Pekka Henttonen

CIDOC CRM – ratkaisu heterogeenisten aineistojen yhteiskäyttöön?

CIDOC Conceptual Reference Model (CRM) on viitemalli, jonka tarkoituksena on mahdollistaa muistiorganisaatioiden tietojen yhteiskäyttöisyys ilman että olemassa olevia metatietoja tai organisaatioiden käytäntöjä joudutaan muuttamaan. Artikkeli esittelee CRM:n perusajatukset ja käy läpi siitä saatuja kokemuksia. CRM:n suurin ongelma on, että se on abstrakti, kompleksinen ja sallii erilaisia tulkinta- ja toteutusvaihtoehtoja. Näitä ongelmia voidaan välttää, jos CRM-pohjaisten palveluiden tuottajayhteisö voi keskustella viitemallin soveltamisesta. Toistaiseksi CRM:n sovellukset ovat olleet rajatumpia kuin mitä tavoite kaikkien muistiorganisaatioiden tietojen yhteiskäyttöisyydestä antaa ymmärtää.

The CIDOC Conceptual Reference Model (CRM) is object–oriented domain ontology for the interchange of rich and heterogeneous cultural heritage information from museums, libraries and archives. The article introduces to the ideas of the CRM and discusses experiences from it. The CRM offers an interesting solution to the co-operation of memory institutions. Its main weaknesses are that it is abstract, complex and allows multiple interpretations in practice. Some of the problems are avoided if the community of metadata and service providers producing CRM –based services is able to discuss together and create common understanding of how CRM is used and expressed technically. So far applications of the CRM have been more limited than the idea of information interchange between museums, libraries and archives suggests.

Address: Email Pekka.Henttonen@uta.fi

Johdanto

Arkistojen, kirjastojen ja museoiden kokoelmat, kuvailujen rakennetta ja sisältöä ohjaavat säännöt sekä kuvailujen tarkkuus ja yksityiskohtaisuus eroavat toisistaan. Siitä huolimatta tarvitaan palveluja, joissa heterogeeniset kokoelmat on integroitu. Näin avulla käyttäjät voivat löytää tarvitsemansa aineistot tietämättä, mitä kokoelmia on olemassa, niiden säilytyspaikkaa tai käytäntöjä, joilla ne on kuvailtu.

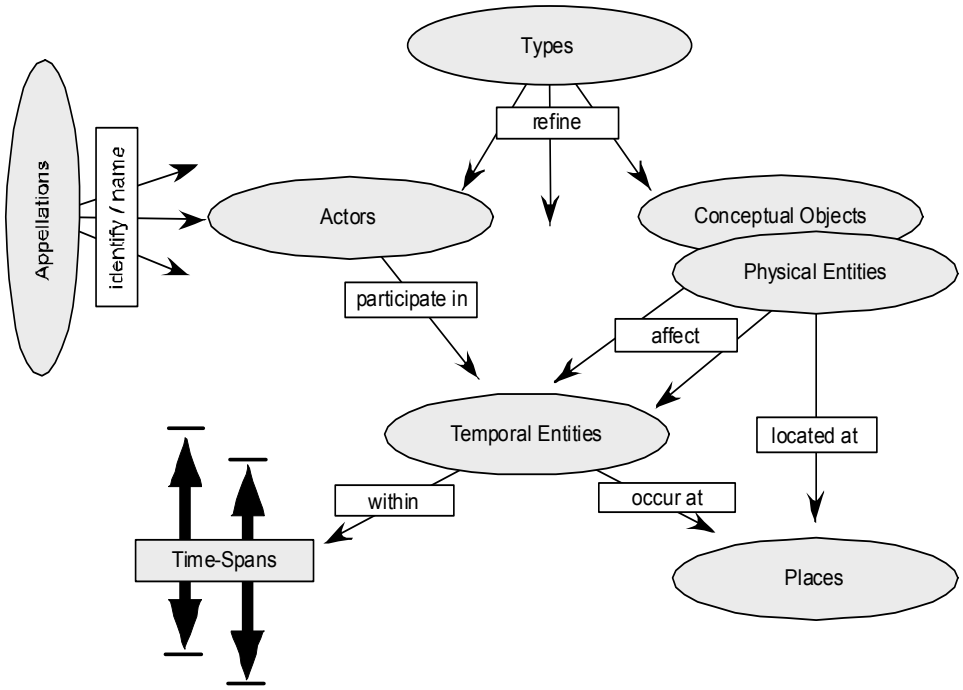
Heterogeenisten metatietojen yhteiskäyttöisyyteen voidaan pyrkiä monin tavoin (vaihtoehdoista ks. Chan ja Zeng 2006). Yleinen ratkaisu on muuntaa kokoelmien metatiedot tiedonhakua varten yleiskäyttöiseen skeemaan (esim. Dublin Core), johon muiden metatietoskeemojen vastaavuus määritellään. Huonona puolena on, että prosessissa menetetään tietoa. Koska vastaavuus perustuu skeemojen alimpaan yhteiseen nimittäjään, eri aineistoja ja tarkoituksia varten kehitettyjä metatietomalleja ei voida tukea. Yhteinen nimittäjä on liian pieni ollakseen riittävä vaativampiin tarpeisiin. Toiseksi, tulokseen yleensä sisältyy kompromisseja, koska rikkaampaa skeemaa on muuten vaikea yhdistää yksinkertaiseen. Kolmanneksi, vaikka tiedonhakuun tarkoitettu yksinkertainen metatietoskeema voi auttaa aineistojen löytämisessä, mikään ei takaa, että käyttäjät tutustuvat myös

alkuperäisiin, rikkaampiin ja ehkä hyödyllisempiin metatietoihin. Niinpä käyttäjät saattavat arvioida sekä aineistoja että säilyttäviä instituutioita vajaan ja epätyydyttävän metatiedon varassa. (Doerr 2003; Gill 2004.) Lisäksi Dublin Core ei riitä kompleksisten tietolähteiden kuvailuun ja vertailuun. Tällaisia ovat esimerkiksi poliittisen historian ja taidehistorian lähteet tai eri tieteenaloilla tuotettu tutkimusdata. (Doerr ja LeBoeuf 2007.)

