

SVEP: DP1

Arbetsrapport 1

Interoperabilitetsfrågor i svenskt perspektiv

Stefan Andersson,

Enheten för digital publicering,
Uppsala universitetsbibliotek

April 2004

Interoperabilitetsfrågor i svenskt perspektiv	3
Sammanfattning	3
Bakgrund	3
Övergripande mål för SVEP	4
De övergripande målen i förhållande till DP1	4
Maximal synlighet.....	5
Långsiktig hållbarhet.....	6
Långsiktig tillgänglighet	7
Tekniska lösningar för elektronisk publicering vid svenska högskolor.....	7
Olika metadataformat som används.....	8
Olika beskrivningsnivåer	9
Kontrollerade vokabulärer	10
Gemensam ämneskategorilista.....	10
Kontrollerade termer för dokumenttyper	11
Kontrollerade namnformer.....	11
Bibliografiska databaser/publikationsdatabaser.....	12
Nästa rapportering	12
Bilagor.....	13

Interoperabilitetsfrågor i svenskt perspektiv

Sammanfattning

Denna arbetsrapport är den första från delprojekt 1, *Interoperabilitet: harmonisering av beskrivningar och beskrivningsformat för fulltextpublicerade dokument*, i SVEP (Svenska högskolans elektroniska publicering).¹ Delprojektet skall leda till en harmonisering av metadatabeskrivningar för elektroniskt publicerade dokument vid svenska universitet och högskolor och ett antal rekommendationer som främjar utbyte av information om dessa dokument och som skapar förutsättningar för nya tjänster.

Rapporten ger en bakgrund till delprojektet, en nulägesbeskrivning samt anger inriktningen på kommande arbete.

Bakgrund

Det BIBSAM²-stödda pilotprojekt *En gemensam portal för akademisk publicering* vid svenska universitet och högskolor³ hade som ett av målen att undersöka möjligheterna att bygga upp en portal för akademisk publicering med hjälp av metadataöverföring genom OAI-PMH (The Open Archives Initiative Protocol for Metadata Harvesting).⁴ Projektet genomfördes på Enheten för digital publicering⁵ vid Uppsala universitetsbibliotek och avrapporterades vid ett seminarium i Uppsala den 5 mars 2003. Där påvisades flera samordningsbehov när det gäller både metadata och tekniska lösningar. Det konstaterades att frågor kring metadata och interoperabilitet kommer att bli allt viktigare i takt med att antalet fulltextpublicerade svenska vetenskapliga publikationer som avhandlingar och forskningsrapporter, men även examensarbeten, snabbt ökar.

Ett antal interoperabilitetsproblem belystes i projektet. De tekniska problemen kring OAI-PMH föreföll relativt enkla att lösa. Större problem var knutna till metadatakvaliteten, framför allt när det gällde tolkning av formatet, beskrivningsnivåer och vokabulärer. En sammanfattning av projektet presenterades vid konferensen ETD 2003⁶ i Berlin.⁷

Projektet lämnade rekommendationer som på lång sikt syftade till att bygga upp en mera konsistent miljö för utbyte, spridning och återvinning av metadata om forskningspublikationer i Sverige. Att samordna dessa frågor skapar även förutsättningar för meningsfulla tjänster på olika nivåer.

Det är viktigt att uppmärksamma dessa frågor så att strategiska val prövas, rekommendationer och samsyn för t. ex. metadataprofiler för olika nivåer på kompatibilitet diskuteras och slutligen väl underbyggda rekommendationer kan ges.

Resultaten från förstudien visar hur viktigt det är med en gemensam strategi som stödjer interoperabilitet. Att bygga upp en teknisk infrastruktur är viktigt. Att denna infrastruktur kan samverka på ett meningsfullt sätt är kanske ännu viktigare. Det räcker inte med att stödja standarder.

¹ Se <http://www.kb.se/bibsam/vetpubl/svep/prjbeskrivning.pdf>

² Se <http://www.kb.se/bibsam/>

³ Müller, Eva et al.: Projekt rapport "En gemensam portal för akademisk fulltextpublicering vid svenska universitet och högskolor": akademisk forskning online. Uppsala, April 2003

⁴ Se <http://www.openarchives.org/OAI/openarchivesprotocol.html>

⁵ Se <http://publications.uu.se/epcentre/>

⁶ Se <http://www.hu-berlin.de/etd2003/>

⁷ Se <http://publications.uu.se/etd2003/papers/afo.pdf>

Tolkningen av dessa standarder går inte att standardisera – där krävs en rad av överenskommelser så att en större konsistens i metadatabeskrivningar kan uppnås.

En fortsättningsansökan gjordes till BIBSAM som så småningom samordnade ett nytt projekt med flera andra projekt i SVEP⁸ (Samordning av den svenska högskolans elektroniska publicering) där det nu ingår som ett av de fem delprojekten under namnet DP1: Interoperabilitet. DP1 samordnas av Enheten för digital publicering vid Uppsala universitetsbibliotek.

Målet för detta delprojekt är att det, på kort sikt, skall leda till en harmonisering av metadatabeskrivningar för elektroniskt publicerade dokument vid svenska universitet och högskolor och ett antal rekommendationer som främjar utbyte av information om dessa dokument och som skapar förutsättningar för nya tjänster. Projektet gör det möjligt att i ett relevant forum diskutera dessa viktiga frågor på ett konkret sätt. Detta skall skapa förutsättningar för att resultaten av projektet blir väl förankrade och kommer att användas i praktiken.

I ett längre perspektiv är förhoppningen att man, genom att dessa frågor diskuteras och konkreta överenskommelser och rekommendationer tas fram och används i praktiken, kommer att ha möjlighet att återanvända resurserna i många olika sammanhang. Denna strategi kommer på sikt leda till att svenska arkiv kan samverka på ett meningsfullt sätt och nya avancerade tjänster kan byggas både på nationell och internationell nivå.

