Digisam

Aggregering och länkade data

– Digisams pilotprojekt

Följande rapport presenterar resultatet av Digisams pilotprojekt om aggregering och länkade data.

Rapporten är producerad av konsult Marie Gustafsson Friberger (Good Measure).

Innehåll

1. Inledning

Digisam har fjorton vägledande principer för arbetet med digitalisering, bevarande och tillgängliggörande av kulturarvet[1]. Enligt princip nummer elva ska semantiska beskrivningar ske på ett standardiserat sätt. En konsekvens av principen är: ”För att främja interoperabilitet på flera nivåer bör redan etablerade semantiska beskrivningar för tvärsektoriell kulturarvsdata återanvändas”. Ontologin CIDOC Conceptual Reference Model (CRM)[2] (ISO 21127:2006) skapar nya möjligheter för kulturarvsinstitutionerna att använda gemensamma ontologiska beskrivningar och att uttrycka dessa som länkade data. Digisams inventering, som redovisas i rapporten “Digitalisering av kulturarvet – nuläge och vägvalsfrågor”[3], visar dock att CIDOC CRM än så länge inte används i någon större utsträckning vid de medverkande institutionerna.

Den här rapporten har tagits fram i ett arbete som syftar till att vidareutveckla modellen för hur digital kulturarvsinformation skulle kunna tillgängliggöras på ett effektivare sätt genom semantiska webbens möjligheter. Som redan nämnts har tidigare rapporter identifierat ett behov av att undersöka om standarden är användbar för att skapa semantisk interoperabilitet mellan svenska arkiv, bibliotek och museer. Projektets initiala formulering var att undersöka olika sätt att strukturera och harmonisera kulturarvsinformation enligt internationella standarder och maskinläsbara format, samt att testa befintliga tjänster för mappning. När projektet hade påbörjats formulerades målen delvis om, för att fokusera mer på att utforska CIDOC CRMs fördelar och nackdelar för respektive område samt identifiera startpunkter för att använda CIDOC CRM.

Denna rapport bygger på tre workshops:

- Workshop 1 (13 juni, 2017) fokuserade på kartläggning av deltagande institutioners format och standarder samt introduktion till CIDOC CRM.
- Workshop 2 (11 september, 2017) fokuserade på fortsatt utvärdering av CIDOC CRM och identifikation av relevanta klasser i CIDOC CRM.
- Workshop 3 (23 oktober, 2017) fokuserade på initial mappning till CIDOC CRM utifrån resultatet från Workshop 2 samt diskussion kring metoder för att tillgängliggöra data.

Representanter från Digisam, K-samsök, Kungliga biblioteket, Musikverket, Nordiska museet, Riksantikvarieämbetet samt Riksarkivet deltog och resultaten presenterades i en delrapport per workshop (bifogade).

---

[1] Vägledande principer för arbetet med digitalt kulturarv. Beslutade av Digisams styrgrupp den 3 april 2014 – http://www.digisam.se/wp-content/uploads/2013/02/Vagledande_principer_for_arbetet_med_digitalt_kulturarv.pdf

[2] http://www.cidoc-crm.org

[3] http://www.digisam.se/wp-content/uploads/2013/05/Digitalisering%20av%20kulturarvet_nulage_och_vagvalsfragor.pdf

2. Bakgrund

2.1 CIDOC CRM

CIDOC CRM är en internationell standardontologi för kulturarvsdata. Bakgrunden till standarden är att olika tjänster ofta använder olika standarder för metadata, vilket bland annat medför att det saknas tjänster där användare på ett sömlöst vis kan söka relaterat material som kommer från flera källor. CIDOC CRM kan ses som ett ”semantiskt klister” för att mediera mellan olika källor till kulturarvsinformation, som publiceras av t ex arkiv, bibliotek och museer.

Standarden innehåller definitioner och formell struktur för att beskriva implicita och explicita koncept och relationer vid kulturarvsdokumentation. Den innehåller grundläggande klasser av entiteter (94 stycken) och egenskaper (*properties)* (168 stycken), ordnade i hierarkier som kan användas vid sökningar för att hitta nya samband. Man behöver inte använda alla dessa klasser och egenskaper, utan en delmängd kan väljas ut. Om begrepp som behövs för en viss domän saknas finns möjlighet att definiera dessa inom ramen för CIDOC CRM. Viktigt att notera är att CIDOC CRM är en händelsecentrerad modell: personer och objekt relateras inte direkt till tid, bara händelser länkas till tid.

CIDOC CRM kan upplevas som komplex och svåröverskådlig. Den kan ge intryck av att den behöver användas i sin helhet, men en fördel är att man kan börja med det mest basala som behövs för att koppla ihop information mellan olika institutioner.

CIDOC CRM Primer[4] är en bra startpunkt för att förstå standarden.

2.2 Länkade data

Länkade data är en uppsättning principer för datapublicering som gör det möjligt att följa samband mellan informationsobjekt. Länkade data bygger på att URI:er används för att identifiera de ting och relationer som beskrivs, att man använder tripplar av subjekt-predikat-objekt enligt standarden Resource Description Framework (RDF), samt att dessa kopplas samman till en graf. Genom att använda gemensam vokabulär går det att ställa frågor till ihopkopplade data och härleda relationer som man inte skulle kunna göra annars. En startpunkt för att läsa mer om länkade data är början av boken “Linked Data: Evolving the Web into a Global Data Space”[5].

[4] http://www.CIDOC-CRM.org/sites/default/files2/CRMPrimer_v1.1_1.pdf
[5] http://linkeddatabook.com

2.3 Metoder för att dela data

Om en organisation har ett dataset som de vill att andra ska kunna bygga applikationer på kan de använda olika metoder för att dela denna datamängd. Några sådana metoder är:

- Att tillhandahålla hela datamängden för nedladdning, en så kallad datadump eller batch. För detta kan flera olika format användas, så som kalkylblad och XML.
- Att använda ett Application Programming Interface (API) där applikationen som använder datasetet anropar på förhand definierade metoder för att hämta en delmängd av datasetet. En av de vanligare metoderna är att använda ett så kallat REST API.
- Att tillhandahålla en SPARQL endpoint, vilket är en möjlig lösning om länkade data används. Här kan frågespråket SPARQL användas för att ställa frågor mot ett eller flera dataset.

3. Kartläggning av standarder

Vid första workshopen kartlades vilka standarder som används vid de institutioner som medverkar i projektet. För ytterligare beskrivningar av dessa standarder, se även digisam.se/om-standarder/ samt Digisams rapport från 2014, "Digital informationshantering och infrastruktur för kulturarvet – Underlag för fortsatt arbete"[6].

Kungliga Biblioteket (KB)

Använda standarder

RDA (Resource Description and Access) – Anvisningar och riktlinjer för katalogisering och auktoritetsarbete. RDA stödjer sig på den konceptuella modellen **IFLA LRM** (tidigare FRBR, FRAD och FRSAD). RDA ersätter Katalogiseringsregler för svenska bibliotek (KRS). Mer om KB:s arbete med RDA finns att läsa på deras webbsida "Om RDA"[7].
Bibframe (Bibliographic Framework Initiative) – Vokabulär och beskrivande modell för bibliografiska beskrivningar för länkade data som utvecklas av Library of Congress. Målet är att ersätta MARC21. Bibframe har tre huvudsakliga abstraktionsnivåer, så kallade kärnklasser: Verk, Instans och Exemplar. Utöver dessa tre kärnklasser definieras tre nyckebegrepp: agent, ämnen och händelser. KB:s arbete med Bibframe beskrivs bl.a. i "Introduktion till nya Libris och XL, del 2"[8].
MARC (MAchine Readable Cataloging) – Föråldrat format för kataloginformation men kommer att finnas kvar.

[6] http://www.digisam.se/images/docs/rapporter/Digital%20informationshantering%20och%20infrastruktur%20for%20kulturarvet.pdf
[7] http://www.kb.se/bibliotek/Metadata/RDA/Om-RDA/
[8] http://www.kb.se/libris/Om-LIBRIS/Introduktion-till-nya-Libris-och-XL2/Del-1/

För bibliografisk beskrivning använder KB ytterligare ett antal vokabulärer, t ex SKOS, FOAF, MADSRDF. För ämnesbeskrivning används främst Dewey decimalklassifikation (DDK) och Svenska ämnesord (sao)

I övrigt används bland annat:
MODS (Metadata Object Description Schema)
EAD (Encoded Archival Description) – används för handskrifter
METS (Metadata Encoding and Transmission Standard)
PREMIS (Preservation Metadata: Implementation Strategies)

Relaterade standarder och kommentarer

IFLA LRM (IFLA Library Reference Model) är den konceptuella modell som RDA och Bibframe baseras på. IFLA LRM kopplar samman och ersätter FRBR, FRAD och FRSAD (se nedan).
FRBR (Functional Requirements for Bibliographic Records) – FRBR delar upp den bibliografiska informationen i olika nivåer: verk, uttryck, manifestation och exemplar.
FRAD (Functional Requirements for Authority Data)[9].
FRSAD (Functional Requirements for Subject Authority Data) – Utökning av FRBR för ämnesauktoriteter.
FRBRoo[10] – En objektorienterad version av FRBR, med mappning från FRBR och FRSAD mot CIDOC CRM.

I nuläget är flera av standarderna på biblioteksområdet i rörelse, med framtagandet av IFLA LRM, samt vidareutveckling av RDA och Bibframe. Även FRBRoo uppdateras för att vara i linje med IFLA LRM[11].

FRBRoo, Bibframe och IFLA LRM varierar något i hur de hanterar identifikation av verk, vilket blir än mer relevant när dessa kopplas till länkade data[12].