CIDOC Conceptual Reference Model (lyhyesti CRM) tarjoaa toisenlaisen vaihtoehdon. CRM on viitemalli, jonka tavoitteena on mahdollistaa heterogeenisten tieteellisten kulttuurihistoriallisten kokoelmien tietojen vaihto ja yhdistäminen ilman, että informaatiota menetetään tai eri muisti-instituutioiden tarvitsee muuttaa kuvailukäytäntöjään. ”Heterogeenisuus” tarkoittaa, ettei alkuperäisten lähteiden tarvitse olla sisällöltään tai rakenteeltaan yhdenmukaisia. ”Tieteellisyys” sitä, että tieto on sisällöltään ja tarkkuudeltaan riittävää tutkimuksen tarpeisiin. ”Kulttuurihistorialliset kokoelmat” määritellään museoiden keräämiksi aineistoiksi, mutta mallin tarkoituksena on eksplisiittisesti palvella myös tietojen vaihtoa museoiden, arkistojen ja kirjastojen välillä (Gill 2004) sekä mahdollistaa esimerkiksi muistiorganisaatioiden tallentamien aineistojen ja tutkimusaineistojen yhdistely. ”Kulttuurihistoriallisuus” on ymmärrettävä mahdollisimman laajasti: kyse voi olla niin poliittisista tai lääketieteellisistä kuin yrityksissä, hallinnossa tai tutkimusprosesseissakin syntyvistä aineistoista (Doerr 2003.) Muodollisesti CRM:n kehitys alkoi v. 1996, mutta se on jatkoa vanhemmille CIDOC-standardeille (Gill 2004). Vuonna 2006 malli hyväksyttiin ISO 21127-standardiksi. (Doerr, Ore, & Stead 2007.) Seuraavassa esittelen CRM:n keskeisiä ajatuksia, sisältöä, taustaa ja siitä saatuja kokemuksia.

CIDOC CRM:n perusajatukset

CRM hyödyntää muistiorganisaatioiden kuvailujen taustalla olevien käsitemallien yhdenmukaisuutta. Ne jäävät ensi silmäyksellä piiloon, koska tietorakenteet on laadittu tukemaan kunkin erityisalueen tietojen keräämistä. Esimerkiksi kreikkalaisen vaasin, kirjeenvaihdon, vedenpinnankorkeusmittausten ja hyönteiskokoelman kuvailutiedot ovat sisällöltään erilaisia. Silti niin ihmisten tekemillä objekteilla



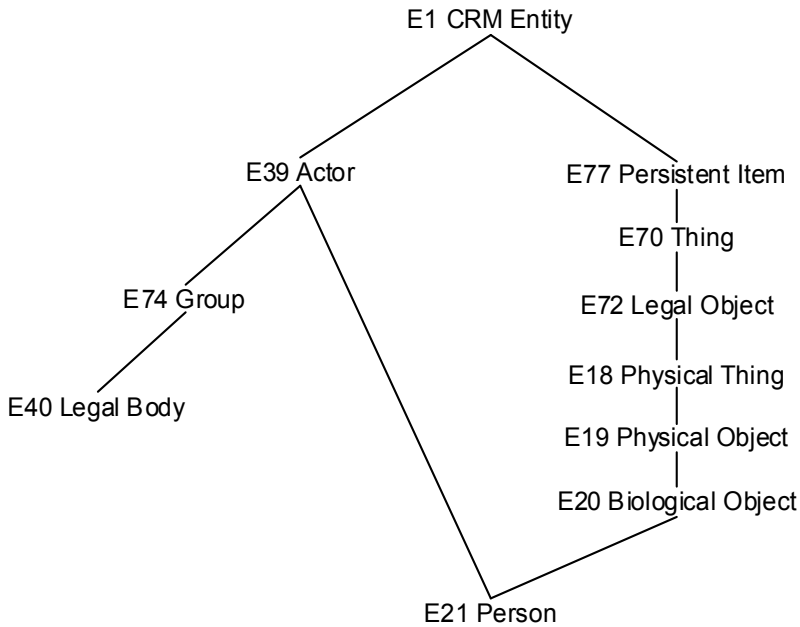
Kuvio 1 CIDOC CRM:n ylimmän tason entiteetit

(vaasit, rakennukset, kirjeenvaihto) kuin luonnonilmiöillä (hyönteiset, vedenkorkeus) on yhteisiä piirteitä. Ne ovat olemassa jonain aikana, jollain alueella, jonkin kulttuurin vallitessa. Niillä on elinkaari, jolla on alkupiste ja ehkä myös loppupiste. Joku on jossain lisännyt objektin kokoelmaan. Hyönteisen siipien väli, veden korkeus ja rakennuksen koko kertovat samasta piirteestä, objektin ulottuvuuksista. (Doerr, Ore ja Stead 2007.)

CRM:n avulla tällaiset erityisalueiden käsitemallien vastaavuudet voidaan määritellä. Useimmissa tapauksissa CRM:stä löytyy sopiva geneerinen, toimintaa, aihetta, objektia, ominaisuutta, paikkaa, aikaa tms. kuvaava entiteetti. Määrittelemällä vastaavuudet CRM:ään saadaan erityisalueiden tiedoista yksi tietämysverkosto, jota voidaan navigoida mihin suuntaan tahansa. Paras tapa esittää tietämysverkosto on käyttää CRM:n käsitteille määrittelemiä suhteita täydentäen sitä erityisalueen terminologialla. (Doerr, Ore ja Stead 2007.) CRM:n ylimmän tason entiteetit on esitetty kuviossa 1.

Kuviosta 1 näkyy yksi CRM:n perusajatuksista, tapahtumakeskeisyys: ”ajalliset entiteetit” liittävät käsitemallin osat toisiinsa. Ne voivat olla lyhytkestoisia tapahtumia tai pitkäkestoisempia prosesseja. Ajallisissa entiteeteissä toimijat, paikat, ideat, fyysiset objektit ja ajankohdat kohtaavat toisensa. Dokumentoitu menneisyys ymmärretään sarjaksi ajassa ja paikassa tapahtuvia ilmiöitä, joihin fyysiset objektit, ideat ja erilaiset toimijat osallistuvat. Esim. julkaisun ilmestyminen on ajallinen entiteetti, johon liittyy niin fyysinen objekti (julkaisu), joukko toimijoita (kirjoittaja, kustantaja, kirjapaino) kuin ideoita, joita julkaisu tai julkaiseminen ilmentää. Samat toimijat, objektit ja ideat liittyvät yleensä myös toisiin tapahtumiin. Siksi mallintamalla yksittäisiä tapahtumia on mahdollista luoda irrallisista faktoista koherentti historianesitys. Ideat tms. käsitteelliset objektit voivat manifestoitua erilaisissa fyysisissä muodoissa samanaikaisesti.