Övergripande mål för SVEP

Projektet ska främja en mer samordnad och kraftfull utveckling av elektronisk publicering av forskares och studenters egna arbeten vid svenska universitet och högskolor. Genom samordning och rådgivning ska alla lärosäten få stöd att lägga ut sina forskares och studenters arbeten elektroniskt. Den elektroniska publiceringen ska ske på ett sätt som främjar maximal synlighet och långsiktig tillgänglighet. Arbetsmetodik och system för publicering ska vara resurssnåla och undvika dubblering av insatser för att bli långsiktigt hållbara.

Projektet strävar efter att underlätta för biblioteken genom att gemensamt definiera standarder för metadata om publikationer, sprida lösningar för långsiktig tillgänglighet, ge råd om tillgängliga verktyg och system för elektronisk publicering samt utveckla nya söktjänster. Genom samordning och standardisering kan också bredare spridning och långsiktig tillgänglighet uppnås.

Projektet är inriktat på fritt tillgängliga vetenskapliga publikationer utgivna vid enskilda svenska universitet och högskolor. Här innefattas även examensarbeten, eftersom de syftar till skolning i vetenskaplig publicering. Utanför avgränsningen faller andra publikationer från universitet och högskolor, t.ex. av mer administrativ karaktär. För examensarbeten skapas en särskild söktjänst riktad mot en nationell publik.⁹

De övergripande målen i förhållande till DP1

De viktigaste begreppen för DP1 i förhållande till SVEP:s övergripande mål är: *maximal synlighet*, *långsiktig tillgänglighet* och *långsiktig hållbarhet*. Sammanfattningsvis kan man säga att effektiviteten i sättet att inhämta metadata har stor betydelse för en långsiktig hållbarhet. Är det för dyrt eller komplicerat att skapa tillräckliga metadata kommer sådana lösningar knappast att överleva i längden. För långsiktig tillgänglighet kommer strukturerade och utförliga metadata att vara av stor betydelse dels för att i framtiden kunna söka och identifiera dokumenten på ett bra sätt dels för att kunna skapa ytterligare arkiveringsmetadata "automatiskt" utifrån beskrivande metadata. "Bestående" identifikatorer kommer också att vara mycket viktiga i detta sammanhang. Maximal synlighet uppnås just nu bäst via allmänna söktjänster. Om nya och bättre söktjänster, t.ex. via OAI-PMH, skall skapas kommer detta att kräva mer strukturerad metadata som kan utnyttjas för verkningsfullare sökningar.

⁸ Se <http://www.svep-projekt.se>

⁹ Se: <http://www.kb.se/bibsam/vetpubl/svep/prjbeskrivning.pdf>

Maximal synlighet

För DP1:s del betyder detta mål att den metadata som skapas via olika system för elektronisk publicering bör kunna synliggöra och sprida information om publikationerna på effektivast möjliga sätt.

I samband med detta har man på forskningsbiblioteken visat stort intresse för the Open Archives Initiative Protocol for Metadata Harvesting¹⁰ (OAI-PMH). Detta är ett enkelt protokoll för kommunikation mellan olika datorer avsett för att skapa ett applikationsoberoende ramverk för interoperabilitet baserat på *metadata harvesting*¹¹. Ramverket innehåller två typer av deltagare:

- *Data Providers* (repositories) administer systems that support the OAI-PMH as a means of exposing metadata; and
- *Service Providers* (harvesters) use metadata harvested via the OAI-PMH as a basis for building value-added services.

Syftet är alltså att söktjänster - mer eller mindre avancerade - skall kunna byggas upp av *service providers* med hjälp av fritt tillgängliga metadataposter från *data providers*. Ett exempel på en sådan tjänst är den särskilda söktjänsten för examensarbeten inom SVEP:s DP3.

Data providers måste erbjuda sina metadataposter i ett okvalificerat Dublin Core-format som är *used as the mandatory "Lowest Common Denominator" metadata record format in OAI-PMH*.¹² Därmed är det inte sagt att detta är det enda format som kan användas. I praktiken kan vilket XML-baserat format som helst användas - och detta snarast uppmuntras eftersom erfarenheter bland annat visar att:

- The DCMES [Dublin Core Metadata Element Set] is not used to its fullest extent (understatement)
- Due to this underutilization of the DCMES, it will be difficult for the OAI community to build relevant cross-resource services based on it¹³
- En undersökning av användningen av Uppsalas webserver för universitetets avhandlingar och andra publikationer (<http://publications.uu.se>) visar att av 1.500.000 nerladdade sidor under ett år (2003) nådde användarna (se även bilaga):
 - 70% via egna bokmärken eller interna länkar inom webbplatsen
 - 20% via Google¹⁴
 - 10% via länkar från externa webbplatser
 - Enstaka sidor (ca 4.500) via bibliotekskataloger eller OAI-baserade söktjänster

Slutsatsen är att det otvivelaktigt effektivaste sättet att uppnå målet maximal synlighet för närvarande är att göra väl utformade vanliga webbsidor för varje dokument (vilket inom parentes sagt här betyder att göra så att det är möjligt att göra effektiva sökningar inom Googles rammar) och att se till att dessa indexerats i Google. Enklaste sättet att göra det senare är att skapa olika typer av bläddringsgränssnitt i publiceringssystemen som sökrobotarna kan gå igenom för att hitta alla dokument. Denna funktionalitet finns inbyggd i exempelvis DiVA, DSpace och Eprints.