Riksantikvarieämbetet

MIDAS[13] (A Manual and Data Standard for Monument Inventories) – MIDAS är en guide för att skapa *inventories*, en detaljerad standard för att skapa full dokumentation av alla aspekter av kulturarv, snarare än en datastandard. Den är ett stöd för den metamodell som används för

---

[9] https://www.ifla.org/files/assets/cataloguing/frad/frad_2009-sv.pdf
[10] IFLA, Definition of FRBRoo: A Conceptual Model for Bibliographic Information in Object-Oriented Formalism: https://www.ifla.org/publications/node/11240
[11] Pat Riva and Maja Žumer (2017) The IFLA Library Reference Model, a step toward the Semantic Web http://library.ifla.org/1763/1/078-riva-en.pdf
[12] Preliminary White Paper från PCC SCS/LDAC Task Group on the Work Entity – http://www.loc.gov/aba/pcc/documents/PoCo-2017/WorkEntitity%20Preliminary%20White%20Paper-2017-09-27.pdf
[13] MIDAS Version 1.1 (Oktober 2012) https://content.historicengland.org.uk/images-books/publications/midas-heritage/midas-heritage-2012-v1_1.pdf/

att utveckla Riksantikvarieämbetets totala informationsarkitektur och kommer att implementeras i takt med övergången till ny teknisk plattform.
K-samsök – Aggregerar information från register om t ex byggnader, samt skördar data från andra institutioner, främst föremål, samlingar och bilder. Ett egenutvecklat protokoll[14] används, men detta är under omarbetning. Objekten i K-samsök får persistenta URI:er som pekar tillbaka till masterposter hos den ägande organisationen. Riksantikvarieämbetet undersöker möjligheterna att använda CIDOC CRM för K-samsök.
Evighetsrunor – Runinskrifter som informationsobjekt

Riksantikvarieämbetet har interna diskussioner kring hantering av periodindelningar och att gå från referenstypologier mot auktoritetslistor.

Nordiska museet

CIDOC CRM – Datamodellen i Primus, det system som Nordiska museet använder, är inspirerad av CIDOC CRM.
Spectrum – Standardiserade processer kring samlingsförvaltning.
Auktoriteter – Intern lösning inspirerad av CIDOC CRM och **EDM** (Europeana Data Model). Mappas mot **SKOS** (Simple Knowledge Organization System), **FOAF** (Friend of a Friend) och CIDOC CRM.
Outline – Äldre klassifikationssystem. Fortfarande i bruk men är inte aktuell för dagens samlingar utan uppdatering.
SIS-CEN terminologier
KulturNav – Nordiska museet förvaltar auktoritetstermer i KulturNav. Från KulturNav hämtar Nordiska museet data till Primus. KulturNav har en egen intern datamodell som inspirerats av modeller i CIDOC CRM och Europeana Data Model. KulturNav tillhandahåller länkade data med persistenta identifierare och går att använda även om Primus inte används.

Riksarkivet

Riksarkivet använder en tredelad informationsbeskrivningsstandard. Var och en av dessa tre delar kan stå på egna ben och relateras till varandra. Kopplat till dessa finns ett antal XML-baserade utbytesformat.
ISAD(G) (General International Standard Archival Description) och dess utbytesformat **EAD** – Beskriver arkivmaterialet, det vill säga själva handlingarna. En logisk beskrivning av arkivets uppbyggnad, den är hierarkisk och allt är beroende av varandra.
ISAAR(CPF) (International Standard Archival Authority Record For Corporate Bodies, Persons and Families) och dess utbytesformat **EAC** – Beskrivning av auktoriteter, det vill säga personer och organisationer, som har koppling till arkiven samt arkivbildare.
ISDF (International Standard for Describing Functions) och dess utbytesformat **EAC-F** (under utveckling) – Beskrivning av verksamheten som arkivbildarna har ägnat sig åt.

[14] http://www.ksamsok.se/resurser/protokoll-och-parametrar/

I övrigt används METS och Premis, samt några mindre kända standarder såsom de för pergamentomslagen.

Den hierarkiska modell som arkiv använder för sina samlingar gör att det vid tidigare försök har varit svårt att översätta arkivbeskrivningar till en objektbaserad konceptmodell. Koncept som arkivbildare saknas ofta när gemensamma standarder tas fram. I Kulturplattform Västernorrland var de områden som var mest intressanta för mappning kontextuella data, dvs. data om platser/händelser/orter/personer osv. som är gemensamma för hela kulturminnessektorn. De händelser som finns med i Riksarkivets material är ofta arkivrelaterade. Andra händelser kan länkas från annat material, på så vis skapas kopplingar till händelser som inte är arkivrelaterade. Dessa erfarenheter kring arkiv och CIDOC CRM finns beskrivna i "Arkivinformation + CIDOC CRM = sant (del 2)"[15].

Musikverket

Musikverket har museum, arkiv och bibliotek. För museets musikinstrumentsamling används MINST, en egenutvecklad Access-databas. För arkivmaterial och övriga föremål används arkivdatabasen CALM[16], som även hanterar föremål. CALM har en hierarkibaserad datastruktur. (oklart vilken standard/vokabulär) som stödjer EAD och OAI-PMH. För biblioteksmaterialet används (se listan för KB ovan för förklaringar av standarder): **FRBR, RDA** (ersätter AACR2-KRS), **MARC21, OAIPMH** (Open Archives Initiative Protocol for Metadata Harvesting) för informationsutbyte och Dublin Core. För klassifikationer används DDK och SAB.

[15] http://www.digisam.se/arkivinformation-cidoc-crm-sant-del-2/
[16] http://alm.axiell.com/sv/losningar/produkter/calm/

4. CIDOC CRM – Utvärdering och startpunkter

Som framgår av Avsnitt 3 används en stor mängd standarder vid de deltagande institutionerna. Varför är det då relevant att diskutera ytterligare en standard, som i nuläget endast används av en av parterna? Förutom den Digisam-princip som nämns i inledningen togs ett antal anledningar upp i projektplanen och vid workshopparna:

- Skapa interoperabilitet mellan olika områden.
- Göra det enklare för experter och allmänhet att röra sig mellan olika domäner genom att möjliggöra mer automatisk sammanföring av olika informationskällor (så att användaren inte behöver vända sig specifikt till var och en av institutionerna).
- Att man måste explicitgöra den interoperabilitet som idag "sitter i huvudet" på domänexperter.
- Tidsbesparing genom att antalet timmar till mappning sjunker då färre antal standarder behöver mappas mot.

Man bör ha i åtanke att användaren av mappad data kan vara både människor och programvara.

Vidare konstaterar workshopdeltagarna att även om metadata och standarder kan framstå som tekniska i sin natur är harmonisering av metadata en informationsfråga – och därmed en verksamhetsfråga snarare än en teknikfråga. Flera gånger kommer diskussionerna in på att förutsättningarna för interoperabilitet med gemensam standard påverkas av att de olika institutionerna har olika behov, beroende bl a på storlek och domän.

CIDOC CRM är föreslagen som en ontologi för att skapa interoperabilitet mellan olika källor inom kulturarvsområdet. Hur väl passar den för att skapa interoperabilitet mellan svenska arkiv, bibliotek och museer? Vilka är möjliga startpunkter och hur går man vidare från dessa?

Vid den andra workshoppen diskuterades fördelar och nackdelar med CIDOC CRM ur perspektiven arkiv, bibliotek, museer och allmän interoperabilitet (se Avsnitt 4.1). Vidare identifierades en delmängd klasser ur CIDOC CRM för att utföra en initial mappning för de deltagande institutionerna (se Avsnitt 4.2 och 4.3). Resultaten av denna mappning presenterades vid den tredje workshopen (se Avsnitt 4.4). Mappningsverktyget 3M presenteras kort, även om det var i mindre fokus i projektet än vad som var planerat (se Avsnitt 4.5). Möjliga vägar framåt i att utvärdera om CIDOC CRM bör användas för att skapa interoperabilitet mellan arkiv, bibliotek och museer presenteras sedan i Avsnitt 5.

4.1 CIDOC CRM ur olika institutioners perspektiv

Enligt det informationsmaterial som finns på webbsidorna för CIDOC CRM är det en lämplig ontologi för att användas till att föra samman data från museer, arkiv och bibliotek. Vid Workshop 1 konstaterades det att standardens lämplighet för arkiv och bibliotek inte är så god som dess skapare anser. Vid Workshop 2 gjordes därför en genomlysning av standardens

fördelar och nackdelar ur perspektiven arkiv, bibliotek och museer. Denna återges nedan och summeras i

Tabell 1.

Det är värt att notera att de flesta institutioner behöver hantera beskrivningar av materialtyper som inte tillhör majoritetsmaterialet. T ex ett arkiv med föremål eller ett museum med arkiv.

Arkiv
Plus: Riksarkivet har tidigare försökt att tillämpa CIDOC CRM-standarden på arkivdata. När det har lyckats bäst har det handlat om udda arkiv, t.ex. medeltida brev.

Minus: När det gäller konventionella arkiv tillämpar arkiven ett hierarkiskt tänkande där man utgår från arkivbildaren som i sin tur har arkiv, med serier och subserier, och därunder akter och underakter o.s.v. Denna typ av beskrivning är inte inbyggd i CIDOC CRM.

Kommentar: Digitalisering av själva innehållet i arkiven är ett mycket stort jobb.

Bibliotek
Plus: KB skulle antagligen kunna beskriva allt i sina samlingar med hjälp av modeller som FRBRoo och CIDOC CRM, men det skulle bli beskrivningar på en så generell nivå att det skulle bli meningslöst i praktiken. Däremot kan det vara möjligt att lyfta in delar av CIDOC CRM-vokabulären i Libris nya länkade data-lösning. Det skulle inte bara främja samarbetet med andra kulturinstitutioner utan även internt för KBs olika samlingstyper. Detta skulle även kunna användas för Musikverkets biblioteksmaterial.

Minus: Det mesta som ingår i bibliotekens samlingar är svårt att beskriva med CIDOC CRM. FRBRoo är ett försök att få ihop två olika modeller men den blir i sig ytterligare en komplex modell att behöva sätta sig in i.

Museer
Plus: På Nordiska museet är man van att arbeta i databaser som bygger på CIDOC CRM. Primus bygger på CIDOC CRM, liksom Kulturnav.

Minus: Musikverket/Musik- och teaterbiblioteket har delsamlingar som är strukturerade på olika sätt, där inget av de befintliga systemen stödjer CIDOC CRM.

Tabell 1. Summering av CIDOC CRM ur olika institutioners perspektiv.