Ajalliset entiteetit yhdistävät toimijat ideoihin ja paikkoihin. Muilla entiteeteillä on yhteys aikaulottuvuuteen vain ajallisten entiteettien kautta. Esimerkiksi tieto ”henkilön syntymäaika” tarvitsee viitemallissa kolme entiteettiä: henkilön (toimija/biologinen olio), syntymän (tapahtuma) ja syntymähetken (joka ei ole diskreetti hetki vaan ajanjakso, kuten kaikki ajoitukset CRM:ssä). (Doerr 2003; Doerr, Ore ja Stead 2007.) Mallin keventämiseksi CRM sallii joukon ”oikopolkuja”. Esimerkiksi objektin koko ei ole tiedossa ilman, että joku on mitannut sen. Niinpä tiedon taustalla on implisiittisesti mittaustapahtuma. Pikalinkin ansiosta tapahtumaa ei tarvitse eksplikoida, vaikka sekin on tarvittaessa mahdollista. (Stead 2008.)



Kuvio 2 Esimerkki CIDOC CRM -entiteettihierarkiasta

Perinteisessä resurssikeskeisessä kuvailussa objektin ominaisuuksien oletetaan olevan stabiileja. Tosin esimerkiksi Dublin Coressa voidaan kuvata objektin synty- tai muutosajankohtaa (”date.created”, ”date.modified”), mutta koska tapahtumia ei ole erotettu omaksi entiteetiksi, vastuita ja tietoa objektin tilan muutoksesta on vaikea yhdistää tähän. Tapahtumakeskeinen ontologia sitä vastoin pystyy kuvaamaan objektien syntyä, käyttöä ja muutosta sekä vastaamaan kysymyksiin ”kuka oli vastuussa, mistä, milloin ja missä”. (Hunter 2003.) Lagoze (2000) ehdottaa tapahtumakeskeisyyttä kaiken kuvailun ja luetteloinnin lähtökohdaksi: tällöin mikä tahansa joukko metatietoa ymmärretään vain tiettyyn yhteisöön ja aikaan sidotuksi näkemykseksi kuvailtavan resurssin tilasta.

CRM kehitettiin ”alhaalta ylös” periaatteella monitieteisissä asiantuntijaryhmissä, joissa oli mukana mm. tietojenkäsittelytieteen, arkeologian, museoiden, taidehistorian, fysiikan ja kirjastotieteen edustajia. Viitemalliin valittujen käsitteiden (entiteettien) tuli olla asiantuntijoiden mielestä erityisalueella keskeisiä. Toinen tärkeä kriteeri oli entiteettien käyttötiheys tietorakenteissa, mikä katsottiin merkiksi todellisista tarpeista. Viitemallin käsitehierarkia laadittiin ajatellen semanttista webiä ja olio-ohjelmointia. Ominaisuudet periytyvät ylemmiltä entiteettiluokilta alemmille. Entiteetti voi periä ominaisuuksia useammalta taholta. (Doerr, Ore ja Stead 2007.)

Entiteetit ja ominaisuudet identifioidaan CRM:ssä kirjaimen ja numeron yhdistelmällä. Entiteettien tunnuksena on ”E”. Kuviossa 2 on esimerkkinä kuvattu osa E21 Person entiteetin periytymishierarkiasta: kaikki ”henkilöt” ovat käsitemallissa sekä toimijoita että eräs pysyvien biologisten olijoiden alalaji. Kaikki viitemallin entiteetit ovat puolestaan CRM:n entiteettejä. (Crofts et al. 2010, xxii-xxiii.)

Myös CRM:n tunnistamilla luokkien suhteilla—joita viitemallissa kutsutaan ”ominaisuuksiksi”—on oma hierarkiansa (Theodoridou et al. 2010). Ominaisuuksien tunnuksena on ”P”. Esimerkkinä ominaisuuksista on tunnistamissuhteiden hierarkia taulukossa 1: CRM:n entiteetit (E1) tunnistetaan nimillä (E41). Yhtenä tämän suhteen muotona on (P87) paikannimien (E44) antaminen paikoille (E53).

CRM siis määrittelee entiteetit sekä niiden välillä mahdolliset suhteet. Yleensä ontologioita laadittaessa lähdetään liikkeelle entiteeteistä, mutta CRM:ää kehitettäessä kiinnitettiin erityistä

Taulukko 1. Esimerkki CRM:n suhdehierarkiasta (Crofts et al. 2010, xxv)

Entity – Domain	**Property**	**Entity – Range**
E1 CRM Entity	P1 is identified by (identifies)	E41 Appellation
E1 CRM Entity	P48 has preferred identifier (is preferred identifier of)	E42 Identifier
E52 Time-Span	P78 is identified by (identifies)	E49 Time Appellation
E53 Place	P87 is identified by (identifies)	E44 Place Appellation
E71 Man-Made Thing	P102 has title (is title of)	E35 Title
E39 Actor	P131 is identified by (identifies)	E82 Actor Appellation

huomiota suhteiden tunnistamiseen. Martin Doerr kutsuu tätä ”ominaisuuskeskeiseksi ontologiaksi”. Suhteet olivat yksi kriteeri sille, mitä käsitteitä otettiin käsitemalliin mukaan. Mukana ovat vain peruskäsitteet, jotka ovat tarpeen suhteiden määrittelemiseksi. Käsitteitä vailla suhteita ei pidetty mallin kannalta kiinnostavina. Taustaajatuksena oli myös, että tutkijat haluavat tiedon ja aineistojen löytämisen ohella ymmärtää asioita. Se tulee mahdolliseksi tekemällä suhteet näkyviksi. Samalla suhteet mahdollistavat olemassa olevan tieteellisen tiedon linkittymisen ja käytön uusiin tarkoituksiin. Paras esimerkki tällaisista uudelleenkäytön mahdollisuuksista on Mendelejevin jaksollinen järjestelmä, jonka pohjana oli 1860–1870 –luvuilla tehtyjen tutkimustulosten systematisointi. (Doerr, Ore ja Stead 2007; Doerr 2003.)

Viitemallissa on 90 entiteettiä ja noin 150 niiden välistä suhdetta. CRM:n käsitemalli on tästä huolimatta melko yksinkertainen. Yhtenä syynä on se, että yksi tapahtuma puretaan useaksi osatapahtumaksi, jotka voidaan koostaa yhteiseksi yläkäsitteeksi eri tavoin. Esimerkiksi aineiston luovutus muistiorganisaatiolle on tapahtuma, joka voi pitää sisällään paikan vaihdoksen ja hoitovastuun muutoksen, mutta näin ei ole välttämättä. Purkamalla tapahtumat yksinkertaisempiin, vapaasti yhdisteltäviin osiin (esim. entiteetteihin ”säilytyspaikan muutos”, ”omistussuhteen muutos”, ”säilytyksestä vastaavan tahon muutos”, jotka voivat esiintyä erikseen tai yhdessä eri yhdistelminä) eri variaatiot voidaan esittää huomattavasti yksinkertaisemmin kuin luomalla jokaisesta erilaisesta tapahtumasta oma entiteetti. Kompleksiset asiat toisin sanoen esitetään yhdistelemällä yksinkertaisia osia, ei kasvattamalla mallin kompleksisuutta. (Stead 2008.)