Än så länge genererar inte söktjänsterna från OAI:s *service providers* annat än marginell användning av DiVA. Dessa tjänsters problem består troligen främst i den ojämna och grunda användningen av Dublin Core (se mer nedan under *olika beskrivningsnivåer*) samt det faktum att de inte alltid erbjuder bläddringsgångar som Google och andra sökrobotar kan utnyttja. Därmed faller

¹⁰ Se <http://www.openarchives.org/OAI/openarchivesprotocol.html>

¹¹ Se <http://www.openarchives.org/OAI/openarchivesprotocol.html#harvester>

¹² Se <http://www.oaforum.org/tutorial/english/page5.htm>

¹³ Ward, J. (2003). A Quantitative Analysis of Unqualified Dublin Core Metadata Element Set Usage within Data Providers Registered with the Open Archives Initiative. Proceedings of the Third ACM/IEEE Joint Conference on Digital Libraries, s. 315-317

¹⁴ Se <http://www.google.com>

de under kategorin "the invisible web", dvs. det som inte hittas via vanliga sökröbotar (och därmed inte av en vanlig internetanvändare). Detta har nu i och för sig uppmärksammats av sådana arkiv som t.ex. OAIster¹⁵ som nyligen har inlett ett samarbete med Yahoo!¹⁶ för att infoga sina harvestade metadataposter i Yahoo!s söktjänst.¹⁷ Dock skall man i detta sammanhang komma ihåg att dessa begränsade metadataposter innehåller en väldigt liten del av den totala information som normalt är sökbar i söktjänsterna, nämligen hela dokumentet. Även DSpace har inlett ett samarbete med en söktjänst¹⁸ - Google - i syfte att göra sina poster mer åtkomliga. Det kommer, som redan nämnts, troligtvis att krävas betydligt mer avancerad funktionalitet i söktjänster som byggs upp via OAI-PMH - eller skapar samsökning på annat sätt - än den enkla nivå som nu finns i existerande tjänster (som OAIster) för att dessa skall erbjuda ett alternativ till Google eller till de informationstjänster biblioteken erbjuder via kommersiella databaser.

I exemplet DiVA exporteras också metadataposter till LIBRIS, det nationella biblioteksdatasystemet, och därifrån sedan vidare till lokala bibliotekssystem. Den stora vinsten här ligger i första hand på registreringssidan och hanteringen av de tryckta publikationerna i biblioteken. När det däremot gäller de elektroniska publikationernas användning har bibliotekssystemen liten betydelse för denna. Däremot kommer de sannolikt, p.g.a. sin relativa stabilitet, att ha större betydelse för målet långsiktig tillgänglighet.

Ett annat sätt att sprida information om (framför allt nya) dokument kan vara RSS¹⁹, ett XML-baserat format för nyhetssyndikering. I DiVA kom under 2003 fler besökare från denna tjänst än från OAI *service providers* och bibliotekskataloger tillsammans.

Långsiktig hållbarhet

För att uppnå målet långsiktig hållbarhet i den elektroniska publiceringen är det viktigt att inhämtandet eller skapandet av ursprunglig metadata sker på effektivast möjliga sätt - helst i samband med publiceringen - istället för att "primärkatalogisera" i efterhand vilket sannolikt dels blir kostsamt dels leder till lägre beskrivningsnivåer. Ett exempel på hur man effektivt kan inhämta metadata direkt från producenten är arbetsflödet inom DiVA publiceringssystem vilket baseras på att författarna skapar strukturerade originalmetadata och -dokument via mallar som sedan ligger till grund för alla övriga utprodukter som genereras; metadataposter såväl som tryckta och elektroniska dokument:

DiVA publishing workflow is based on the concept that data is entered in a structured form only once. The resulting structured document is then used throughout the entire chain for generating other products. For example, different metadata formats (a MARC record for the library catalogue or a record disseminated within the OAI framework); parts of the full text publication (the cover and the title page); or the presentation layer on the web are all created directly from the data delivered by the author.

The author creates the original document using word processor templates (or LaTeX macros). Both electronic and traditional print publications are created from a single source—the master file, which is produced from a file delivered by an author and created in a template for word processing. In this way, both the metadata and the structure of the document are marked up.²⁰

Genom att använda registreringsgränssnitt med kontroller av originaldata, exempelvis att institutioner, serier eller dokumenttyper väljs från kontrollerade listor eller att ISBN kontrolleras, är det möjligt att förenkla registreringen av strukturerade data samtidigt som man kan minska antalet felaktigheter.

Skapandet av sådana registreringsfunktioner ligger utanför DPI:s område men är viktiga att ha i åtanke när man bedömer olika möjliga tänkbara beskrivningsnivåer: ju bättre publiceringssystemen fungerar i detta avseende desto bättre metadata, när det gäller både struktur och tillförlitlighet, kan produceras.

¹⁵ Se <http://www.oaister.org/o/oaister/>

¹⁶ Se <http://www.yahoo.com>

¹⁷ Se <http://www.umich.edu/news/index.html?Releases/2004/Mar04/r031004>

¹⁸ Se <http://chronicle.com/free/2004/04/2004040901n.htm>

¹⁹ Se <http://blogs.law.harvard.edu/tech/rss>

²⁰ Se <http://www.dlib.org/dlib/november03/muller/11muller.html>

Långsiktig tillgänglighet

De elektroniska dokumentens tillgänglighet i ett långtidsperspektiv behandlas inom SVEP främst i DP2: *Långtidsbevarande/ett arbetsflöde och tekniska lösningar för arkivering av elektroniskt publicerade dokument vid svenska universitet och högskolor*, vars mål är att skapa en infrastruktur som bygger på gemensamma rutiner för överföring och lagring av dokument mellan en lokal producent och det nationella digitala arkivet. Genom att dokument levereras till ett arkiv på Kungl. biblioteket, KB, säkras tillgång till dokument både nu och i framtiden.

Inom ramen för DP2 har två arbetsgrupper skapats. En grupp är inriktad på mjuka frågor, dvs. organisations- och metadatastandarder, rekommendationer och deras tolkning. Just nu arbetar denna grupp med tolkning av Open Archival Information System, OAIS-standard²¹ för SVEP och identifiering av de metadata som är specifika för bevarande. Ett samarbete med Riksarkivet har också etablerats. En stor del av den metadata som kommer att krävas för att uppnå målet långsiktig tillgänglighet kommer inom DP2 att hämtas från rekommendationerna för beskrivande metadata i DP1. Beskrivande metadata kommer också att kunna ligga till grund för "automatiskt" skapande av för DP2 arkiverings-specifika metadata.

Den andra gruppen består huvudsakligen av systemutvecklare. Den undersöker tekniska lösningar för identifiering av resurser. Gruppen kommer att lämna ett förslag på implementering av tjänster, baserade på en prototyp till en URN:NBN resolver, som leder till att åtkomstgaranti till dokument publicerade vid svenska universitet och högskolor kan ges. Det kan inte nog poängteras hur viktiga de "bestående" länkarna är, exempelvis i förhållande till OAI-PMH-modellen eller bibliotekssystemen där metadata helt kopplas bort från själva fulltextdokumenten och enda kopplingen är länken.