	Fördelar	**Nackdelar**
Allmän interoperabilitet	De högre nivåerna av vokabulären innehåller koncept som är rimliga startpunkter för ett gemensamt urval. Här beskrivs även informationsobjekt, dvs. inte bara fysiska objekt som ofta finns främst hos museer.	CIDOC CRMs komplexitet. I grunden utvecklat för museer.
Arkiv	I de experiment som gjorts, kan standarden fungera väl med mindre arkiv.	Passar sämre för den hierarkiska modell som används i arkivens standarder.
Bibliotek	KB: Kan lyftas in KB:s vokabulär för länkade data där relevant. Musikverket: Möjligt att lägga till fler samband (som koppling till nationalitet, som ofta inte fångas upp av klassifikation eller ämnesord), t ex för noter, än vad som används idag.	Mycket som är viktigt för bibliotek beskrivs ej i CIDOC CRM. $FRBR_{OO}$ är ytterligare en komplicerad modell.
Museer	Används i Primus som har många museianvändare. Eftersom standarden är framtagen av internationell branschorganisation är det sannolikt att fler museisystem stödjer, eller kommer att stödja. Tar upp samlingsobjekt, föremål och processer kring föremål, ägarskap.	Domänen i sig är mer disparat – även om man tror att man är överens så är man kanske inte det. CIDOC CRMs komplexitet.

4.2 Urval för initial mappning

Ett första steg i en konkret utvärdering av CIDOC CRM som semantiskt klister mellan de deltagande institutionerna är att identifiera klasser ur CIDOC CRM där det förväntas finnas ett stort överlapp. Det är alltså inte fråga om att tillämpa hela standarden, utan att ta avstamp i ett urval av klasser som bedöms som särskilt relevanta. Utifrån detta finns senare möjlighet att identifiera om det saknas något i standarden som behövs för att skapa interoperabilitet.

Vid Workshop 2 utgick deltagarna från ett specifikt användningsområde: att hitta allt om en person, plats/område eller händelse. Utifrån en initial lista med koncept (se Tabell 2) som skulle kunna vara relevanta identifierades några huvudkategorier av koncept som särskilt intressanta som startpunkter i mappningsarbetet (se Tabell 3).

Tabell 2. Initial lista med koncept.

Koncept	**Eventuell kommentar**
Aktör (Personer och organisationer)	Aktör som begrepp är förhållandevis gemensamt över domänerna.
Platser	Sällan entydigt, men bra ingång för användare. Se längre diskussion nedan.
Processer/händelser	
Koncept	Kan bli mycket omfångsrikt
Saker/objekt	
Verk	
Tid/tidsperiod (ev. med koppling till plats)	En tidsperiod är ofta även geografiskt begränsad (t.ex. frihetstiden)
Knowledge organiation system (KOS)	Hierarkiska eller grafbaserade

Vidare diskussion om plats: Plats är sällan entydigt, även om plats ofta är en bra ingång för slutanvändare (som t ex ofta söker på sin födelseort). Vad gäller plats kan man lägga sig på både generell och detaljerad nivå. Inom kulturmiljöområdet är just plats väldigt intressant för utvecklare. Även på DigitaltMuseum är det orter som användare söker på i första hand.

Tabell 3. De koncept som valdes ut som startpunkter för mappningsarbetet.

Aktör
Tid
Saker/objekt (även informationsobjekt)
Plats

Trots att plats inte är entydigt och varierar i precision (som diskuteras ovan) så är det relevant att inkludera även i ett första steg. Plats är viktigt för allmänheten, en vanlig ingång till material i alla samlingar. Det kan dessutom vara en värdefull ingång för att illustrera skillnader mellan domäner.

4.3 Urval av CIDOC CRM-klasser för initial mappning

I smågrupper identifierades vilka klasser i CIDOC CRM som speglar koncepten i Tabell 3. Utöver dessa har ett fåtal klasser från CIDOC CRM som inte togs upp vid mötet lagts till i listorna nedan. För mer information om de olika klasser som nämns nedan, se dokumentationen för senaste versionen av CIDOC CRM[17].

Aktör

- ***E39 Actor*** och dess subklasser ***E21 Person*** och ***E74 Group*** (samt dess subklass ***E40 LegalBody***)

Tid

- ***E4 Period***
- ***E12 Production***
- ***E49 Time appellation***
- ***E63 Beginning of existence***
- ***E65 Creation***
- ***E66 Formation***
- ***E67 Birth***
- ***E68 Dissolution***
- ***E69 Death***

Kommentarer:

- Det är svårt att hitta hur man beskriver när ett föremål har använts.
- Det händelsebaserade synsättet i CIDOC CRM är annorlunda än vad som används av t ex KB. Det går dock att mappa så att det som de normalt uttrycker med en egenskap mappas till en specifik sorts händelse.

Saker/objekt

- **E70 Thing**, har flera klasser som ärver från den, där man behöver identifiera relevant nivå för denna initiala mappning. En fördel med CIDOC CRM är att man kan börja med E70 och sedan ge mer detaljer på sikt.
- ***E24 Physical Man-Made Thing***
- ***E73 Information Object***
- ***E78 Collection***

Platser

- ***E53 Place***
- ***E44 Place appellation***
- ***E48 Place name***

[17] http://www.cidoc-crm.org/Version/version-6.2.1

4.4 Utvärdering av mappningsmöjligheter mot den initiala delmängden

Utvärderingen nedan bygger på diskussioner vid Workshop 3. Utöver detta skickade Nordiska museet och K-samsök även in det underlag som hade efterfrågats inför workshoppen, med konkretisering av vad CIDOC CRM-klasserna kan mappas mot hos dem, samt eventuella kommentarer på dessa mappningar.

Kungliga biblioteket: För de flesta identifierade klasser funkar det ganska bra och en möjlig väg framåt är att lägga till dem där vi exponerar länkade data. Det är svårast att mappa mot klasser som rör tid, eftersom sättet som biblioteksstandarder och CIDOC CRM representerar detta skiljer sig åt (se kommentar i Avsnitt 4.3). Libris mappas mot Bibframe.

Musikverket: Musikverkets biblioteksmaterial går in i Libris, därför följer de den mappning KB gör. Utöver detta finns flera databaser med olika samlingar för arkiv och museum, registrerade på olika sätt. Musikverket har viss möjlighet att anpassa sig efter hur andra museer och arkiv gör, men har svårt att stödja standarder som deras system inte hanterar. Eftersom de inte har egna utvecklare innebär ytterligare en standard (och med den förknippad mappning för att möjliggöra export) ett stort merarbete.

Nordiska museet: Primus, som Nordiska museet använder, har ett schema i systemets databas, och ett annat schema för det som publiceras utåt i Digitalt museum. För denna mappning användes det interna systemets schema. Strukturen för CIDOC CRM är bekant då de också sorterar utifrån person-plats-tid samt relationer däremellan. Vad gäller Aktör använder Nordiska museet en namnfunktion, där namnen relateras till händelser eller andra objekt med olika anknytningskoder, t ex tillverkare, brukare, tillverkare av original, beställare, producent. Dessa var svårare att koppla till CIDOC CRM-klasser rakt av.

RAÄ (K-Samsök): Det finns begrepp i K-Samsök som går att mappa direkt mot CIDOC CRM. En del av det som är klasser i CIDOC CRM är attribut i K-Samsök. För vissa begrepp har K-Samsök mer nyanser än CIDOC CRM, och för en del andra begrepp är det tvärt om. K-Samsök är intresserade av CIDOC CRM oberoende av detta projekt och kommer i första hand att försöka följa CIDOC CRM rakt av. Om detta inte går kommer en mappning att skapas.

Riksarkivet: Analysen är gjord utifrån beskrivning av datainnehållet i Arkis (Riksarkivets arkivinformationssystem) där olika typer av innehåll beskrivs systematiskt. ”Aktör” är ganska enkelt att arbeta med. Tidsperioder anges och har olika dateringstyper för arkivmaterial. ”Tid” – här täcker arkiven inte in så mycket. ”Plats” finns täckning för, information förses ofta med geografisk plats och koordinater. ”Objekt” är också svårt på grund av arkivens strikta hierarkiska beskrivning. Normalt beskrivs inte ett enskilt objekt, beskrivningarna är på en mer aggregerad nivå. Det finns dock utrymme att beskriva enskilda objekt, även om det normalt inte görs, vilket gör detta till ett praktiskt problem.

De problem med att mappa mot tid som KB tar upp ovan är sannolikt generella för flera institutioner, det syns t ex även i mappningsunderlaget som K-samsök delat. Detta beror på det händelseorienterade perspektivet i CIDOC CRM, där metadata kring tid och personer inte

kopplas direkt till E70 Thing, utan detta görs via olika typer av händelser (E5 Event). Det som listas som under rubriken tid i förra avsnittet är klasser som ärver från Event. Dessa händelser kan ses ligga närmre verkliga skeenden och är ofta underförstådda i metadata. Att explicitgöra dessa kan skapa tydlighet samtidigt som det kräver resurser.

4.5 Mappningsverktyg

Det finns ett antal verktyg för att mappa befintliga format mot CIDOC CRM. Två exempel är 3M (Mapping Memory Manager)[18] och Karma[19].

3M är ett verktyg för att mappa mellan olika datascheman. Verktyget är öppen källkod och man kan titta på andra användares mappningar (liksom de har tillgång till det man själv skapar). Även om verktyget är mer öppet vad gäller datascheman så är det vanligaste syftet att mappa mot CIDOC CRM.

Det är viktigt att mappningar görs av personer som förstår innehållet i de data som mappas, snarare än teknikförståelse, även om det är att föredra att personer med båda kompetenser är delaktiga. Ofta framgår inte den explicita innebörden i databaser utan förtydligas genom mjukvara, gränssnitt, utbildning, manualer och expertkunskap.

Resultatet av att använda 3M är mappningar beskrivna på ett för ändamålet utvecklat XML-format. Dessa mappningar kan sedan användas av X3ML[20] för att generera RDF baserat på mappningen och underliggande data.