CRM sallii (Doerr ja Kritsotaki 2006; Doerr, Ore ja Stead 2007):

- olioiden identifioimisen niiden reaalimaailman nimillä
- olioiden luokittelun
- osa–kokonaisuussuhteiden ilmaisemisen (sekä käsitteellisten että fyysisten olioiden purkamisen osiin, ml. ajalliset entiteetit, henkilöiden muodostamat ryhmät (toimijat), paikat ja ajat)
- persistenttien olioiden osallistumisen ajallisiin entiteetteihin
- objektien ja toiminnan ja sen tulosten vuorovaikutuksen sekä
- tieto-objektien viittaamisen reaalimaailman olioihin (”aboutness”).

CRM ei ole preskriptiivinen tai tieto- tai metatietostandardi (Gill 2004). Se kuvaa, mitä muistiorganisaatiot jo tällä hetkellä dokumentoivat. Viitemalli ei määrittele, mitä muistiorganisaatioiden tulisi kuvailla tai mitä terminologiaa tulisi käyttää. Osoituksena on, että vaikka dokumentissa, jossa CRM:n luokat ja suhteet määritellään, ilmaistaan myös suhteiden kardinaliteetti, kardinaliteettirajoitteilla on vain informatiivinen merkitys: ne eivät kuulu itse standardiin. Kaikki suhteet ovat ei-pakollisia ja toistettavissa. (Crofts et al. 2010, i, xxii–xxiii.)

Olemassa olevia kuvailutietoja pidetään lähtökohtaisesti CRM:ssä valideina, mutta samalla niiden katsotaan heijastavan omaa syntykontekstiaan sekä muuttuvaa ja joskus ristiriitaistakin historiallista tietämystä. Entiteettien tyypit ja käytetyt luokitusjärjestelmät ovat CRM:ssä, paitsi tapa strukturoida informaatiota, myös osa historiallista todellisuutta. Niinpä muistiorganisaatioiden tuottama dokumentaatio ja erilaiset tavat luokittaa todellisuutta ovat samanaikaisesti sekä ontologian avulla ilmaistavia sisältöjä että sen puitteissa kuvattavia kohteita. ”Nimet” ovat mallissa oma luokkansa, koska niidenkin käyttö on sidottu tiettyyn historialliseen kontekstiin. Nimien monitulkintaisuutta ei pidetä ongelmana vaan osana sitä muuttuvaa tietämystä, jota CRM mallintaa. Muistiorganisaatioiden kuvailuinformaatio nähdään osaksi ylläpidon, täydentämisen ja säilyttämisen diskurssia. Jotta sitä voidaan ymmärtää ja päivittää, alkuperäinen dokumentaatioyksikkö on säilytettävä. (Doerr, Ore ja Stead 2007; Doerr 2003.)

Yhteiskäyttöisyydelle merkityksettömiä instituutiospesifejä tietoja, jotka liittyvät esim. kokoelmanhallintaan, ei ole huomioitu, ei myöskään epädiskreettiä matemaattista informaatiota. Oletuksena on, että liiketoiminnan tai tutkimuskokeiden dataa voidaan yhdistää vain, jos se on ontologisesti samalla tasolla, ts. koskee samanlaisia ilmiöitä. (Doerr, Ore ja Stead 2007.)

CIDOC CRM:n laajennuksia

Vaikka CRM:stä on tässä tekstissä puhuttu viitemallina sen abstraktiuden ja tekstipohjaisuuden vuoksi, CRM voidaan tulkita myös ydinontologiaksi, jossa määritellään keskeiset peruskäsitteet. Toinen esimerkki ydinontologiasta on ns. ABC ontologia, jolla pyritään tapahtumakes-

keisesti (kuten CRM:ssä) sisällönkuvailuun ja oikeuksien hallintaan kehitettyjen metatietomallien yhteiskäyttöisyyteen. ABC ontologian avulla voidaan esimerkiksi tallenteen sisältöä, käyttöoikeuksia ja teknisiä ominaisuuksia kuvaavat metatiedot yhdistää. (Hunter 2003; Hunter ja Darren 2000; Lagoze ja Hunter 2002). ABC on muistuttaa sisällöltään ja tarkoitukseltaan CRM:ää, joten myös näiden kahden ydinontologian harmonisoimista on suunniteltu (Hunter 2002; Doerr, Hunter ja Lagoze 2003). Samankaltaisilta vaikuttaneissa käsitteissä on kuitenkin paljastunut eroja (Kalinichenko et al. 2003).

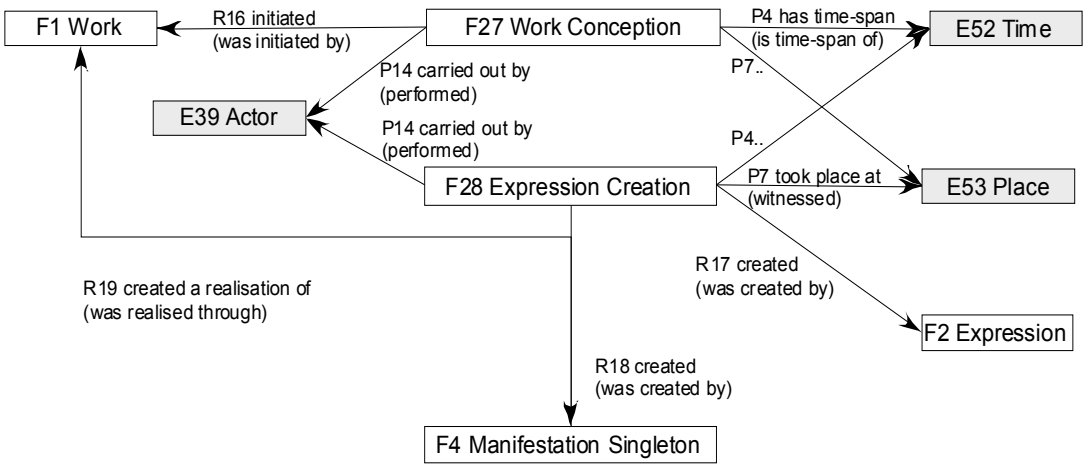
CRM:n kehitystyössä oli tavoitteena, että sitä voidaan soveltaa erityisaloilla luomalla peruskäsitteistä tarkempia tyyppejä ilman että määriteltyihin suhteisiin tarvitsee koskea. (Doerr, Ore ja Stead 2007; Doerr 2003). Tällä tavoin syntyneitä CRM:n laajennuksia ovat ainakin ontologia digitaalisten objektien alkuperän kuvaamiseen (Theodoridou et al. 2010) sekä CRM-EH, English Heritage Extension, jota on yhdessä erityisalueen tesaurusten kanssa sovellettu arkeologisten kaivausten dokumenttien automaattisessa annotoinnissa (Vlachidis et al. 2010).

