SCIRES-IT
SCIentific RESearch and Information Technology
Ricerca Scientifica e Tecnologie dell'Informazione
Vol 1, Issue 2 (2011), 87-128
e-ISSN 2239-4303, DOI 10.2423/i22394303v1n2p87

UN APPROCCIO "SOCIALE" E ONTOLOGICO ALLA CATALOGAZIONE

Oreste Signore[1]

[1] C.N.R e Ufficio Italiano W3C - CNR-ISTI - Area della Ricerca di Pisa, Pisa, Italia (oreste.signore@gmail.com)

Abstract

La catalogazione è basata sulla conoscenza e sulla comprensione dei legami interdisciplinari tra gli elementi informativi. L'avvento del Web ha modificato profondamente i meccanismi di accesso all'informazione, mentre Web 2.0 e Semantic Web sono due evoluzioni del Web che stanno rivoluzionando i canoni di gestione del patrimonio informativo. Il primo spostando l'accento sul contributo portato dagli utenti, il secondo enfatizzando il ruolo dei dati e dell'interconnessione tra fonti diverse di conoscenza. Il web ha fatto prestare maggiore attenzione all'interoperabilità semantica e ha portato all'adozione di schemi comuni di metadati, mentre il Web 2.0 ha spostato l'accento sull'aspetto "sociale" e sulle folksonomy. Entrambi questi approcci presentano delle limitazioni intrinseche, che vengono superate dall'adesione a principi del Semantic Web e quindi all'adozione di un approccio ontologico, con il superamento della catalogazione centrata sull'oggetto. È possibile coniugare i vantaggi dell'approccio "sociale" (partecipazione degli utenti, ampliamento delle fonti di informazione) con quelli puramente ontologici (formalizzazione e condivisione della conoscenza, possibilità di dedurre nuova conoscenza) mediante un'opportuna opera di mapping della scheda catalografica e l'implementazione di ambienti sociali che utilizzino le tecnologie del Semantic Web. Un'esperienza in merito è stata condotta in un progetto.

1. Catalogazione e conoscenza: il modello italiano

Il principio guida della catalogazione è la *conoscenza*, sia riguardo allo specifico oggetto che a tutti gli altri elementi che possono aiutare a comprendere le complesse relazioni che lo legano ad altre discipline. In alcuni casi solo la completa conoscenza del contesto storico, politico e religioso può far comprendere il valore e il messaggio di un oggetto. Tuttavia, questa conoscenza è quasi sempre appannaggio degli studiosi e degli esperti, e raramente viene comunicata agli altri. Di conseguenza, molti di noi possono percepire solo in maniera ridotta il reale valore e il significato di un oggetto facente parte del patrimonio storico, culturale e ambientale. Chiunque abbia mai avuto la fortuna di visitare un museo in compagnia di uno studioso[1], che può comunicare i dettagli del contesto culturale in cui è nato l'oggetto può comprendere la differenza rispetto ad una normale e frettolosa visita da turista.

Il modello[2] adottato dall'Istituto Centrale per il Catalogo e la Documentazione (ICCD) come standard nazionale a metà degli anni 1980, e base per la catalogazione elettronica, a prima vista può apparire semplicemente un altro schema "piatto", strutturato in un sistema di

[1] Personalmente ebbi occasione di visitare il Museo del Louvre, in un giorno di chiusura, in compagnia di Oreste Ferrari, all'epoca Direttore dell'ICCD, con il quale ho avuto il piacere e l'onore di collaborare.

[2] Le line guida seguite nella definizione dello standard sono riportate in [PAPALDO1986] e [SIGNORE1986]. Vedi anche [SIGNORE2009].

paragrafi (insieme di campi), campi e sottocampi, ordinati gerarchicamente e regolati da vincoli di obbligatorietà (assolute e di contesto), ripetibilità (applicabili a singoli sottocampi, campi o ad interi paragrafi), uso di vocabolari controllati (aperti o chiusi).

Tuttavia, chiunque sia familiare con le metodologie di progettazione di basi di dati potrà facilmente comprendere come, in termini generali, le entità siano state modellate come paragrafi, gli attributi (eventualmente multivalore) in campi (eventualmente ripetibili), gli attributi aggregati in campi strutturati. Appare anche evidente come l'identificazione di una sequenza di campi, con la caratteristica di essere ripetibile (eventualmente per gruppi di campi) e i riferimenti ad "authority files" può essere vista come la "linearizzazione" di un testo non lineare. In altri termini, come una lettura del modello concettuale in cui, partendo dall'oggetto, si percorrono tutti i cammini presenti nel modello. Va inoltre rilevato uno sforzo significativo nel mantenere una certa coerenza tra le varie aree disciplinari, grazie all'utilizzo dello stesso nome per campi semanticamente equivalenti.