I projektets initiala formulering ingick mer ingående tester av 3M. Vid Workshop 2, där tester av 3M var planerade, hade mindre än hälften av deltagarna förberett material i enlighet med utskickade instruktioner. Workshopen fokuserade istället på en vidare utvärdering av CIDOC CRM och initiala mappningar mellan en mindre delmängd av vokabulären mot deltagande institutioners vokabulär. Detta valdes även pga. att vi i detta första skede inte är intresserade av att göra en större mappning av dataset mot CIDOC CRM, utan snarare att välja ut en mindre del och identifiera var hos de deltagande institutionerna dessa återfinns.

Det finns fler verktyg för att mappa mot CIDOC CRM. Om mappning med verktyg blir relevant i framtiden bör även verktyget Karma[21] undersökas. Detta verktyg användes av ett av projekten som beskrivs nedan i punkt 1 i Avsnitt 5.

[18] Vertyget 3M nås via http://139.91.183.3/3M/
[19] http://usc-isi-i2.github.io/karma/
[20] https://github.com/delving/x3ml
[21] http://karma.isi.edu

5. Vägar framåt med CIDOC CRM

Detta pilotprojekt innebär ett steg i att undersöka om CIDOC CRM och standarder kopplade till den kan användas för att skapa interoperabilitet mellan arkiv, bibliotek och museer. Trots möjligheterna med standarden finns det, som beskrivs i Avsnitt 4.2, ett antal hinder kopplat till dess användande. Många av dessa har att göra med att standarden i första hand är framtagen för museer.

De praktiska förutsättningarna för CIDOC CRM behöver alltså utredas vidare för att konstatera om det är en realistisk väg för att nå interoperabilitet inom kulturarvsområdet eller inte. Nedan föreslås ett antal möjliga, av varandra oberoende, vägar för vidare utredning av CIDOC CRM.

1. Sammanställ erfarenheter från andras användning av CIDOC CRM. Det finns fortfarande relativt få sammanställningar av erfarenheter av att använda CIDOC CRM i mer omfattande sammanhang. Det har nyligen publicerats erfarenheter från större projekt, bl.a. för att länka data kring första världskriget[22] och kring amerikansk konst[23]. Den förstnämnda beskriver användandet av CIDOC CRM och länkade data för att länka samman flera dataset kring första världskriget och delar dels erfarenheter av att använda CIDOC CRM, dels en prototyp som byggts på datasetet. Det andra projektet presenterar erfarenheter från att skapa länkade data från fjorton amerikanska konstmuseer, genom vilket ett antal verktyg skapats och tillgängliggjorts. Även en webbapplikation[24] för att utforska den sammanlänkade konsten utvecklades. Några av erfarenheterna från de två projekten tas upp i denna rapport, men t ex är beskrivningar kring arbetsflöden samt länkning av instanser relevanta att undersöka vidare.

2. Samla en arbetsgrupp (ca tre till fyra personer), från ett urval av institutioner samt Digisam för fortbildning inom CIDOC CRM. Det kan bl.a. handla om att delta i CIDOC CRMs sommarskola och konferens tillsammans. För att använda standarden behövs mer kunskaper i hur den bäst appliceras. Deltagandet i grupp ger möjlighet att föra diskussioner om standardens möjligheter och begränsningar på en högre nivå. Deltagande på plats ger möjlighet att diskutera erfarenheter med institutioner från andra delar av världen som kanske står inför liknande avväganden.

3. Genomför en liten, väl avgränsad, utvecklingspilot för att undersöka praktiska aspekter av att använda CIDOC CRM som semantiskt klister för data från två till fyra aktörer där arkiv, bibliotek och museer finns med. Lägg in data i gemensam triplestore (en databas med tripplar, dvs den modell som används för länkade data) snarare än att kombinera flera SPARQL endpoints (i ett första läge är det viktigare att titta på specifika möjligheter och utmaningar

[22] Mäkelä, E., Törnroos, J., Lindquist, T. et al. (2017) WW1LOD - An application of CIDOC CRM to World War 1 Linked Data. Int J Digit Libr. 18: 333. https://doi.org/10.1007/s00799-016-0186-2
[23] Knobock et al. (2017) Lessons Learned in Building Linked Data for the American Art Collaborative. In Proc. International Semantic Web Conference 2017. https://iswc2017.ai.wu.ac.at/wp-content/uploads/papers/MainProceedings/382.pdf
[24] http://browse.americanartcollaborative.org

med att använda CIDOC CRM över flera domäner/institutioner, snarare än hur data tillgängliggörs för andra). En utvecklingspilot skulle kunna innehålla följande steg:

a. Överväg att arbeta nära en specifik användargrupp (t ex en forskargrupp, en skolklass, släktforskarförening, etc.) för att identifiera ett väl avgränsat område samt hur de skulle vilja utforska dess data. Identifiera "competency questions" – exempel på frågor de skulle vilja ställa till det underliggande materialet.
b. Om man väljer att arbeta nära en användargrupp, välj ett område där det finns hyfsad mängd digitaliserat material, så att användarna av piloten även kan komma till t ex ett digitalt dokument, bild eller bild på ett föremål, snarare än en metadatapost.
c. Identifiera en mindre mängd klasser och egenskaper ur CIDOC CRM (och dess släktingar, som $FRBR_{OO}$) som är relevanta för domänen. Utgå ifrån de som har identifierats i denna rapport (se Avsnitt 4.3).
d. Identifiera gemensamma principer för identifierare för auktoriteter. Utöver detta kan man även göra mappningar, se steg F.
e. Exportera relevanta data i RDF-format från respektive institution som har att göra med det valda området enligt de klasser och egenskaper som valts ut. För vissa deltagare kan det finnas inbyggt systemstöd för detta. För andra bör möjligheten att använda 3M och X3ML alternativt Karma för detta undersökas (se Avsnitt 4.5).
f. Samla exporterade data i gemensam triplestore. Sannolikt kommer det att finnas behov av att skapa mappningar (med hjälp av owl:sameAs och SKOS) mellan resurser som är relaterade eller identiska. För detta finns olika verktyg, t ex det mer generella Silk[25] och det som beskrivs i Knobock et al. (2017) (se förslag 1).
g. Använd tillgängliga faceted browsers och query-miljöer för att utforska data tillsammans med användare. Ett annat alternativ är att bygga enklare webbgränssnitt specifikt för syftet.
h. Utvärdera användandet av CIDOC CRM, kvalitet på sökresultat och användarupplevelser av att arbeta med denna typ av struktur och material. Undersök särskilt om CIDOC CRM klasserna var ändamålsenliga samt om det fanns skillnader i hur olika institutioner tolkade dem.

4. Undersök i vilken mån museer (samt K-samsök) kan använda CIDOC CRM för att skapa interoperabilitet mellan sina samlingar. Några museer använder redan i nuläget system som kan mappas mot CIDOC CRM. K-samsök undersöker möjligheterna att använda CIDOC CRM. Om delar av CIDOC CRM ska användas för interoperabilitet med andra är det värdefullt att i första hand undersöka interoperabilitetsmöjligheterna för de som har lättast att använda modellen.

5. För de institutioner som inte redan tillgängliggör länkade data, överväg att använda KulturNav, som stödjer CIDOC CRM. Även om KulturNav är kopplat till Primus så är det möjligt att använda fritt även för de som inte använder Primus. Material kan sedan hämtas via deras API. Fördelarna är att mindre arbete behöver läggas på utvecklingssidan och att materialet blir tillgängligt i ett vidare sammanhang. En stor nackdel är att institutioner inte

[25] http://silkframework.org

själva kan förvalta identifikatorerna, då de skapas utifrån domännamnet, som ägs av KulturNav.

6. Kungliga biblioteket skulle, som de föreslagit vid workshopparna, kunna lyfta in delar av CIDOC CRM-vokabulären i sina länkade data.

7. Om det upplevs som relevant för biblioteksområdet kan befintliga mappningar mellan CIDOC CRM och Bibrame respektive IFLA LRM undersökas.

8. Om det upplevs som relevant kan Riksarkivet undersöka om den tillgängliggjorda mappningen av EAD och CIDOC CRM[26] är användbar. Riksarkivets representanter har även föreslagit följande startpunkter vid workshopparna:

- Börja med kontextuella data.
- Kan man koppla teman och därmed skapa broar? Det behöver inte handla om specifika dokument, CIDOC CRM har åtta fysiska nivåer och man skulle kunna tänka sig ett antal arkivdokument som en hel grupp.
- Sondera valde att representera alla nivåer i arkiv som separata objekt, är det relevant här?

[26] http://old.CIDOC CRM.org/workshops/finland_helsinki_20102801/N13_28Jan2010%20Christos%20Papatheodorou.pdf

6. Avväganden i valet mellan REST API och/eller SPARQL endpoint

För att tillgängliggöra data finns flera olika metoder, t ex REST API och SPARQL endpoints (se Avsnitt 2.3 för en kort introduktion till dessa metoder). Som vid alla teknikval finns fördelar och nackdelar.

Fördelar med API är att det är ett etablerat sätt att tillgängliggöra data. En nackdel med ett gemensamt API är att det riskerar att bli en tjänst baserad på minsta gemensamma nämnare. API är ofta utvecklade för specifika syften och att man måste sätta sig in i varje nytt API som man vill använda.

En fördel med SPARQL är att det är mer generellt än ett API, då man jobbar direkt mot data snarare än ett kontrakt. En annan fördel med SPARQL är möjligheten att ställa frågor över olika system, samt att man inte behöver ladda ner data från flera ställen. Nackdelar är att SPARQL är mindre känt och långsammare. Det är även svårare att få statistik i hur det anropas.

Av de institutioner som är med i detta projekt är det bara KB som tillhandahåller en SPARQL endpoint[27]. K-samsök har ett API men ingen SPARQL endpoint, något som eventuellt kommer det att finnas i nya K-samsök.

Ett steg mot att använda länkade data är att KB lägger in CIDOC CRM-klasser i sin vokabulär och exponerar dem via sin SPARQL endpoint, vilket K-samsök kan använda. Detta är ingen djup ansats, men kan ses som en stor förbättring över nuläget. En annan väg som kan undersökas är om Wikidata[28] är en möjlig plattform, både att använda som gemensam nämnare och som en möjlig nod vid federering.