Keskeisin CRM:n laajennus on FRBRoo joka on bibliografisen kuvailun käsitemallin, FRBR:n (Functional Requirements for Bibliographic Records), CRM yhteensopiva, oliomallin mukainen tulkinta: ”oo” -lyhenteen lopussa tulee sanoista ”objektiorientoitunut”. Käynnissä on myös prosessi, jossa myös kahdesta FRBR:ää täydentävästä IFL-standardista, auktoriteettidataa mallintavasta FRADista (Functional Requirements for Authority Data) ja sisällöllisiä teemoja koskevasta FRSADista (Functional Requirements for Subject Authority Data) ollaan toteuttamassa CRM-yhteensopivia, oliopohjoisia malleja. Myöhemmin tavoitteena on liittää kokonaisuuteen arkistoalan käsitemallit. FRBR:llä ja FRBRoo:lla on erilaiset päämäärät ja painotukset. FRBR:n entiteetti–suhde -malli on tarkoitettu ensisijaisesti tietorakenteiden esittämiseen, kun taas FRBRoo pyrkii kuvaamaan tietorakenteiden taustalla olevaa maailmaa. FRBR:ssä tapahtumia ei huomioitu, mutta FRBRoo:ssa kohteena ovat myös prosessit. (Bekiari, Doerr ja Le Bœuf 2010; Doerr ja LeBoeuf 2007.) FRBRoo:n kehitystyön tuloksena oli täydennyksiä myös CRM:ään.

Kuvio 3 Teoksen (Work) ja ajan (Time) suhde FRBRoossa (Riva, Doerr & Žumer 2009)

Kuviossa 3 on esimerkkinä työn ja ajan suhdetta kuvaava kaavio FRBRoo:sta (Riva, Doerr ja Žumer 2009). Harmaat laatikot kuvaavat entiteettejä, jotka ovat sellaisinaan CRM:ssä. FRBRoo:n omia ovat entiteetit, joiden tunnuksena on F, ja suhteet, joiden tunnuksessa on R. Kuten kaaviosta huomaa, entiteettien oikea tulkinta vaatii usein niiden kuvausten lukemista, nimet tunnuksineen eivät siihen riitä.

FRBRoo:n laatimisprosessia vaikeutti se, ettei FRBR huomioi tapahtumia. Siksi monet attribuuteista liittyivät entiteetteihin, johon ne eivät



Kuvio 3 Teoksen (Work) ja ajan (Time) suhde FRBRoossa (Riva, Doerr & Žumer 2009)

oikeasti kuulu. Tämä teki niiden semantiikasta epäselvän. Työryhmälle oli esimerkiksi yllätys, etteivät julkaisuaika ja paikka välttämättä liity teoksen painamiseen. Kehitystyössä ilmeni myös, että jokainen julkaisu koostuu käytännössä joukosta erilaisia teoksia (teksti, kuvitus, toimittajat, taitto jne.), vaikkei FRBR sitä huomioikaan. Tämä oli otettava FRBRoo:n mallinnuksessa. Koska FRBRoo on kevyisiin sovelluksiin liian kompleksinen ja vaikea omaksua, siitä on tarkoitus tuottaa oma ydinmäärittely. (Doerr ja LeBoeuf 2007.)

Bountourin et al. (2010) mukaan CRM:ää voitaisiin käyttää integroitaessa julkishallinnon tietovarantoja, esimerkiksi rakennettaessa yhteisiä portaaleja tai suunniteltaessa yhteistä arkkitehtuuria. Asiakirjahallinta on tehtävä- ja prosessilähtöistä, mikä voi helpottaa harmonisointia CRM:n kanssa. Arkistohakemistojen sisältöä rakenteisena tekstinä määrittävän EADin (Encoded Archival Description) mappaus CRM:ään on myös tehty. CRM:n arvioitiin soveltuvan arkistojen kuvailutietoihin melko hyvin. Ongelmia tuotti eniten EADin epämääräisyys, ei tietojen sisältö. (Theodoridou ja Doerr 2001.) Myös museoissa on törmätty siihen, että koodaus datassa palvelee enemmän esitystapaa kuin sisällön esittämistä. (Doerr 2003).

Saadut kokemukset

CRM:n kehittäjien oma arvio viitemallista on myönteinen: viitemalli on museolähtökohdista huolimatta yllättävän geneerinen. Positiivisena pidetään viitemallin muuttumista yhä kompaktimmaksi ja muutostarpeen vähenemistä samalla kun uusia skeemoja on analysoitu. Tätä nykyä CRM on hyvin stabiili eikä siihen viimeisen kymmenen vuoden aikana ole juuri tehty muutoksia. Tämän katsotaan osoittavan viitemallin olevan käsitteellisesti viimeistelty. CRM:n sanotaan näyttäneen toimivuutensa myös käytännössä. (Doerr, Ore ja Stead 2007; Theodoridou et al. 2010.)

Tätä nykyä CRM:n kotisivulla olevassa bibliografiassa on yli 120 julkaisua.[1] Monissa julkaisuista viitemalliin tosin vain viitataan lyhyesti, mutta kokemustakin viitemallista alkaa siis olla. Arvio viitemallin toimivuudesta käytännössä saa osin tukea muilta tutkijoilta, joskin soveltamisessa on havaittu myös ongelmia. Kokonaisarviointia vaikeuttaa, että CRM:ää voidaan käyttää eritasoisesti ja monin tavoin. Viitemallia voidaan tarkastella abstraktisti tai ainakin etenemättä konkreettisen sovelluksen asteelle (ks. esim. Øyvind 2008; Marins et al. 2007; Janowicz 2007). Perusideoita voidaan soveltaa hyödyntämällä viitemallia sen enempää. Käsitteellisten ja fyysisten entiteettien, toimijoiden, paikkojen ja aikojen yhdistäminen on kiinnostanut varsinkin GIS-datan, arkeologisen aineiston ja kaivosteollisuuden historian parissa toimivia (esim. Hiebel, Hanke ja Hayek 2010; Katsianis et al. 2008; Vlachidis et al. 2010; Nussbaumer and Haslhofer 2007b; Holmen, Ore ja Eide 2003). Viitemallin avulla on yhdistetty antiikin kuvakokoelmia (Kurtz et al. 2009), multimediaa (Goodall et al. 2004; Addis et al. 2003), filosofisia tekstejä (Pasin and Motta 2009; Pasin, Motta ja Zdrahal 2007), musiikkiaineistoja (Eggen 2007) ja CRM-yhteensopivia museotietokantoja (Elbekai and Rossiter 2005). Myös taitelijaelämäkertojen dynaamisesta generointia nettiaineistoista on kokeiltu (Millard et al. 2003; Alani et al. 2003).