Tekniska lösningar för elektronisk publicering vid svenska högskolor

Föregående års förstudie²² innehöll också en genomgång av de tekniska lösningar som används för elektronisk publicering vid svenska universitet och högskolor. Denna undersökning visade att omfattningen på publiceringen inte var så stor som förväntat och att endast ett fåtal hade tagit steget ut mot fulltextpublicering i större skala.

De tekniska lösningar som användes kunde grupperas i:

- lokala egenutvecklade system (baserade på antingen ett databassystem eller rena webbsidor).
- DiVA (utvecklat av Uppsala universitetsbibliotek i ett samarbetsprojekt mellan ett antal nordiska universitet).
- Open source-lösning (EPrints).

Det förstnämnda alternativet var det klart vanligaste (och användes i första hand för publicering av uppsatser och examensarbeten). I förstudien föreslogs att en oberoende utvärdering av befintliga och tillgängliga lösningar skulle kunna bidra till att BIBSAM eller andra organ skulle kunna ge "välbelagda rekommendationer baserade på t.ex. komplexitet på tjänsterna, framtidspotential och kostnader".²³

Delvis kommer en sådan utvärdering nu att göras inom SVEP:s delprojekt 4 (DP4: E-publicering) som:

Förutom att vi skapar denna webbplats [<http://www.svep-projekt.se>] för att sprida information om projektet och elektronisk publicering inom SVEP, kommer vi att fungera som resurspersoner för rådgivning, och under våren 2004 göra en sammanställning av olika publiceringsprogramvaror.²⁴

²¹ Se <http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf>

²² Müller, Eva et al., s. 6-9

²³ Ibid., s. 8

²⁴ Se <http://www.svep-projekt.se/e-publishing/>

Delprojekt 4 leds av Lunds universitetsbibliotek. Nämnda (i skrivande stund ännu opublicerade) sammanställning av publiceringsprogramvaror omfattar fem olika system, tre open source-lösningar (1-3), DiVA (se ovan) samt ett lokalt utvecklat system baserat på Lotus-Notes²⁵:

1. EPrints²⁶
2. DSpace²⁷
3. CDSware²⁸
4. DiVA²⁹
5. Lokalt egenutvecklat system vid Blekinge tekniska högskola

En slutsats som kan dras från förstudien är att den stora mängden olika egenutvecklade system troligen var den största enskilda orsaken till de interoperabilitetsproblem som där uppdagades. Dessa problem kan i princip sammanfattas som:

- Olika tolkningar av formatet (samma information i olika "fält" eller tvärtom)
- Inga "katalogiseringsregler" (Svensson, Anders och Anders Svensson likställda)
- Olika beskrivningsnivåer (ojämn standard, data saknas)
- Felaktig information (eller ej fungerande länkar)
- Ej kompatibla vokabulärer (icke enhetliga namnformer, ämneskategorier, typer)

Värt att notera är att en viss förskjutning nu tycks ha skett av högskolornas och högskolebibliotekens intresse från individuell lokal egenutveckling mot de mer generella system som de fyra första i ovanstående lista representerar. Antalet deltagare i DiVA-samarbetet har t.ex. mer än fördubblats sedan pilotstudien genomfördes. Möjligen har SVEP:s aktiviteter inom delprojekt 4 (rådgivning) och 5 (workshops) redan haft en viss påverkan på detta.

En aktuell utredning som stöder detta antagande är *Rapport angående elektronisk publicering vid Växjö universitet*³⁰ (publicerad 2004-01-25). I Växjö:s rapport är de alternativ som anges som tänkbara för universitetets e-publiceringsverksamhet, i linje med ovanstående, antingen en open source-lösning (EPrints) eller DiVA. I vilket fall som helst kommer en ökad användning av mer generella lösningar att, åtminstone sett ur ett strikt metadataperspektiv, i sig själv underlätta samverkan mellan olika system.

Olika metadataformat som används

Det torde stå klart att metadata kommit att spela en allt större roll, och att valet av katalogiseringsformat kan få mycket stor betydelse för samlingarnas framtida åtkomlighet. De olika samordningsverktyg som utvecklats marginaliserar inte formatet, utan gör det istället än mer viktigt. Det format biblioteket väljer måste vara flexibelt och tillåta interoperabilitet. Då katalogisering över huvud taget är en avsevärd investering måste formatet också väljas med hänsyn till framtida användning och behov.³¹

I de generella publiceringssystemen som ingår i SVEP DP4-studien används tre olika metadataformat som grundformat:

- DC - Dublin Core³² (Eprints och DSpace)
- MARC 21 - MACHine-Readable Cataloging³³ (CDSware)

²⁵ Se <http://www.lotus.com/products/product4.nsf/wdocs/noteshomepage>

²⁶ Se <http://www.eprints.org/>

²⁷ Se <http://www.dspace.org/>

²⁸ Se <http://cdsware.cern.ch/>

²⁹ Se <http://www.diva-portal.se/about.xsql?lang=sv>

³⁰ Se <http://www.bib.vxu.se/eprints/rapport.pdf>

³¹ Björkhem, Miriam and Lindholm, Jessica (2000) *Metadata för det digitala biblioteket*, Master Thesis, Lund University. BIVILs skriftserie 2000:7. (<http://www.kult.lu.se/bivil/publikationer/fulltext00/2000-7.pdf>)

³² Se <http://www.dublincore.org/>

- DiVA dokumentformat³⁴ (DiVA)

DC och MARC är välkända inom bibliotekssektorn och kräver här ingen närmare presentation. Utomstående kan få en utförlig beskrivning av de båda formaten i uppsatsen *Metadata för det digitala biblioteket* av Miriam Björkhem och Jessica Lindholm.³⁵

DiVA dokumentformat (Document Format) avviker från de övriga två eftersom det är ett generellt format för att skapa hela XML-baserade dokument och inte bara metadataposter. DiVA dokumentformat har skapats vid Enheten för digital publicering vid Uppsala universitetsbibliotek eftersom inget av de format som undersöktes kunde möta de krav som ställdes på det publiceringssystem som skulle utvecklas av enheten:

... one of the goals of the DiVA project was to create a workflow where information from the authors' original documents could be reused to extract metadata for various purposes and, ultimately, to extract the complete document in XML.