För att hantera att olika utvecklare har olika behov är en möjlighet att tillhandahålla både REST API och SPARQL endpoint. REST API:er ger kortare svarstider men enbart på en smalare och på förhand definierad mängd frågor, medan en SPARQL endpoint kan ge mer precisa svar på och där frågor kan ställas på ett ej på förhand specificerat vis, men med en längre svarstid. Det finns även lösningar för att tillhandahålla länkade data via REST API (t ex Elda[29]).

Oavsett vilket alternativ som väljs så behövs en plan för hur de ska börja användas då båda är till för utvecklare snarare än slutanvändare. Marknadsföring av tillgängliggjorda data kan t ex göras via portaler, hackathon och samarbeten med andra organisationer, t ex högskolor. En idé är att anordna ett nytt kulturarvshack som visar hur man kan använda sig av ihopkopplad information. Hacket behöver i så fall använda förvaltade data, så att dataförsörjningen som sådan inte är en del av hacket.

---

[27] http://libris.kb.se/sparql
[28] http://www.wikidata.org
[29] http://epimorphics.github.io/elda/current/index.html

7. Diskussion och slutsatser

7.1 Sammanfattning och diskussion av genomförande

Denna rapport sammanfattar resultatet av ett pilotprojekt för att framförallt undersöka möjligheterna att använda standarden CIDOC CRM för att dela semantiskt interoperabla data mellan arkiv, bibliotek och museer. Även tekniker rörande metoder för att tillgängliggöra data har ingått i projektet, men i mindre omfattning. Inom ramen för projektet genomfördes tre workshops med deltagare från Digisam, K-samsök, Kungliga biblioteket, Musikverket, Nordiska museet, Riksantikvarieämbetet samt Riksarkivet.

Vad som initialt kunde förstås som en förutsättningslös diskussion om en föreslagen standards möjligheter att mappas mot olika domäners metadata landade lika mycket i diskussioner om för vilken domän standarden tagits fram, skillnader mellan domäners syn på sina underliggande data snarare än likheter, samt de praktiska förutsättningarna för att använda den. Skillnader i synsätt på hur data förvaltas vid deltagande institutioner, deltagarnas roller och resursmässiga överväganden har också konsekvenser för synen på CIDOC CRM. Författarens bild är att flertalet deltagare inte hade möjlighet att lägga den tid utanför workshopparna som anges i projektplanen. Detta återspeglas delvis i resultaten ovan: rapporten presenterar i större grad de problem som finns med att använda CIDOC CRM och i mindre grad mappningserfarenheter. I tidigare utkast av projektplanen fanns en första workshop med för att introducera projektet, vilket nog hade behövts för att diskutera bl.a. vision, syfte, Digisams roll, möjliga projektresultat samt deras konsekvenser. Denna ströks för att ge deltagarna mer tid åt de praktiska delarna.

De huvudsakliga projektresultaten är:

- Standarden CIDOC CRM upplevs som en komplicerad standard i ett landskap där det redan finns många standarder. Ur perspektivet museer upplevs den som intressant, för arkiv och bibliotek fattas centrala perspektiv. Se Avsnitt 4.1.
- Det är möjligt att använda endast en delmängd av CIDOC CRM, där projektdeltagarna identifierat koncept relaterade till Aktör, Tid, Objekt och Plats som mest relevanta. Utifrån dessa har ett subset av klasser ur CIDOC CRM identifierats och mappningsmöjligheterna för de olika institutionerna har identifierats. Se avsnitt 4.2-4.4.
- Givet CIDOC CRMs särställning är det trots allt relevant att vidare undersöka möjligheterna att använda den. Några möjliga vägar framåt är att genomföra en utvecklingspilot och att sammanfatta lärdomar från några nyare internationella projektresultat. Se fler förslag i Avsnitt 5.

7.2 Diskussion

Om man tar ett steg tillbaka från CIDOC CRM och ser till det mer generella målet att göra det möjligt att koppla ihop kulturarvsdata som finns utspritt över flera institutioner, kan ett antal frågor ställas. Vad behöver lösas för att detta ska ske, och för vem sker det? Är CIDOC CRM rätt väg att gå, och löser det rätt problem?

Inom den semantiska webben finns sedan länge två olika läger med avseende på hur pass logiskt och filosofiskt avancerade ontologier som bör användas, dessa brukar kallas för "scruffys" och "neats". Neats tycker att det är viktigt att den ontologi som används är filosofiskt välgrundad och med avancerad logisk uppbyggnad, medan scruffys tycker att det är väsentligast att man utgår ifrån en modell som funkar bra nog och blir använd.

CIDOC CRM är en "neat" ontologi. Dess omfattning, logiska koherens och genomtänkta modell kan ses både som en fördel och en nackdel. Ska den användas kommer det att behövas hyfsat omfattande fortbildning som endast går att få genom ett fåtal instanser. Om man ser till ett återanvändningsperspektiv är detta en ännu större nackdel: utvecklare och forskare som behöver arbeta nära modellen kommer också behöva lägga ner tid på att förstå den, alternativt använda den utan att förstå helheten. En annan större nackdel är att workshopdeltagare från arkiv och bibliotek inte finner att CIDOC CRM beskriver deras material på ett adekvat sätt.

Vid workshopparna togs ett förslag fram på en liten delmängd av CIDOC CRM som kan användas som utgångspunkt i dess användande. Det är en liten delmängd klasser utifrån koncepten aktör, tid, objekt och plats, beskrivna i Avsnitt 4.3. Utifrån detta föreslås ett antal vägar framåt i Avsnitt 5, bl.a. en utvecklingspilot för att testa värdet av denna delmängd i praktiken.

Det är alltså en hyfsat liten delmängd av CIDOC CRM som har valts ut. Avsnitt 4.3 listar 19 av 94 klasser, men om man ser till utvärderingen är inte alla relevanta. Fördelar med att använda denna delmängd är att det är en internationell standard och att det är möjligt att koppla på fler delar i framtiden.

Å andra sidan, givet att det i första hand bara är en delmängd av CIDOC CRM som är relevant som semantiskt klister, är en annan väg att gå att använda en mindre avancerad ontologi. Ytterligare en väg är att skapa en tjänst där de huvudsakliga standarderna från arkiv, bibliotek och museers mappas mot varandra på en nivå som är relevant för det tänkta användningsbehovet.

För att skapa ett återanvändbart digitalt kulturarv finns det två andra aspekter, utöver vilken ontologi man använder för klasser och relationer, som kanske är ännu viktigare: Sammanlänkning av identiteter på instansnivå och att identifiera prioriterade användarfall.

Sammanlänkning av identiteter på instansnivå (t ex för specifika personer, platser, verk, etc.) har inte varit en del i detta projekt, men har förts fram som en viktig del i att skapa interoperabilitet mellan institutionerna (enligt några deltagare viktigare än det som diskuterats inom projektet). Dessa behöver antingen matchas genom speciella verktyg eller genom att gemensamma ämnesord, taxonomier och auktoritetsposter används. Nyligen publicerade projektresultat (se första förslaget i Avsnitt 5) om att använda CIDOC CRM för att föra samman dataset kring första världskriget, konstaterar att det är betydligt mer resurskrävande att nå denna nivå, och att skillnader på modellnivå är relativt små i jämförelse. Detta pekar

mot att jämfört med CIDOC CRM kan auktoritetsposter och mappningar kan vara viktigare, mer resurskrävande och samtidigt lättare att enas kring. Relativt tidigt i ett sådant arbete behöver frågan om förvaltning hanteras.

Att identifiera prioriterade användarfall är viktigt för att vägleda beslut som rör semantisk interoperabilitet. Upprepade gånger konstaterade workshopdeltagarna att det var svårt att komma framåt eftersom användarfallet inte var klart. För att säga om CIDOC CRM eller någon annan ontologi bör användas behövs mer kunskap om vad som ska uppnås för vem genom det semantiska klistret. Detta skulle kunna ta formen av ett antal användarfall som har identifierats som prioriterade.

Bilaga 1 – Kartläggning av deltagande institutioners format och standarder samt introduktion till CIDOC CRM (Delrapport 1)

Marie Gustafsson Friberger, projektledare

Denna rapport är resultatet av den första av tre workshops kopplat till projektet. Workshopens mål var att undersöka användning av CIDOC CRM som gemensam ontologi för kulturarvsinformation som en ”översättning” av olika datamodeller som tillämpas inom kulturarvet. Workshopen kartlade även deltagande institutioners format och standarder. Projektet som helhet presenteras och analyseras i en avslutande rapport.

Workshop 1, 13 juni 2017 kl. 13-16, Riksantikvarieämbetet, Stockholm
Närvarande: Martin Malmsten, KB, Stina Degerstedt, KB, Sandra Åberg, Nordiska museet, Pär Johansson, Musikverket, Birger Stensköld, RA, Mårten Johansson, RA, Kerstin Jonsson, RAÄ, Marcus Smith, RAÄ, Henrik Summanen, Digisam, Moa Ranung, Digisam, Marie Gustafsson Friberger, konsult

1. Introduktion till projektet

Projektet syftar till att vidareutveckla modellen för hur digital kulturarvsinformation skulle kunna tillgängliggöras på ett effektivare sätt genom semantiska webbens möjligheter. Pilotprojektet undersöker olika sätt att strukturera och harmonisera kulturarvsinformation enligt internationella standarder och maskinläsbara format, samt testa befintliga tjänster för detta.

I samband med presentationen av projektet diskuterades dess syfte och upplägg. Varför är CIDOC CRM i fokus snarare än en mer öppen ansats? Tidigare projekt har identifierat ett behov av att undersöka CIDOC CRM närmre, för undersöka om standarden är relevant för svenska arkiv och museum, i vilken mån man kan använda en modell samt hur genomförbar den är.

2. Introduktion till CIDOC CRM

CIDOC Conceptual Reference Model (CRM) är en internationell standard för kulturarvsdata. Bakgrunden till standarden är att olika tjänster ofta använder olika standarder för metadata, vilket bland annat medför att det saknas tjänster där användare på ett sömlöst vis kan söka relaterat material som kommer från flera källor. De tjänster som finns bygger ofta på tidskrävande manuell mappning mellan källorna metadatabeskrivningar. CIDOC CRM kan ses som ett ”semantiskt klister” för att mediera mellan olika källor till kulturarvsinformation, som publiceras av t ex museum, bibliotek och arkiv.