Yleensä viitemalli jää tulosten raportoinnissa melko vähälle huomiolle. Tämä kertonee siitä, että se on palvellut tarkoitustaan tyydyttävästi. Tosin usein tietojen yhdistäminen on tapahtunut vain ”paperilla”, jolloin viitemalliin liittyvät ongelmat ovat voineet jäädä piiloon. Suurimpia käytössä olevia CRM-pohjaisia tietojärjestelmä lienee kuvataiteen CLAROS-tietokanta (ks. http://www.clarosnet.org), jossa on yli 20 miljoonaa tietuetta. CLAROSissa viitemallin sanotaan toimineen tietojen yhdistämisessä hyvin (Kurtz et al. 2009). Monitieteisessä tutkimuksessa CRM on ”hyvin hyödyllinen malli”, joka on toteutettavissa relaatiotietokantana (Hiebel, Hanke ja Hayek 2010).

Vaikka viitemalli on tätä nykyä viimeistelty, yksi käsitteellinen ongelma siihen liittyy. Viitemallin E55 Type -entiteetti on tarkoitettu rajapinnaksi, jonka avulla kontrolloidut asiasanastot ja luokitukset voidaan tuoda siihen mukaan. E55 Type -entiteetit edustavat käsitteitä, joilla CRM:n entiteettejä luokitetaan. (Crofts et al. 2010, xvii, 25). Epäselvää on, milloin E55 Typen sijasta olisi käytettävä alaluokkaa: ”taitelija” voi esimerkiksi olla yhtä hyvin E21 Person -entiteettiluokan ”tyyppi” kuin sen oma alaluokka. On myös todettu, että yhteiskäyttöisyys vaatii tässä yhteisen tesauruksen tai luokituksen käyttöä. (Hiebel, Hanke ja Hayek 2010).

Viitemallin keskeisin vaikeus liittyy kuitenkin siihen, että se on abstrakti, kompleksinen ja

sallii erilaisia toteutuksia. Aina ei ole helppoa tulkita viitemallia ja ratkaista, miten se vastaa käsillä olevan informaatiota (Eggen 2007, 86). Tarvittaisiin ohjeistusta siitä, miten reaalimaailman ilmiöt vastaavat viitemallin luokkia (Hiebel, Hanke ja Hayek 2010). Metatietojen määritteleminen suhteessa CRM:ään vaatii sekä metatietojen että viitemallin semantiikan hyvää tuntemusta (Nussbaumer, Haslhofer ja Klas 2010), varsinkin kun useinkaan kyse ei ole yhden vastaavan entiteetin löytämisestä viitemallista, vaan sellaisen viittausketjun luomisesta, jolla ilmaistaan metatiedon semanttinen sisältö. Esimerkiksi esinettä kuvaava metatietokenttä ”ajanjakso” voi kääntyä CRM:ksi seuraavasti: ”E84 Information Carrier – P8 witnessed – E4 Period – P1 is identified by – E41 Appellation (value = object.period)” (Addis et al. 2005.)

CRM:n perusongelma seuraakin juuri tästä: metatiedot on ensin ”käännettävä” viitemallin kielelle ja sitten ilmaistava teknisesti. Kumpaankaan ei ole olemassa valmiita malleja tai välineitä, mikä johtaa viitemallin epäyhtenäiseen soveltamiseen. Viitemalli on esitetty tarkoituksellisesti vain tekstimuodossa, jotta sen riippumattomuus ontologioiden kuvauskielistä olisi ilmeinen. (Doerr, Ore ja Stead 2007; Doerr 2003.)[2] Ongelma havaittiin selvästi BRICKS-projektissa, jossa rakennettiin eri museoiden numismaattiset kokoelmat yhdistävää tietokantaa. Projektin lopputulema on hyvin kriittinen CRM:ää ja sen mahdollisuuksia kohtaan: CRM haittaisi tietojen yhteiskäyttöisyyttä, koska se toi metatietojen monitulkintaisuuteen yhden ongelman lisää. CRM:n tavoitteena oli, että eri yhteisöjen tuottamia metatietoja voitaisiin käyttää yhdessä ilman että yhteisöt muuttavat toimintaansa tai neuvottelevat keskenään. Koska CRM:n tulkinta ja tekninen toteutus kuitenkin ovat avoinna, tavoite ei toteudu. Koska eri yhteisöt voivat tulkita ja toteuttaa CRM:ää eri tavalla, sama metatieto voi tulla esitetyksi viitemallin puitteissa ja teknisesti eri tavoin. Tällöin ei ole taetta sille, että eri yhteisöjen semanttisesti identtiset metatiedot saavat lopulta teknisesti saman ilmaisun tai että samaan tekniseen ilmaisuun ei päädyttäisi alkuperäisten metatietojen ollessa semanttisesti erilaisia. Vaikka olemassa olevien tietovarantojen yhdistämisessä on näin ongelmia, CRM saattaa toimia paremmin sellaisten uusien ontologioiden yhteensovittamisessa jotka laaditaan ottaen se huomioon. (Nussbaumer, Haslhofer ja Klas 2010; Haslhofer ja Nussbaumer 2009; Nussbaumer ja Haslhofer 2007b, 2007a.)

CRM-yhteensopivat tietojärjestelmät voivat siis olla keskenään yhteensopimattomia. Viitemallin kompleksisuus ja monitulkintaisuus on ongelma luotaessa yhteiskäyttöistä tietovarantoa. Sama ongelma tulee vastaan myös tiedonhaussa, jossa viitemalli on sopivasti piilotettava tiedonhakijoilta. Esimerkiksi roomalaisen kolikon löytäminen BRICKS-tietokannasta vaati seuraavaa ketjua: ”E22 Man-Made Object – P108 was produced by – E12 Production Event – P10 falls within – E4 Period – P1 is identified by – E49 Time Appellation – *Roman”*. (Nussbaumer ja Haslhofer 2007b.) Tarvitaan ”käyttöliittymiä CRM-lukutaidottomille” (Hiebel, Hanke ja Hayek 2010).

