpedia.org/>

<sup>5</sup> <http://www.freebase.com/>

<sup>6</sup> <http://www.geonames.org/>

från detta projekt, som kan fungera som ett pilotprojekt för svenska länkade data. Vi föreslår även att etablera en praxis kring tillhandahållande av länkade data som har anknytning till Sverige, varvid dessa då registreras på ett sätt som gör det enklare för andra att återfinna data. Vi kommer att förlita oss på [thedatahub.org](http://thedatahub.org) för att på så sätt automatiskt integreras i det internationella länkade öppna data molnet och ge bättre synlighet för svenska initiativ.

## **Syfte och mål**

Ett av huvudsyftena med projektet är att demonstrera värdet av att publicera länkade data. Dessa värden innefattar (1) ökad datainteroperabilitet, (2) data som är förberett för uppgifter och användningsområden ej påtänkta vid publiceringen och (3) format som möjliggör automation av dataåtkomst och användning. Vi anser att denna nya paradigm, d.v.s. *länkade data*, har potential att revolutionera vårt sätt att se på öppna data och hur öppna data kommer att publiceras i Sverige i framtiden. Sveriges tradition av offentlighet ger oss en unik position i världen när det gäller öppenhet, vilket utgör en perfekt bas för att bli en världsledande nation inom *länkade öppna data*.

Den första av projektets målsättningar är att etablera en central datakälla i Sverige som kan agera katalysator för andra initiativ kring länkade data i Sverige, och en andra målsättning är att etablera en svensk portal för länkade öppna data där information om vilka länkade data som finns i Sverige kan samlas, i kombination med lättillgänglig information om hur man publicerar sådana data. Projektet tillgängliggör offentliga öppna data (*tema A* i utlysningen) genom att tillgängliggöra SCBs klassifikationsdatabas som länkade data med stabila identifierare i form av URI:er. Projektet innefattar även insatser som möjliggör omfattande tillgängliggörande av datakällor (*tema D*) genom att skapa en nationell portal med information kring länkade data, för att stödja svenska aktörers publicering av länkade datakällor både i Sverige och internationellt.

## **Nytta och skalbarhet**

Att publicera data enligt principerna för länkade data ger en markant ökning av värdet av data från enskilda datakällor, genom att möjliggöra automatisk identifiering och åtkomst, länkar mellan olika datamängder, möjlighet till automatisk tolkning av datamängders betydelse (genom ontologier), samt interoperabilitet med datamängder från hela världen. Datakällan som vi fokuserar på att göra till länkade data, SCBs klassifikationsdatabas, innehåller en stor mängd spännande fakta om Sverige, exempelvis vilka tätorter, kommuner, och myndigheter som finns, såväl som olika klassificeringar av yrken, produkter, branscher osv. som ofta används i den offentliga debatten. Då en del av det som finns i klassifikationsdatabasen har sitt ursprung i internationella standarder kommer projektet att knyta an till bredare initiativ, särskilt om de redan är exponerade som länkade data. För de delar som är specifika för Sverige och underhålls av redaktörer på SCB är det även intressant att undersöka möjligheten att länka till andra svenska initiativ som t.ex. Svenska Ämnesord<sup>7</sup> som underhålls av Kungliga Biblioteket. De länkade data som produceras kommer att driftsättas av SCB med identifierare som indikerar ursprunget från SCB tydligt. Detta, tillsammans med information om

---

<sup>7</sup> <http://www.kb.se/katalogisering/Svenska-amnesord/>

datastabilitet, kommer att skapa en god grund för andra utvecklare i Sverige att skapa mervärde genom att länka till dessa data. T.ex. kommer en sådan enkel sak som en uppdaterad maskinläsbar lista på vilka myndigheter, kommuner och landsting<sup>8</sup> som finns i Sverige kunna leda till betydande förenklingar och förbättrad kvalitet i andra system, som därigenom snabbt kan påvisa mervärdet av länkade öppna data. I ett längre perspektiv är det viktigt att klassificeringsdatabasen har god tillgänglighet och svarstid om andra system ska kunna förlita sig på denna information. Då länkade data är en förlängning av den vanliga webben ges goda möjligheter till standardiserade lösningar för skalbarhet, t.ex. genom en webbcache.

Portalen för svenska länkade data kommer att tillgängliggöras via Linköpings Universitet, medan datakällorna i sig länkas och beskrivs i en nyskapad grupp på thedatahub.org. Portalen kommer att vara en wiki, för att på så sätt minimera det underhåll som krävs genom en enskild administratör och istället göra portalen till en levande mötesplats genom bidrag från dess användare, på så sätt kan vi garantera ett långsiktigt tillhandahållande av beskrivningarna. Användandet av thedatahub.org medför att datakällorna beskrivs på ett standardiserat sätt, m.h.a. VOID ontologin<sup>9</sup>, och lätt kan återfinnas av användare och utvecklare världen över.