We evaluated a number of existing schemas [metadata schemas and DocBook³⁶ and TEI³⁷ for the encoding of documents]. Unfortunately, none of the schemas we evaluated met all our requirements for DiVA. Many limitations were found, not only in the granularity of the description, but also in the ability to express relationships and hierarchies and in the extensibility of the schemas. Consequently, we decided to develop a new schema: the DiVA Document Format.

The DiVA publishing workflow makes it possible to capture data at a deep level of granularity. We didn't want to lose the ability to capture this structured data; it is still relatively easy to produce structures with a lower granularity level from those with more structured data. It is more complicated—and in many cases it is impossible—to do it in the opposite way, i.e., to produce structured data from non-structured information.

Another DiVA requirement involved making the structured DiVA Document Format compatible with a number of metadata schemas and standards. The idea was to be able to easily generate other formats from the basic, structured document format. In the context of producing other schemas from the basic schema, the granularity of the description is not enough. In many cases, it is necessary to be able to express relationships and hierarchies.

Metadata recommendations (for example, in the area of the rights metadata and preservation metadata) are still under development, and new standards and recommendations will become available over time. Therefore, one of the requirements for the DiVA format was that it should be extensible.³⁸

Olika beskrivningsnivåer

Av de tre formaten, DC, MARC 21 respektive DiVA dokumentformats metadatadel, skiljer sig otvivelaktigt DC från de övriga genom sin betydligt enklare uppbyggnad: femton, icke-hierarkiskt ordnade, olika element varav i praktiken oftast bara några används i någon större utsträckning. I uppsatsen *A Quantitative Analysis of Dublin Core Metadata Element Set (DCMES) Usage in Data Providers Registered with the Open Archives Initiative (OAI)*³⁹, som undersökte DC-användningen hos 100 OAI data providers, anges att åtta element genomsnittligt användes per post och att fem av de femton elementen (creator, identifier, title, date och type) användes i 71% av posterna.

Det grundläggande problemet med DC, sett i ett långsiktigt perspektiv, är den grunda beskrivningsnivån vilket medför att om detta är det ursprungliga lagringsformatet blir det besvärligt att lägga till ytterligare struktur eller information på ett enkelt sätt, medan det motsatta - att göra ett enklare format utifrån ett mer strukturerat - normalt sett är okomplicerat. Detta har sedan tidigare också konstaterats av flera, se t.ex. Sten Hedbergs *Bruket av metadata enligt Dublin Core: Principer, teknik och tillämpningar utanför Sverige*:

³³ Se <http://www.loc.gov/marc/>

³⁴ Se <http://publications.uu.se/schema/ddf/>

³⁵ Björkhem, Miriam and Lindholm, Jessica (2000) *Metadata för det digitala biblioteket*, Master Thesis, Lund University. BIVILs skriftserie 2000:7. (<http://www.kult.lu.se/bivil/publikationer/fulltext00/2000-7.pdf>)

³⁶ Se <http://www.docbook.org/>

³⁷ Se <http://www.tei-c.org/>

³⁸ Se <http://www.dlib.org/dlib/november03/muller/11muller.html>

³⁹ Se http://www.foar.net/research/mp/Jewel_Ward-MPaper-November2002.pdf

... Det gemensamma i dessa redovisningar är de svårigheter som går tillbaka på att DC är ett grovt format medan de olika MARC-formaten är väsentligt mera detaljerade. Medan det alltså skulle gå lätt att översätta en MARC-post till DC ... klarar man inte den andra riktningen, till MARC från DC, inte ens om samtliga de ovan skisserade utvidgningarna och typ-tilläggen till DC-fältens namn utnyttjades. Scheme-tillägg, analys av textsträngar och av repetitioner av fält måste dessutom tillgripas.⁴⁰

Förutom den grunda beskrivningsnivån är det ostrukturerade innehållet i DC-fälten ett minst lika stort problem vid skapandet av mer strukturerade tjänster. I en OAI-PMH-baserad tjänst som OAIster är strukturerad sökning⁴¹ möjlig endast i fyra fält:⁴²

- **Title:** This will look for titles of books, articles, journals, audio files, etc., in the title field of the digital resource record.
- **Author/Creator:** This will look for authors of books, creators of paintings, institutions responsible for a pamphlet, etc., in the author/creator field of the digital resource record.
- **Subject:** This will look for words or phrases that have been used to describe the topical nature of a digital resource. These are subjects used by the resource publisher to classify the resource.
- **Resource Type:** This will look for certain kinds of resource types, i.e., text, image, audio, video. This is not a fully comprehensive search due to limitations in normalizing this field.

Ett exempel på svårigheten att konstruera en strukturerad sökning på författare med DC som underlag kan illustreras av en sökning i OAIster på namnet "Stefan Andersson". Denna resulterar i en träff där data är lagrade som "Stefan Andersson".

En sökning på "Andersson Stefan" däremot resulterar i tolv andra träffar där några innehåller författaruppgiften lagrad som "Andersson, Stefan" men man hittar även poster med flera författare lagrade som en sträng i samma fält, t.ex.: "Arne Andersson, Stefan Nilsson".