BRICKS-tietokannan tapauksessa yksinkertaisemmat keinot metatietojen yhdistämiseen olisivat luultavasti tuottaneet paremman tuloksen. Lagoze ja Hunter (2002) toteavat ABC ontologian pohjalta, että tuki monimutkaisille kyselyille väistämättä lisää kustannuksia ja käsin tehtävää työtä. Siksi kustannusten ja hyötyjen suhdetta on tarkasti pohdittava: usein on tarkoituksenmukaisempaa kuvailla suuri joukko aineistoja yksinkertaisella ”pidgin metadatalla” kuin pieni joukko rikkaasti ilmaisuvoimaisen ydinontologian avulla.

Johtopäätökset

CRM:ää on toistaiseksi tutkittu ja käytetty rajoitetusti. Vaikka viitemalli on ajateltu palvelemaan kirjastojen, arkistojen ja museoiden aineistojen yhteiskäyttöisyyttä, sitä ei ole sovellettu tällä alueella. Arkistoilla ei tosin ole tarjota FRBRoo:n kaltaista CRM-yhteensopivaa käsitemallia, mutta myös kirjasto- ja museoaineistoja yhdistävät palvelut puuttuvat. Tutkimuksessa on aukko myös käyttäjälähtöisen tutkimuksen alueella: CRM:ää on tarkasteltu lähinnä mallinnuksen tai teknisten toteutusten näkökulmasta. Tutkimus, jossa tarkasteltaisiin sen avulla rakennettujen palveluiden sopivuutta tutkimusprosessiin tai tutkijoiden tiedonhakuun, puuttuu.

Suomessa Kansallisen Digitaalisen Kirjaston (KDK) kaltaisten hankkeiden luulisi olevan kiinnostuneita CRM:stä. Viitemallin mielenkiintoisimpia mahdollisuuksia on uusien yhteyksien syntyminen: käyttäjä ei löydä vain samansisältöisen metatiedon yhteen liittämiä esineitä tai dokumentteja, hän löytää objekteja,

jotka liittyvät ideoihin, toimijoihin tai paikkoihin, jotka puolestaan liittyvät uusiin ideoihin, esineisiin tai tietoaineistoihin (jne). Vastaavantapaista navigointia on toteutettu yhden kokoelman sisällä (ks. Mulholland, Collins ja Zdrahal 2005). KDK:n rikas aineisto tarjoaisi tutkimus- ja kehityshankkeille erinomaisen alustan.

Viitemallin suurin ongelma liittyy välimatkaan teknisen toteutuksen ja alkuperäisten heterogeenisten aineistojen välillä: ajatus siitä, että CRM:n avulla tämä kuilu voitaisiin ylittää helposti—ilman että aineistoja tuottavat yhteisöt neuvottelevat keskenään—vaikuttaa liian kunnianhimoiselta. KDK:n ympärille syntynyt yhteisö voisi kuitenkin olla foorumi, jolla tulkinnoista ja käytännöistä sovitaan yhdessä. Tällöin vältettäisiin ainakin osa viitemalliin soveltamiseen liittyvistä käytännön ongelmista.[3]

Viitteet

[1] http://www.cidoc-crm.org/references.html

[2] Nyt CIDOC CRM:n nettisivulla on luonnoksia siitä, miten viitemallin mukaiset tiedot voidaan esiittää RDF:n ja OWL:in avulla.

[3] Kiitos Mika Nymanille (Synapse Computing) kommenteista artikkeliversioon.

Lähteet

Addis, M., M. Boniface, S. Goodall, P. Grimwood, S. Kim, P. Lewis, K. Martinez, and A. Stevenson. 2003. "SCULPTEUR: Towards a New Paradigm for Multimedia Museum Information Handling." *Proceedings of Semantic Web ISWC 2870* (November):582 – 596.

Addis, M. J., S. Goodall, P. H. Lewis, K. Martinez, P. A. S. Sinclair, F. Giorgini, C. Lahanier, J. Stevenson, M. Cappellini, L. Serni, and R. Rimaboschi. 2005. Searching and Exploring Multimedia Museum Collections Over the Web In *EVA, 14 - 18 Mar 2005*. Palazzo dei Congressi, Florence, Italy.

Alani, Harith, Sanghee Kim, David E. Millard, Mark J. Weal, Paul H. Lewis, Wendy Hall, and Nigel R. Shadbolt. 2003. "Automatic Extraction of Knowledge from Web Documents." *2nd International Semantic Web Conference - Workshop on Human Language Technology for the Semantic Web abd Web Services, Sanibel Island, Florida, USA, 20 - 23 Oct 2003.*

Bekiari, Chryssoula, Martin Doerr, and Patrick Le Bœuf. 2010. *FRBR object-oriented definition and mapping to FRBR$_{ER}$ (version 1.0.1)*.

Bountouri, Lina, Christos Papatheodorou, and Manolis Gergatsoulis. 2010. "Modelling the public sector information through CIDOC conceptual reference model." *Lecture Notes in Computer Science* no. 6428:404-413.

Chan, Lois Mai, and Marchia Lei Zeng. 2006. "Metadata interoperability and standardization. A study of methodology. Parts I and II." *D-Lib Magazine* no. 12 (6).

Crofts, Nick , Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff. 2010. *Definition of the CIDOC Conceptual Reference Model, January 2010. Version 5.0.2*.

Doerr, Martin. 2003. "The CIDOC Conceptual Reference Module. An ontological approach to semantic interoperability of metadata." *AI Magazine* no. 24 (3):75–92.

Doerr, Martin, Jane Hunter, and Carl Lagoze. 2003. "Towards a core ontology for information integration." *Journal of Digital Information* no. 4 (1).

Doerr, Martin, and A. Kritsotaki. 2006. Documenting events in metadata. In *The 7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST*, edited by M. Ioannides, D. Arnold, F. Niccolucci and K. Mania.

Doerr, Martin, and Patrick LeBoeuf. 2007. "Modelling Intellectual Processes: The FRBR - CRM Harmonization." In *Digital Libraries: Research and Development*, edited by Costantino Thanos, Francesca Borri and Leonardo Candela, 114-123. Springer Berlin / Heidelberg.

Doerr, Martin, Christian-Emil Ore, and Stephen Stead. 2007. The CIDOC Conceptual Reference Model. A New Standard for Knowledge Sharing. Paper read at Tutorials, posters, panels and industrial contributions at the 26th International Conference on Conceptual Modeling - ER 2007, at Auckland, New Zealand.