Det är värt att notera att de projektsökande anordnade den första öppna träffen i Sverige om länkade data på KTH den 17 april 2012<sup>10</sup>, vilket gav en uppslutning på cirka 30 personer med många fler som ville sluta upp men inte hade möjlighet (mötet sändes även via Bambuser för att möjliggöra deltagande på distans). Under träffen diskuterades möjligheten att sätta upp en nationell portal med fokus på länkade data i Sverige. Samstämmighet rådde om att de projektsökande skulle initiera processen och därefter bjuda in fler att delta. Vi håller för troligt att detta innebär både att initiativet kommer att mottas positivt av många intressenter i Sverige och att en community snabbt kommer bildas för att hålla portalen uppdaterad i ett längre perspektiv. Portalen kommer att stödja aktörer intresserad av länkade data i Sverige, inklusive såväl utvecklare (på myndigheter och företag), akademiker såväl som allmänintresserade.

### **Genomförande och parter**

Projektet pågår från och med 2012-08-20 till och med 2012-12-21 (4 månader). Tabell 1 beskriver de aktiviteter som ska utföras, ansvariga och involverade parter, tidsåtgång och projektmånad när aktiviteten slutförs. Medverkande parter är Linköpings universitet (LiU), Malmö högskola (Mah), Statistiska Centralbyrån (SCB) samt MetaSolutions AB (MS) – detaljer i Tab. 2. Projektets totala budget är på 399.800kr, utöver lönekostnader för respektive part (timmar enligt Tab. 1) och relaterade indirekta kostnader innehåller projektets budget resekostnader om 13.800kr för ett projektmöte och besök vid SCB i Örebro (aktivitet 1-2).

SCBs främsta uppgifter är att förse kunder med statistikunderlag för beslutsfattande, forskning m.m. Detta görs i huvudsak på uppdrag av regeringen och olika myndigheter, men även för privata kunder. Delar av den insamlade statistiken publiceras offentligt på SCBs

---

<sup>8</sup> Detta var en av de saker som föreslogs på träffen om länkade data i Sverige den 17 april 2012.

<sup>9</sup> <http://vocab.deri.ie/void>

<sup>10</sup> <http://ldsv.eventbrite.com/>

webb och i tryckta publikationer, medan annat material kan fås not betalning. Statistik och metadata presenteras och struktureras med hjälp av de klassifikationer som SCB tillhandahåller. **MetaSolutions AB** är ett företag avknoppat från forskning vid Kungliga Tekniska Högskolan (KTH) i Stockholm. Företaget fokuserar på lösningar för avancerad informationshantering, som t.ex. att sammanföra semantik med storskaliga tillämpningar på webben. MetaSolutions jobbar med att exploatera resultaten från ett antal framgångsrika forskningsprojekt och använder i huvudsak öppen källkod och länkade (öppna) data vilket gynnar både industri och akademi.

**Tabell 1:** Aktivitetsplan med involverade parter (ansvarig i fetstil), timmar samt milstolpar.

Aktivitet	Parter	Arbetsinsats i timmar	Milestone (måndad)
1) Hitta eller designa vokabulärer som ska användas för SCB	<b>MS + SCB</b>	40+40	M1
2) Anpassa SCBs webbplattform så att den producerar länkade data mha vokabulärerna	<b>SCB + MS</b>	100+60	M4
3) Dokumentera och lägga upp på thedatahub.org	<b>LiU</b>	8	M4
4) Sätta upp portalen, inkl. struktur	<b>LiU</b>	40	M2
5) Sätta upp grupp på thedatahub.org	<b>Mah</b>	10	M1
6) Producera/översätta tutorials och riktlinjer	<b>Mah + LiU + MS</b>	60+50+20	M3
7) Samla länkar och goda exempel	<b>LiU + Mah</b>	20+20	M3
8) Marknadsföra portalen	<b>LiU + MS</b>	8+8	M4
9) Projektledning	<b>LiU</b>	20	-

**Tabell 2:** Medverkande parter

Part	Org-nr	Arbetsställe
Linköpings Universitet (LiU)	202100-3096	Institutionen för Datavetenskap
Malmö högskola (Mah)	202100-4920	Institutionen för Datavetenskap
Statistiska centralbyrån (SCB)	202100-0837	SCB, Örebro
MetaSolutions AB (MS)	556854-0701	MetaSolutions, Saltsjöbaden

Konkreta leverabler inkluderar länkade öppna data (SCBs klassificeringsdatabas), en wikiportal för svenska länkade data, samt material om länkade data på svenska. Portalen kommer att tillgängliggöras från Institutionen för Datavetenskap, LiU (en existerande webbserver kommer användas, varvid denna inte genererar någon kostnad i projektet). Institutionen tar ansvar för att tillhandahålla portalen även efter projektet. Materialet i portalen ska vara på svenska och innehålla följande: bakgrundsmaterial om länkade data och format, instruktioner, hjälpsidor och "tutorials" för publicering, exempel på svenska länkade data och applikationer, länksamlingar med pågående och tidigare projekt, länkar till forum, länk till en grupp på thedatahub.org för svenska datakällor, "önskelista" för datakällor som användare skulle vilja se som länkade data.