Kontrollerade vokabulärer

I bakgrundsstudien uppmärksammades även behovet av att utarbeta rekommendationer lämpliga för forskningsdokument för kontrollerade vokabulärer och setstrukturer i OAI-PMH. Vad gäller områdena *kontrollerade vokabulärer* respektive *kontrollerade namnformer* i DP1 har dessa, efter det att projektbudgeten slutligen fastställts, skurits ner i förhållande till den ursprungliga planen. Insatserna kommer nu att fokuseras på tre huvudområden:

Gemensam ämneskategorilista

Inom DiVA-samarbetet har de deltagande högskolorna och universiteten enats om att använda en gemensam ämneskategorilista för att i DiVA-portalen kunna skapa ett enkelt sammanhållet och hierarkiskt ämnesbaserat bläddringsgränssnitt typ Yahoo.⁴³ Ämneskategorierna används också i DiVA för att erbjuda ämnesspecifik selektiv harvesting av metadataposter (s.k. sets) via OAI-PMH.⁴⁴ I detta sammanhang bör det betonas att ämneskategorierna främst är tänkta att fylla dessa två funktioner och inte ersätta mer förfinade ämnesord eller klassifikationskoder som även dessa naturligtvis kan tillfoga bibliografiska poster.

Inom det nationella samarbetet kring system för redovisning av vetenskaplig publicering (se vidare nedan) har också en samordning av en ämneskategorilista diskuterats. Både DiVA-portalen och den bibliografiska databasen (publikationsdatabasen) vid Uppsala universitet bygger i grunden sin kategoriindelning på Statistiska Centralbyråns (SCB) *Nationella förteckning över forskningsämnen*.⁴⁵

⁴⁰ Sten Hedberg (192001 [sic]) <http://www.kb.se/nvb/Metadata/meta7.htm>

⁴¹ Se <http://oaiSTER.umdl.umich.edu/cgi/b/bib/bib-idx?c=oaiSTER;page=simple>

⁴² Se <http://oaiSTER.umdl.umich.edu/o/oaiSTER/help.html#ti>

⁴³ Se <http://www.diva-portal.se/subjectindex.xsql>

⁴⁴ Se <http://www.openarchives.org/OAI/openarchivesprotocol.html#Set>

⁴⁵ Se http://www.scb.se/templates/Standard____24458.asp

SCB har haft ett regeringsuppdrag att utveckla forskningsstatistiken så att en ämnesuppdelad redovisning av statistiken över forskning och utveckling blir möjlig.

DiVA-portalens ämnesträd skapades ursprungligen som en del av ett projekt för att fastställa en standard för metadatamärkning av webbsidor på Uppsala universitet. I detta projekt bestämdes att standardiserade vokabulärer skulle användas i så utsträckning som möjligt för att garantera semantisk kompatibilitet.

Det ingick som en särskild del att undersöka vilken eller vilka ämnesordstesaurusar som skulle kunna användas. Denna del av projektet samordnades också med universitetsbibliotekets webbgrupps arbete med att ämnesindela elektroniska resurser, framförallt elektroniska tidskrifter och databaser samt med arbetet att bygga upp en studentportal med visning av biblioteksresurser för studenter.

För att kunna erbjuda en bläddringsstruktur för användare som vill kunna söka sig fram bland resurser knutna till universitetets ämnen, utan att känna till organisationen, fanns ett önskemål om en begränsad vokabulär. Genom att använda en sådan blev det möjligt att göra länknings mellan olika typer av resurser och göra dem sökbara i ett ämnesträd. Vokabulären skulle spegla forsknings- och utbildningsämnen vid Uppsala universitet. Det fanns också önskemål om att de ämnesrubriker som används i universitetets utbildningskatalog skulle finnas med i vokabulären.

Resultatet blev då en lista över ämneskategorier som utgick från SCB:s nationella förteckning men som innehöll en rad tillägg, speciellt inom humaniora och samhällsvetenskap där SCB:s lista håller en mycket grund nivå. Denna utökade lista har sedan utan vidare anpassning infogats i DiVA-systemet.

I och med att fler och fler universitet (dels med olika ämnesprofiler dels i andra nordiska länder) har anslutit sig till DiVA-samarbetet har frågan om en revidering av ämneskategorilistan kommit upp. Därför bildades i mars 2004 en arbetsgrupp inom detta samarbete för en översyn av denna.

Samtidigt har frågan, som redan nämnts, aktualiserats inom det nationella samarbetet kring system för redovisning av vetenskaplig publicering. Därför kommer vi inom DP1 att försöka samordna dessa båda andra initiativ med SVEP med avsikten att nå en så bred och användbar lösning som möjligt. Ett konkret förslag är tänkt att föreligga i december 2004. Det handlar också om hur man i så fall organiserar underhållet av en sådan gemensam lista. Utgångspunkten är att en gemensam kärna av kategorier skall kunna användas av så många som möjligt samtidigt som lokala påbyggnader skall kunna göras djupare i en trädstruktur.

Kontrollerade termer för dokumenttyper

Inte minst i arbetet med att skapa den särskilda söktjänsten för "examensarbeten" inom SVEP:s delprojekt 3 har frågan rests om vad för slags dokument det egentligen är som publiceras.⁴⁶

Mer precisa definitioner av begrepp som "examensarbete" och "uppsats" (och vilka dokument som är vilket) är önskvärda och ett enhetligt användande av liknande termer i svenska metadaposter vore till stor nytta. Detta gäller även i ett internationellt perspektiv. De söktjänster för avhandlingar som finns på de svenska universitetsbibliotekens webbplatser använder t.ex. omväxlande termerna *dissertations* och (*doctoral*) *theses* för avhandlingar i sina engelska versioner. Om man föreställer sig att metadata från de svenska söktjänsterna sprids ut i världen via - exempelvis - OAI-PMH vore det fördelaktigt om det fanns en enhetlig engelsk vokabulär som beskriver olika typer av forskningsdokument från de svenska högskolorna. Även ur arkiveringssynpunkt (SVEP DP2) kommer det att vara av intresse att veta varför ett visst dokument har skapats, exempelvis för att få en viss examen.

Därför kommer vi inom DP1 att utarbeta ett förslag till definitioner av de svenska begreppen med engelska översättningar.

Kontrollerade namnformer

När det gäller kontrollerade namnformer kommer vi inom DP1 att koncentrera oss på att ta fram rekommendationer för hur institutionsnamn kan normaliseras. Anledningen är att dessa utgör en mycket viktig sökingång som är svår att skapa på ett bra sätt om inte informationen är väl strukturerad.