Standarden innehåller definitioner och formell struktur för att beskriva implicita och explicita koncept och relationer vid kulturarvsdokumentation. Den innehåller grundläggande entiteter (62 stycken) och relationer (149 stycken), vilka är ordnade i hierarkier vilket kan användas vid sökningar för att hitta nya samband. Man behöver inte använda alla dessa entiteter och relationer, utan en delmängd kan väljas ut. Om begrepp för en viss domän saknas finns möjlighet att definiera dessa inom ramen för CIDOC CRM. Viktigt att notera är att CIDOC CRM är en händelse-centrerad modell: personer och objekt relateras inte till tid, bara händelser länkas till tid. I presentationen ingår ett antal exempel för att illustrera värdet av CIDOC CRM och hur en beskrivning kan se ut.

CIDOC CRM Primer är en bra startpunkt för standarden.

3. Diskussion kring CIDOC CRM

Harmonisering av metadata är en informationsfråga snarare än en teknikfråga – därmed en verksamhetsfråga. Deltagande institutioner har olika behov, t ex beroende på storlek och domän. Mer detaljer kring detta i nästa avsnitt.

CIDOC CRM är utvecklat från museernas perspektiv men gör anspråk på att vara applicerbar även på bibliotek och arkiv. Det är oklart i vilken mån bibliotek och arkiv framgångsrikt använt CIDOC CRM eller relaterade standarder. CIDOC CRM kan upplevas som komplex och oöversiktlig.

Frågor framåt:

- Vad är det mest basala som behövs för att kunna koppla ihop information mellan institutioner? Ett förslag: ID - aktör - aktivitet. Även: Platser, händelser, personer och typer av objekt.
- I vilken mån måste beskrivningar göras på domännivå? Det vill säga hur mycket uttrycker CIDOC CRM i sig?
- Hur undvika minsta gemensamma nämnare-problemet samt risken att alla mappar mot CIDOC CRM men att resultatet ändå inte hänger ihop?
- Hur kan man bäst ta hänsyn till användarens (där det finns flera grupper) behov när man väljer hur data mappas?

4. Kartläggning av standarder vid medverkande institutioner

Nedan är den kartläggning som gjordes av representanter för de medverkande institutionerna under workshopen. För fylligare beskrivningar av flera av dessa standarder, se även digisam.se/om-standarder/ samt Digisams rapport från 2014, Digital informationshantering och infrastruktur för kulturarvet – Underlag för fortsatt arbete.

Kungliga biblioteket

FRBR (Functional Requirements for Bibliographic Records) – FRBR delar upp den bibliografiska informationen i olika nivåer: verk, uttryck, manifestation och exemplar. *Kommentar: hög nivå, upplevs som svårimplementerad.*
RDA (Resource Description and Access) – Anvisningar och riktlinjer för katalogisering och auktoritetsarbete. RDA använder sig av FRBR-modellen för att strukturera information.
Bibframe (Bibliographic Framework Initiative) – Vokabulär och beskrivande modell, kopplat till länkade data, utvecklas av Library of Congress.
FRSAD (Functional Requirements for Subject Authority Data)
FRBR-LRM (FRBR Library Reference Model) – Kopplar samman bland annat FRBR Och FRSAD. Kompatibel med CIDOC CRM(?)
MARC (föråldrat men kommer finnas kvar)
KB använder också många andra vokabulärer, de är för bibliografisk beskrivning.
I övrigt används
EAD (Encoded Archival Description) – används för handskrifter
METS (Metadata Encoding and Transmission Standard)
PREMIS (Preservation Metadata: Implementation Strategies)

Relaterat: FRBRoo – en objektorienterad version av FRBR, med mappning av mapping från FRBR och FRSAD mot CIDOC CRM

Relevanta kategorier: dokumentation (händelser/tidpunkt kopplat till detta), tidsperioder, ämnesord

Riksantikvarieämbetet

MIDAS (Manual and Data Standard for Monument Inventories) – Typning av informationen, händelser, dokumentation. Tillkom genom DAP (Digital Arkeologisk Process).
K-Samsök – Aggregerar information från register om till exempel byggnader, samt data vi skördar från andra institutioner - främst föremål, samlingar, bilder. *Kommentar: Kan ge sken av att vara masterdata, men är det inte. Tittar på CIDOC CRM för K-Samsök.*
Evighetsrunor – runinskrifter som informationsobjekt

Riksantikvarieämbetet har interna diskussioner kring hantering av periodindelningar och att gå från referenstypologier mot auktoritetslistor.

Relevanta kategorier: lämningar, byggnader, personer, föremål, samlingar, bilder.

Nordiska museet

CIDOC CRM genom Primus
Spektrum – Standardiserade processer kring samlingsförvaltning.
Auktoriteter – Intern lösning inspirerad av **CIDOC CRM** och **EDM**. Mappas mot **SKOS**, **FOAF**, **CIDOC CRM**.
Outline – Äldre klassifikationssystem. Fortfarande i bruk men är inte aktuell för dagens samlingar utan uppdatering.
SIS-CEN terminologier
Har Primus och därigenom KulturNav. KulturNav har en egen intern datamodell som inspirerats av modeller i CIDOC CRM och Europeana Data Model. Många museer vill kunna plocka upp information från Getty vocabularies.

Riksarkivet

Tredelad informationsbeskrivningsstandard. Var och en av dessa tre delar kan stå på egna ben och relateras till varandra.
ISAD(G) (General International Standard Archival Description) – Beskriver arkivmaterialet, det vill säga själva handlingarna. Det är en logisk beskrivning av uppbyggnaden, den är hierarkisk och allt är beroende av varandra.
ISAAR(CPF) (International Standard Archival Authority Record For Corporate Bodies, Persons and Families) – Beskrivning av auktoriteter, det vill säga personer och organisationer, som har koppling till arkiven, arkivbildare.
ISDF (International Standard for Describing Functions) – Beskrivning av verksamheten som arkivbildarna har ägnat sig åt.

Kopplat till dessa finns ett antal XML-baserade utbytesformat: **EAD** för ISADG, **EAC** för ISAAR-CPF, samt **EAC-F** för ISDF (under utveckling). EAD och EAC har även lokala tillämpningar. Ansluter sig även till METS och Premise. Har också några mindre kända standarder såsom de för pergamentsomslagen.

Kommentarer:

- RA och KB har haft utbyten kring att försöka jämka ihop kontextuella data (ej själva samlingarna).
- I kulturplattform Västernorrland var de områden som var mest intressanta för mappning kontextuella data, dvs. data om platser/händelser/orter/personer osv. som är gemensamma för hela kulturminnessektorn.
- Svårt att översätta arkivbeskrivningar till en objektsbaserad konceptmodell.

Relaterat till CIDOC CRM

- Arkiv har hierarkisk modell för sina samlingar, det vill säga har med arkivbildare, något som museum och bibliotek inte efterfrågar. Därför saknas arkivbildare ofta när gemensamma standarder tas fram.
- Arkivsektorn har sedan 90-talet standarder för arkivdata (multilevel descriptions)
  Finns det mappningar av dessa till CIDOC CRM?
- Händelser i Riksarkivets material är arkivrelaterade. Andra händelser kan länkas från annat material, på så vis skapas kopplingar till händelser som ej är arkivrelaterade.

Musikverket

Musikverket har museum, arkiv och bibliotek. För museets material används en egenutvecklad databas. För arkivmaterialet används arkivdatabasen CALM (oklart vilken standard/vokabulär). För biblioteksmaterialet används (se listan för KB ovan för förklaringar av standarder): **FRBR, RDA** (ersätter AACR2-KRS), **MARC21, EAD,** OAIPMH (Open Archives Initiative Protocol for Metadata Harvesting) för informationsutbyte och Dublin Core. För klassifikationer används DDK och SAB.

5. Övrigt

Applikation/demo: Värdet i att ha en demo, som visar på värdet av att bygga på CIDOC CRM. Fritextsökning funkar till viss gräns, men användaren måste göra dessa sökningar i separata tjänster.

Persistenta identifierare (PID):

- Beslut kring var PID ska vara för att vara långsiktigt är en verksamhetsfråga, snarare än en teknisk fråga.
- Organisationen och dess process gör det persistent, inte var/hur det görs.
- Fördel att använda slumpad sträng, som inte inbjuder till att ge organisationsinfo i ID.

Sekretess/PUL:

- Museer har ofta avtal kring användning av personuppgifter när de får en gåva.

Det finns ofta tydlig aktivitet och aktör när RA tar emot mer material digitalt, men sekretess och PUL etc. för levande personer gör att detta inte delas.

Bilaga 2 – Fortsatt utvärdering av CIDOC CRM och identifikation av relevanta termer (Delrapport 2)

Marie Gustafsson Friberger, projektledare

Om projektet: Projektet ska vidareutveckla modellen för hur digital kulturarvsinformation skulle kunna tillgängliggöras på ett effektivare sätt genom semantiska webbens möjligheter. Pilotprojektet undersöker olika sätt att strukturera och harmonisera kulturarvsinformation enligt internationella standarder och maskinläsbara format, samt testa befintliga tjänster för detta.

Denna rapport är resultatet av den andra workshopen inom ramen för projektet, där målet fortsatt utvärdering av CIDOC CRM, identifikation av relevanta termer ur vokabulären att utgå från samt presentation av ett mappningsverktyg. Projektet som helhet presenteras och analyseras i en avslutande rapport.