Eggen, Lars Gunnar. 2007. *Ontologibasert musikkmetadata*, Institutt for datateknikk og informasjonsvitenskap, Norges teknisk-naturvitenskapelige universitet, Trondheim.

Elbekai, Ali S., and Nick Rossiter. 2005. Virtual exhibitions framework: utilisation of XML data processing for sharing museum content over the web Paper read at CIDOC Annual Conference, at Zagreb.

Gill, Tony. 2004. Building semantic bridges between museums, libraries and archives: The CIDOC Conceptual Reference Model. *First Monday* 9 (5–3 May), http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1145/1065.

Goodall, S., P. Lewis, K. Martinez, P. Sinclair, M. Addis, C. Lahanier, and J. Stevenson. 2004. Knowledge-Based Exploration of Multimedia Museum Collections. In *European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (EWIMT)*. London.

Haslhofer, Bernhard, and Phillipp Nussbaumer. 2009. "CIDOC CRM in Practice." *Solutions*:1-25.

Hiebel, Gerald, Klaus Hanke, and Ingrid Hayek. *A relational database structure and user interface fro the CIDOC CRM with GIS integration presented in the 22th CIDOC CRM SIG meeting, Nuremberg, Germany, December 20-22, 2010 [PowerPoint esitys]* 2010 [cited 22.3.2012. Saatavissa http://www.cidoc-crm.org/docs/Hiebel_crm_sig_2010.ppt (viitattu 22.3.2012).

Holmen, Jon, Christian-Emil Ore, and Oyvind Eide. 2003. "Documenting two histories at once: Digging into archaeology." *Proceedings of CAA 2003, 8 -12 April 2003*.

Hunter, Jane. 2002. Combining the CIDOC CRM and MPEG-7 to describe multimedia in museums.

Hunter, Jane. 2003. "Enhancing the semantic interoperability of multimedia through a core ontology." *IEEE Transactions on Circuits and Systems for Video Technology* no. 13 (1):49-58. doi: 10.1109/tcsvt.2002.808088.

Hunter, Jane, and James Darren. 2000. The application of an event-aware metadata model to an online oral history project.

Janowicz, Krzysztof. 2007. Towards a Similarity-Based Identity Assumption Service for Historical Places. In *GIScience'06. 4th international conference on Geographic Information Science* .

Kalinichenko, L.A., M. Missikoff, F. Schiappelli, and N. Skvortsov. 2003. "Ontological Modeling." *Proceedings of the 5th Russian Conference on Digital Libraries RCDL2003, St.-Petersburg, Russia, 2003*.

Katsianis, M., S. Tsipidis, K. Kotsakis, and A. Kousoulakou. 2008. "A 3D digital workflow for archaeological intra-site research using GIS." *Journal of Archaeological Science* no. 35 (3): 655-667. doi: 10.1016/j.jas.2007.06.002.

Kurtz, Donna, Greg Parker, David Shotton, Graham Klyne, Florian Schroff, Andrew Zisserman, and Yorick Wilks. 2009. CLAROS – bringing classical art to a global public. In *E-SCIENCE '09. Fifth IEEE International Conference on e-Science*.

Lagoze, Carl. 2000. Business unusual: how "event-awareness" may breathe life into the catalog? Paper read at Bicentennial Conference on Bibliographic Control for the New Millennium, November 15-17 2000, at Library of Congress.

Lagoze, Carl, and Jane Hunter. 2002. "The ABC ontology and model." *Journal of Digital Information* no. 2 (2).

Marins, A. , M. A. Casanova, A. Furtado, and K. Breitman. 2007. Modeling Provenance for Semantic Desktop Applications. In *SBC 2007*.

Millard, D. E., H. Alani, S. Kim, M. J. Weal, P. Lewis, W. Hall, D. De Roure, and N. Shadbolt. 2003. "Generating Adaptive Hypertext Content from the Semantic Web." *Proceedings of 1st International Workshop on Hypermedia and the Semantic Web, Nottingham, UK, July 2003*.

Mulholland, Paul, Trevor Collins, and Zdenek Zdrahal. 2005. Spotlight browsing of resource archives. In *Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*. Salzburg, Austria: ACM.

Nussbaumer, Philipp, and Bernhard Haslhofer. 2007a. CIDOC CRM in Action - Experiences and Challenges. In *11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL07)*, edited by László Kovács, Norbert Fuhr and Carlo Meghini. Budapest, Hungary: Springer Berlin / Heidelberg.

Nussbaumer, Philipp, and Bernhard Haslhofer 2007b. Putting the CIDOC CRM into Practice - Experiences and Challenges.

Nussbaumer, Philipp, Bernhard Haslhofer, and Wolfgang Klas. 2010. Towards Model Implementation Guidelines for the CIDOC Conceptual Reference Model. Technical Report June 2010 TR-20100603. Vienna: Universität Wien. Department of Distributed and Multimedia Systems.

Pasin, Michele, and Enrico Motta. 2009. "Ontological requirements for annotation and navigation of philosophical resources." *Synthese* no. 182 (2): 235–267. doi: 10.1007/s11229-009-9660-3.

Pasin, Michele, Enrico Motta, and Zdenek Zdrahal. 2007. Capturing knowledge about philosophy. In *K-CAP 2007, Whistler, BC, 28-31 October 2007*.

Riva, Pat, Martin Doerr, and Maja Žumer. 2009. "FRBRoo: enabling a common view of information from memory institutions." *ICBC* no. 38 (2): 30–34.

Stead, Stephen. 2008. *Tutorial for ISO-21127*: CIDOC CRM.

Theodoridou, M., Y. Tzitzikas, M. Doerr, Y. Marketakis, and V. Melessanakis. 2010. "Modeling and querying provenance by extending CIDOC CRM." *Distributed and Parallel Databases* no. 27 (2): 169-210. doi: 10.1007/s10619-009-7059-2.

Theodoridou, Maria, and Martin Doerr. 2001. Mapping of the Encoded Archival Description DTD element set to the CIDOC CRM. In *ICS-FORTH, June 2001*.

Vlachidis, A., C. Binding, D. Tudhope, and K. May. 2010. "Excavating grey literature. A case study on the rich indexing of archaeological documents via natural language-processing techniques and knowledge-based resources." *Aslib Proceedings* no. 62 (4-5):466-475. doi: 10.1108/00012531011074708.

Øyvind, Eide. 2008. "The exhibition problem. A real-life example with a suggested solution." *Literary and Linguistic Computing* no. 23 (1):27–37.