⁴⁶ Diskussion på SVEP:s mailinglist (<http://www.lub.lu.se/svep/vidare/>)

I praktiken har detta problem också redan visat sig i uppbyggnaden av den nya söktjänsten för examensarbeten i SVEP DP3 där strängbaserad information i ett enda fält gör en sådan ingång svår att åstadkomma⁴⁷ även om rekommendationer finns för hur textsträngen skall formateras.⁴⁸

Idealfallet är att det går att skapa tjänster med även "organisatorisk" bläddring. Som exempel på en sådan funktion kan nämnas en prototyp som skapades inom projektet *Standard för metadata på Uppsala universitets webbplats*⁴⁹ där universitetets webbsidor kan bläddras fram via organisationsstrukturen. För att detta skall fungera krävs en hierarkisk struktur på metadatabeskrivningen i vilken över- respektive underordnade institutioner/avdelningar kan inordnas.

Dessutom bör registreringen av denna information kopplas till någon typ av centralt underhållet register (som ett universitets adresskatalog). I t.ex. DiVA-systemet väljs universitetets institutioner från kontrollerade listor vilket också gör det möjligt att i olika exportformat exempelvis plocka fram olika korrekta språkformer (svenska eller engelska), olika antal nivåer i beskrivningen eller lägga till adressuppgifter beroende på vart posterna skall skickas.

Vidare finns det även här en koppling till metadata för långtidslagring (DP2) vad gäller ursprunglig utgivare, rättigheter m.m.

Inom DP1 kommer vi att föreslå en modell för hur detta kan gå till. Denna modell skall enligt planeringen presenteras i april 2005.

Bibliografiska databaser/publikationsdatabaser

Den 28 januari 2004 hölls i Uppsala ett möte om nationellt samarbete kring system för redovisning av vetenskaplig publicering (i Uppsala benämnt den *bibliografiska databasen* men på andra universitet ofta *publikationsdatabaser*). BIBSAM välkomnade på detta möte att den fortsatta diskussionen av bibliografiska format rörande dessa system sker inom SVEP-projektet. I utvecklingsprogrammet för Svenskt nätbibliotek föreslås även att det sker en samordning av lokala publikationsdatabaser.

Ur protokollet från mötet den 28 januari:

Deltagarna i tekniksessionen var överens om att lokala, decentraliserade system kommer att behövas på grund av de olika behoven av uppföljning vid olika universitet. Vi finner det inte sannolikt att detta kan lösas inom en snar framtid med hjälp av ett nationellt centralt system och ett gemensamt postformat för analysposter.

Dock behövs samverkan för utbyte av poster bibliografisk information för att på bästa sätt kunna exponera universitetens publicering. Formatet för utbyte av bibliografisk information bör samordnas med rekommendationer som tas fram inom ramen för BIBSAM:s SvEP-projekt för att främja elektronisk publicering i Sverige. Formatet för utbyte bör följa rekommendationen OAI - Open Access Initiative - och posterna bör kunna flyttas mellan systemen med hjälp av OAI-PMH, OAI:s protokoll för postutbyte.

Det beslutades således att ett möte, främst ägnat åt formatfrågor, skulle hållas i SVEP:s regi under våren 2004 med syfte att ta fram ett bibliografiskt minimiformat inom DP1. Detta möte kommer att arrangeras i Uppsala den 23 april. Förutom det bibliografiska formatet kommer en samordning av ämneskategorier att diskuteras.

Nästa rapportering

Nästkommande arbetsrapport är planerad till juni 2004. Denna skall omfatta en konkret översyn av metadata för olika typer av publikationer i förhållande till olika nivåer på tjänster.

⁴⁷ Diskussion på SVEP:s mailinglist (<http://www.lub.lu.se/svep/vidare/>)

⁴⁸ Metadatamodell för Arkiv Ex - Sverige (<http://www.svep-projekt.se/masters-theses/Metadatamodel/>)

⁴⁹ Se <http://publications.uu.se/metadata/org.xsql?lang=sv>

Bilagor

1. SVEP DP1: Tidsplanering 2004-2005
2. Statistik <http://publications.uu.se/> år 2003

SVEP DP1: Tidsplanering 2004-2005

Aktivitet/resurser		2004												2005											
Nr	Aktivitet	D.*	Jan	Feb	Mar	Apr	Maj	Jun	Jul	Aug	Sep	Okt	Nov	Dec	Jan	Feb	Mar	Apr	Maj	Jun	Jul	Aug	Sep		
1	Projekttid.	50					♦				♦							♦							
2	Interoper.	20				♦																			
3	Pub.-typ.	55					♦																		
4	Vokabulär	20												♦											
5	Namnform.	50																♦							
6	Rek. 1	40																	♦						
7	Rek. 2	40																	♦						
8	Webbpl.	25																							
9	Spridning	25																							
10	Rapporter	25					♦											♦							

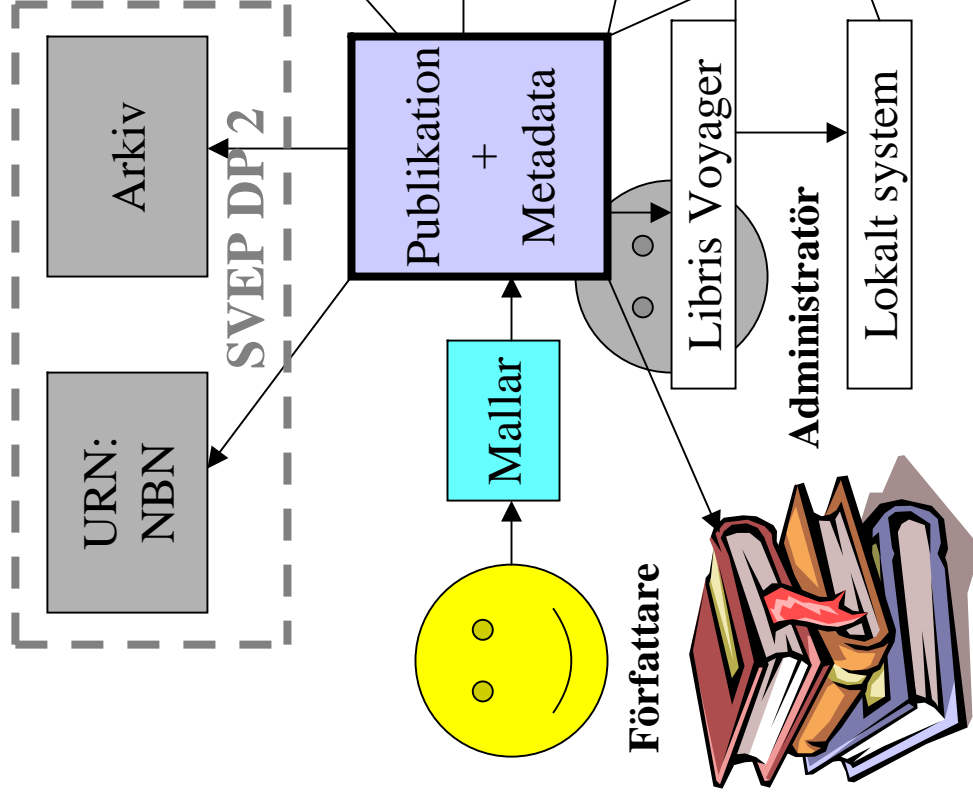
♦ = Leveranser (se nedan), Blå ruta = aktiviteten pågår