Workshop 1, 11 september 2017 kl. 13-16, Riksantikvarieämbetet, Stockholm
Närvarande: Marie Gustafsson Friberger, Henrik Summanen (Digisam), Antonio Molin (Digisam), Moa Ranung (Digisam), Marcus Smith (Riksantikvarieämbetet), Pär Johansson (Musikverket), Sandra Åberg (Nordiska museet), Stina Degerstedt (Kungliga biblioteket), Martin Malmsten (Kungliga biblioteket), Maria Carlsson (Riksantikvarieämbetet), Birger Stensköld (Riksarkivet)
Förhinder: Kerstin Jonsson (Riksantikvarieämbetet), Mårten Johansson (Riksarkivet)

1. Varför mappa mot gemensam vokabulär?

Workshopen inleddes med en diskussion om varför det är relevant att mappa mot en gemensam vokabulär. En del av detta är att skapa interoperabilitet mellan olika områden och att göra det enklare både för experter och allmänhet att röra sig mellan olika domäner. En fördel med detta arbete är att man måste explicitgöra interoperabilitet som idag "sitter i huvudet" på domänexperter. Om man ser till användandet av data så är mappningen nödvändig för att automatiskt kunna kombinera olika informationskällor och sammanföra information från flera källor (så att användaren inte behöver vända sig specifikt till var och en av institutionerna).

Jämför med ett byggvaruhus: Det finns en variation av sortiment, man kan specialisera sig på vissa varor, man kanske väljer mer aparta beståndsdelar om man är proffs för att man vill bygga något väldigt speciellt. Viktigt är att alla beståndsdelar passar ihop. Ibland går det att kombinera delar ingen trodde att man kunde kombinera. En gemensam vokabulär kan möjliggöra detta.
Man bör ha i åtanke att användaren av mappad data kan vara både människor och programvara.

2. CIDOC CRM för olika domäner

Hur väl funkar CIDOC CRM för allmän interoperabilitet, arkiv, bibliotek och museer? Diskussionen summeras i en tabell sist i avsnittet.

Arkiv
Plus: Riksarkivet har tidigare försökt att tillämpa CIDOC CRM-standarden på arkivdata. När det har lyckats bäst har det handlat om udda arkiv, t.ex. medeltida brev. När det gäller konventionella arkiv tillämpar arkiven ett hierarkiskt tänkande där man utgår från arkivbildaren som i sin tur har arkiv, med serier och subserier, och därunder akter och underakter o.s.v. Hos arkiven kommer vi sällan till att kunna registrera på dokumentnivå. Born digital-dokument innebär stora möjligheter för en maskin att ta hand om dem. Det är svårare med pergament o.s.v., då det är svårt att OCR-läsa. I forskarsalarna digitaliseras mycket på egen hand, och de företag som digitaliserar arkiv tar betalt.

Minus: Riksarkivet har svårt att skildra hierarkier. Kan man koppla teman och därmed skapa broar? Det behöver inte handla om specifika dokument, CIDOC CRM har åtta fysiska nivåer och man skulle kunna tänka sig ett antal arkivdokument som en hel grupp. När Riksarkivet har skickat information via Lido till Europeana har mottagarsidan efterlyst rikare data.

Kommentarer: Sondera valde att representera alla nivåer i arkiv som separata objekt. Digitalisering av själva innehållet i arkiven är ett mycket stort jobb.

Bibliotek
Plus: Inkluderar man FRBR fungerar i princip allt. Kungliga biblioteket försöker beskriva allt man har, men beskrivningarna sker på mycket generell nivå. Det kan vara intressant att se närmare på CIDOC CRM-vokabulären och det vore intressant att få in delar av det i en länkade data-lösning. Kungliga bibliotekets länkade data-vokabulär kan lyftas in där relevant. CIDOC CRM innebär en möjlighet att lägga till fler samband än de som används idag.
Minus: Mycket av det som bibliotek vill beskriva finns inte i CIDOC CRM. Det går förmodligen att modellera, men kommer vara krävande. FRBRoo utgår från en komplicerad modell som man behöver lära sig. Om man, enligt modellen för länkade data, skriver in att en sak är samma som en annan sak, så kanske det fungerar att inte mappa så väldigt mycket fastän man beskriver saken olika på olika håll.

Museer
Plus: På Nordiska museet är man van att arbeta i databaser som bygger på CIDOC CRM. Primus bygger på CIDOC CRM, liksom KulturNav. Musikverket/Musik- och teaterbiblioteket har delsamlingar som är strukturerade på olika sätt. Allt är inte registrerat med CIDOC CRM i tankarna.
Minus: Är man mer disparata än man tror i sina beskrivningar?

De flesta institutioner behöver hantera hur beskriva materialtyper som inte tillhör majoritetsmaterialet. T ex att arkiv har föremål, eller museer har arkiv.

Tabellen nedan summerar diskussionen.

	+	-
Allmän interoperabilitet	Lågt hängande frukt finns i de högre nivåerna av vokabulären. Här beskrivs även informationsobjekt, dvs. inte bara fysiska objekt som ofta finns främst hos museer.	Komplexitet I grunden utvecklat för museer.
Arkiv	I de experiment som gjorts, passar bra med mindre arkiv.	Passar sämre för den hierarkiska modell som används i arkivens standarder.

Bibliotek	KB: Kan lyfta in KB:s vokabulär för länkade data där relevant. Musikverket: möjligt att lägga till fler samband, t ex för noter, än vad som används idag.	Mycket som är viktigt för bibliotek beskrivs ej i CIDOC CRM. FRBRoo är ytterligare en komplicerad modell.
Museer	NM: Används i Primus. Tar upp föremål och processer kring föremål, ägarskap.	Domänen i sig är mer disparat – även om man tror att man är överens så är man kanske inte det. Komplexitet.

3. Första steg i att använda CIDOC CRM

Deltagarna enades om att ha ett användningsområde i åtanke för mappningarna: att hitta allt om en person, plats/område eller händelse. Utifrån en initial lista med termer som skulle kunna vara relevanta identifierades några huvudkategorier av koncept som "lågt hängande frukt", startpunkter i mappningsarbetet. I smågrupper identifierades sedan var i CIDOC CRM dessa återfanns, och sedan har författaren lagt till ytterligare några relevanta termer från CIDOC CRM.

Ett steg på vägen mot det tänkta användningsområdet är att de deltagande institutionerna identifierar sina motsvarigheter till de termer som kommer fram till har stort överlapp mellan institutionerna. Det viktiga är inte att försöka få in allt i CIDOC CRM-ramverket, utan att fundera över om det även finns något annat som behövs utöver CIDOC CRM-standarden för att kunna komma framåt. Risken med att använda samma stora modell är att man använder den olika.

3.1 Koncept att utgå från

Koncept	**Kommentar**
Aktör (Personer och organisationer)	Actor som begrepp är förhållandevis gemensamt över domänerna
Platser	Sällan entydigt, men bra ingång för användare. Se längre diskussion nedan.
Processer/händelser	KB saknar beskrivna processer.
Koncept (sväljer mycket)	På KB ryms mycket inom termen koncept.
Saker/objekt	
Verk	
Tid/tidsperiod (ev med koppling till plats)	En tidsperiod är ofta även geografiskt begränsad (t.ex. frihetstiden)
Knowledge organisation system (KOS)	Hierarkiska eller grafbaserade

Vidare diskussion om plats: Plats är sällan entydigt, även om plats ofta är en bra ingång för slutanvändare (som t.ex. ofta söker på sin födelseort). Vad gäller plats kan man lägga sig på både generell och detaljerad nivå, vilket är krångligt. Söker man t.ex. på en socken får man inte så många träffar, jämfört med om man söker på hela Europa. Det vore värdefullt att ta upp detta i rapporten för att illustrera skillnader. Inom kulturmiljöområdet är just plats väldigt intressant för utvecklare. Även på DigitaltMuseum är det orter som användare söker på i första hand.

Utifrån detta så väljer workshopdeltagarna att gå vidare med följande:

- Aktör
- Tid
- Saker/objekt
- Plats

Trots att plats inte är entydigt och varierar i precision (som diskuteras ovan) så är det relevant att inkludera även i ett första steg. Plats är viktigt för allmänheten, en vanlig ingång till material i alla samlingar. Kan dessutom vara en värdefull ingång för att illustrera skillnader mellan domäner.

3.2 Var återfinns identifierade termer i CIDOC CRM?

I små grupper ägnar sig deltagarna åt att hitta gemensamma nämnare över domänerna i relation till CIDOC CRM-standarden. Vad är gemensamt i de olika systemen?

För mer information om de olika entiteter som nämns ovan, se dokumentationen för senaste versionen av CIDOC CRM (i nuläget 6.2.1).

Aktör
E39 Actor och dess subklasser ***E21 Person*** och ***E74 Group*** (samt dess subklass ***E40 LegalBody***)

De relationer som finns listade för E39 upplevs som relevanta.

Tid
E4 Period
E12 Production
E49 Time appellation
E63 Beginning of existence
E65 Creation
E66 Formation
E67 Birth
E68 Dissolution
E69 Death

Det är svårt att hitta en kod för när ett föremål har använts (t.ex. en viss period under vilken ett objekt har använts). CIDOC CRM är händelsebaserat, vilket inte är hur t ex KB brukar beskriva. Dock går det att mappa saker som ett event. ”E65 (creation)” skulle t.ex. kunna peka på att en författare säger att något är en bok. ”E63” skulle kunna peka på att en bok är utgiven.

Saker/objekt
E70 Thing, har flera klasser som ärver från den, där vi behöver komma fram till vad som passar bäst för oss. Men, fördelen med CIDOC CRM är att vi kan börja med E70 om vi vill. Hur identifiera relevant nivå?

E24 Physical Man-Made Thing
E73 Information Object
E78 Collection

Plats
E53 Place
E44 Place appellation
E48 Place name

5. Mapping Memory Manager (3M)

Verktyget 3M, Mapping Memory Manager, är ett open source verktyg som presenteras som ett exempel på verktyg. I verktyget kan man se andras mappningar. Det är den som har domänförståelse som i första hand bör göra mappningen. I 3M ska man kunna plocka fram ett dataset i XML, och genom programmet 3M Engine skapa RDF-filer som använder URI:er man specificerat i mappningen.

Ett annat verktyg för att mappa mot CIDOC CRM är Karma. Detta kräver dock att användaren skriver Python-kod i vissa delar av gränssnittet.

Stellar är ett verktyg för att mappa CSV-data mot framför allt CIDOC CRM. Verktyget genererar RDF XML.