*Resursfördelning totalt i dagar

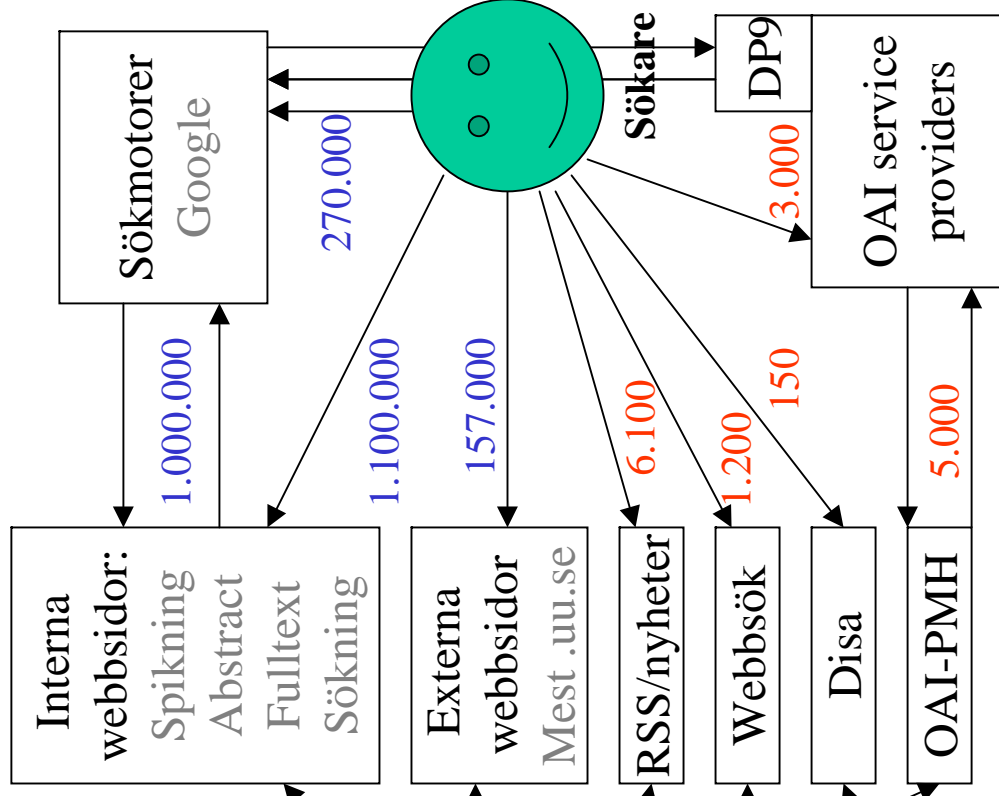
Nr	Namn	Leveranser	Resurser totalt (dagar)																					
			UU	LU	LU*	Libris	IDA	LIU	LUTH	SU	UMU	Totalt												
1	Projektleddning	Löpande + Maj + Okt 2004, Apr 2005: Delprojektmöte	26	4	4	Y	Y	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	50
2	Interoperabilitetsfrågor i svenskt perspektiv	Apr 2004: Arbetsrapport	14	1	1	Y	Y	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20
3	Metadata för olika typer av publikationer	Jun 2004: Arbetsrapport	49	1	1	Y	Y	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	55
4	Kontrollerade vokabulärer	Dec 2004: Arbetsrapport	14	1	1	Y	Y	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20
5	Kontrollerade namnformer	Apr 2005: Arbetsrapport	38	2	2	Y	Y	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	50
6	Rek. som främjar interoperabilitet	Jun 2005: Dokument	34	1	1	Y	Y	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	40
7	Rek. för nivåer på interoperabilitet	Jun 2005: Dokument	34	1	1	Y	Y	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	40
8	Webbplatsen	Löpande uppdatering	25	0	0	N	N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25
9	Resultatspridning	Maj 2004 + Apr 2005: Expertmöte; Sep 2005: Artikel	13	2	2	Y	Y	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	25
10	Avrapportering	Jun + Dec 2004, Jun + Sep 2005: Statusrapport	13	2	2	Y	Y	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	25
Totalt:			260	15	15	0	0	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	350

Resursfördelningen omfattar hela projektet (även redan använd tid under år 2003). Libris/IDA (Kungl. biblioteket) Y = deltar, N = deltar ej.
 UU = Uppsala universitetsbibliotek, LU = Lunds universitetsbibliotek (Biblioteksdirektionen), LU* = Lunds universitetsbibliotek (Medicinska biblioteket), LIU = Linköpings universitetsbibliotek, LUTH = Luleå tekniska högskolas bibliotek, SU = Stockholms universitetsbibliotek, UMU = Umeå universitetsbibliotek.

PUBLICERING OCH LAGRING



SPRIDNING



HARVESTING

Tryckta publikationer

<http://publications.uu.se> - statistik år 2003