Att utvärdera 3M var ett av projektets originalmål. Dock har ingen workshop fördjupat sig i detta, utan fokus har varit på vidare utvärdering av CIDOC CRM och initiala mappningar mellan en mindre delmängd av vokabulären mot deltagande institutioners vokabulär. Detta eftersom vi i detta första skede inte är intresserade av att göra en större mappning av dataset mot CIDOC CRM, utan snarare att välja ut en mindre del och identifiera var hos de deltagande institutionerna dessa återfinns.

Bilaga 3 – Resultat av initial mappning och metoder för att tillgängliggöra data (Delrapport 3)

Marie Gustafsson Friberger, projektledare

Om projektet: Projektet ska vidareutveckla modellen för hur digital kulturarvsinformation skulle kunna tillgängliggöras på ett effektivare sätt genom semantiska webbens möjligheter. Pilotprojektet undersöker olika sätt att strukturera och harmonisera kulturarvsinformation enligt internationella standarder och maskinläsbara format, samt testa befintliga tjänster för detta.

Denna rapport är resultatet av den tredje och avslutande workshopen inom ramen för projektet, där målet var att gå igenom hur deltagande institutionerna mappat sina vokabulär mot de termer som identifierades vid förra workshopen, att diskutera hur gemensamma data bäst kan tillgängliggöras samt hur man möjliggör att användbara tjänster byggs på dessa. Projektet som helhet presenteras och analyseras i en avslutande rapport.

Workshop 1, 23 oktober 2017 kl. 13-16, Riksantikvarieämbetet, Stockholm
Närvarande: Marie Gustafsson Friberger (konsult), Henrik Summanen (Digisam), Moa Ranung (Digisam), Maria Carlsson (Riksantikvarieämbetet) Marcus Smith (Riksantikvarieämbetet), David Haskiya (Riksantikvarieämbetet), Pär Johansson (Musikverket), Sandra Åberg (Nordiska museet), Martin Malmsten (Kungliga biblioteket), Mårten Johansson (Riksarkivet)
Förhinder: Stina Degerstedt (Kungliga biblioteket), Birger Stensköld (Riksarkivet), Kerstin Jonsson (Riksantikvarieämbetet)

1. Mappning mot CIDOC CRM för ett urval av termer

Vid Workshop 2 identifierades ett antal termer ur CIDOC CRM som särskilt intressanta. I ett första steg valde vi att inte undersöka tillhörande relationer. Dessa termer beskrivs i Rapport 2 och har att göra med Aktör, Tid, Objekt och Plats. Inför Workshop 3 hade alla deltagare fått underlag för att undersöka för sina representationer.

KB: För de flesta identifierade termer funkar det ganska bra. Det var svårast att mappa mot ”tid”, även om KB beskriver saker som sker i en viss tid. Libris mappas mot Bibframe. Oklart om Bibframe är mappat mot CIDOC CRM.

Musikverket: Musikverkets biblioteksmaterial går in i Libris, därför följer de den mappning KB gör. Finns även flera databaser med olika samlingar för arkiv och museum, registrerade på olika sätt, där Musikverket antingen kan anpassa sig efter andra museer och arkiv, men där det i vissa fall kan vara relevant att göra egna mappningar.

Nordiska museet: Primus, som Nordiska museet använder, har ett schema i systemets databas, och ett annat schema för det som publiceras utåt i Digitalt museum. För denna mappning användes det interna systemets schema. Strukturen för CIDOC CRM är bekant då de också sorterar utifrån person-plats-tid. Vad gäller Aktör definieras namn med olika typer: person, organisation, namngivet objekt eller händelse m.m. Namnen relateras till händelser eller andra objekt med olika anknytningskoder, t.ex. tillverkare, brukare, tillverkare av original, beställare, producent. Detta kan mappas, men alla namntyper är t.ex. inte aktörer.

RAÄ (K-samsök): Det finns begrepp i K-samsök som går att mappa direkt mot CIDOC CRM. En del av det som är termer i CIDOC CRM är attribut i K-Samsök. För vissa begrepp har K-samsök mer nyanser än CIDOC CRM, och för en del andra begrepp är det tvärt om. K-samsök är intresserade av CIDOC CRM oberoende av detta projekt och kommer i första hand att försöka följa CIDOC CRM rakt av. Om detta inte går kommer en mappning att skapas.

Riksarkivet: Analysen är gjord utifrån beskrivning av datainnehållet i Arkis (Riksarkivets arkivinformationssystem) där olika typer av innehåll beskrivs systematiskt. ”Aktör” är ganska enkelt att arbeta med. Tidsperioder anges och har olika dateringstyper för arkivmaterial. ”Tid” – här täcker arkiven inte in så mycket. ”Plats” finns täckning för, information förses ofta med geografisk plats och koordinater. ”Objekt” är också svårt på grund av arkivens strikta hierarkiska beskrivning. Normalt beskrivs inte ett enskilt objekt, beskrivningarna är på en mer aggregerad nivå. Det finns dock utrymme att beskriva enskilda objekt, även om det normalt inte görs, vilket gör detta till ett praktiskt problem.

Övrigt:

- Små institutioner, t.ex. mindre museer, länsmuseer m.m., har inget stöd vad gäller mappning.
- Det viktigaste är att länkar skapas mellan de olika domänernas data.
- Viktigt att även identifiera nivåer av identifierare.
- Tillåter CIDOC CRM fiktiva ting som ämnesord (t.ex. Midgård)?
- Finns en mappning från Lido till CIDOC CRM?

2. Tjänster på aggregerade data

Den stora nyttan med att tillhandahålla data är att andra än de som producerar data kan använda den. Arbetet med att tillhandahålla data externt kan även vara allmänt kvalitetshöjande.

En politisk utmaning med tjänster som bygger på aggregerade data är att det kan vara oklart för användare vem som tillhandahållit data och i vilken mån de som bygger tjänsterna kan förklara hur data har kombinerats. Vad ska synas, källan eller tjänsten?

En öppen ansats till vem användaren innebär en utmaning vad gäller att tillhandahålla och länka data, då det bl a är svårare att identifiera startpunkter.

Utmaningar för att bygga användbara tjänster på aggregerade data:

- Rankning – För visning av relevanta resultat inom ett visst område.
- Hantering av dubbletter.
- Uppdateringar och underhåll, vilket blir än mer relevant när data från flera institutioner används.
- Identifikation av data av sämre kvalitet så att dessa ej visas för användaren.
- Presentationen av resultaten på ett användbart och lättförståeligt vis där användaren ändå kan få en helhetsbild.

3. Lärdomar från Sondera

Sondera (sondera.kb.se) är en tjänst där användaren söker i arkiv, bibliotek samt ljud och bild via en sökingång i stället för tre. Den bygger på NAD, LIBRIS och SMDB. Resultaten presenteras i tre parallella träfflistor. Träffarna är oftast metadata snarare än digitala resurser. Tjänsten lanserades 2009.

Sondera har mycket få användare. Det var däremot värdefullt som samarbetsprojekt mellan de ingående institutionerna. Arbetet med Sondera gick fort. Tanken från början att använda aktör, plats

etc. liknande det som diskuteras i detta projekt. Det fanns inte möjlighet då och man utgick istället direkt från samlingarna. Om man hade samlat materialet runt exempelvis aktör, plats etc. så hade resultatet för Sondera kanske blivit bättre.

4. SPARQL endpoint och/eller API

För att tillgängliggöra data finns flera olika metoder, t ex REST API och SPARQL endpoints. Som vid alla teknikval finns fördelar och nackdelar. För ett REST API (Application Programming Interface), hämtas data genom att anropa på förhand definierade metoder. Om man använder länkade data är en SPARQL endpoint en möjlighet, där frågespråket SPARQL kan användas för att ställa frågor mot ett eller flera dataset.

Fördelar med API är att det är ett etablerat sätt att tillgängliggöra data. En nackdel med ett gemensamt ett gemensamt API är att det riskerar att bli en tjänst baserad på minsta gemensamma nämnare. API är ofta utvecklade för specifika syften och att man måste sätta sig in i varje nytt API som man vill använda.

Fördelar med SPARQL är att det är mer generellt än ett API, då man jobbar direkt mot data snarare än ett kontrakt. En annan fördel med SPARQL är möjligheten att ställa frågor över olika system, samt att man inte behöver ladda ner data från flera ställen. Nackdelar är att SPARQL är mindre känt och långsammare. Det är också svårare att få statistik i hur det används.

Av de institutioner som är med i aggregeringspiloten är det bara KB som tillhandahåller en SPARQL endpoint (http://libris.kb.se/sparql). K-Samsök har ett API men ingen SPARQL endpoint, något som eventuellt kommer det att finnas i nya K-Samsök.

Ett steg mot att använda länkade data är att KB lägger in CIDOC CRM termer i sin vokabulär och exponerar dem via sin SPARQL endpoint, vilket K-Samsök kan använda. Detta är ingen djup ansats, men kan ses som en stor förbättring över nuläget. En annan väg som kan undersökas är om Wikidata är en möjlig platform, både att använda som gemensam nämnare och som en möjlig nod vid federering.

En väg framåt är att tillhandhålla både REST API och SPARQL endpoint för att tillfredsställa olika behov. Detta kan ses som att både en snabbare och ”dummare” infrastruktur och en federerad och långsammare men ”smartare” (mer precis) dito? Det finns även lösningar för att tillhandhålla länkade data via REST API (t ex Elda). Oavsett vilket alternativ som väljs så behövs en plan för hur de ska börja användas då båda är till för utvecklare snarare än slutanvändare.

Tillgängliggjorda data behöver marknadsföras. Detta kan t ex göras via portaler, hackathon och samarbeten med andra organisation, t ex högskolor. En idé är att anordna ett nytt kulturarvshack som visar hur man kan använda sig av hopkopplad information. Det behövs ett hack med förvaltade data, där dataförsörjningen som sådan inte är en del av hacket.

Digisam är ett samordningssekretariat för digitalisering, digitalt bevarande och digitalt tillgängliggörande av kulturarvet.

Samordningssekretariatet är en av regeringen beslutad verksamhet vid Riksantikvarieämbetet.

www.digisam.se