Il modello costituiva un compromesso tra i requisiti posti dal mondo accademico e dalle comunità di ricerca, che richiedevano informazioni esaustive e dettagliate, e le esigenze dettate dalle necessità amministrative di disporre di un modello sufficientemente generale in grado di adattarsi a una ampia varietà di oggetti, astraendo dalle loro differenze. Il modello proposto può essere considerato anche un modo per *rappresentare la conoscenza*, considerato che all'epoca non per tutti i campi erano disponibili thesauri o "authority files", e l'uso di campi strutturati risultò un modo per rappresentare la conoscenza oggi verrebbe modellata con thesauri multidimensionali (faceted thesauri). Il modello di oggetto poteva anche essere visto come un *modello interoperabile per raccogliere informazioni*.

Il modello prestava grande attenzione alla definizione di diversi tipi di relazioni tra gli oggetti, portando così alla definizione di tre tipi di oggetto: semplice, complesso e aggregazione di oggetti, supportando anche dei meccanismi di ereditarietà tra i componenti (madre/figlia) di un oggetto complesso. In sostanza, venivano identificate delle relazioni di tipo "verticale" tra i componenti di un oggetto complesso, e di tipo "orizzontale" per l'aggregazione di oggetti. In seguito fu deciso di consentire la codifica esplicita di relazioni semanticamente più ricche tra oggetti. A parte alcuni aggiornamenti di dettaglio, il modello messo a punto a metà degli anni 1980 è quello tuttora in uso, anche se è opportuno segnalare due elementi critici: la percezione di uno schema rigido, con una eccessiva frammentazione dell' informazione, destinato evidentemente alla compilazione e fruizione da parte di esperti

del settore, e la complessità e lentezza di aggiornamento di vocabolari controllati, authority files e thesauri.

Più importante, però, è menzionare alcuni limiti intrinseci, determinati dall'approccio adottato per la catalogazione. Per prima cosa l'approccio restava ancorato alla compilazione di "una scheda per ogni oggetto", anche se l'oggetto veniva modellato in maniera più ricca di quella usuale. Questa visione "oggettocentrica" è all'origine della codifica ridondante, e potenzialmente incoerente, di informazioni (ad es. l'autore) e non consente di rappresentare adeguatamente relazioni semanticamente ricche.

Va evidenziato che queste limitazioni non derivano dal progetto originale, che era basato sulle metodologie di progettazione concettuale di basi di dati, ma dalla scelta di rappresentare le informazioni sotto forma di "scheda di catalogazione", che enfatizza il ruolo dell'oggetto di classificazione, ma impone una linearizzazione della visita dello schema concettuale, con la possibilità di esprimere solo le relazioni binarie tra l'oggetto e le altre entità, e non consente, mantenendo una complessità ragionevole, di esprimere le relazioni interdisciplinari. Di conseguenza, la conoscenza dell'esperto non viene espressa in modo formale, e resta nascosta all'utente non specialista, mentre solo gli esperti, grazie alle loro competenze e conoscenze, possono individuare i collegamenti con le altre discipline, ponendo l'oggetto nel suo contesto storico e culturale.

2. Web, Web 2.0 e Semantic Web

2.1 La "Web revolution"

Il Web è esploso a metà degli anni 1990, ed è stato all'origine di una autentica rivoluzione dei metodi tradizionali di accesso all'informazione [COYLE 2007]. Nel passato gli utenti accedevano all'informazione partendo da fonti "ufficiali" come biblioteche, cataloghi di musei, e simili, mentre adesso essi partono quasi sempre da una query generica sul web, e poi seguono i link, alla ricerca delle informazioni rilevanti. Di conseguenza, il ruolo dei siti centrali e istituzionali è fortemente ridimensionato, perché l'architettura del web è intrinsecamente decentralizzata, e assumono importanza due aspetti: l'interoperabilità tecnica, che è garantita dai protocolli del web, e *l'interoperabilità semantica*, che dovrebbe consentire di combinare conoscenza disponibile su siti diversi. Quest'ultimo è senz'altro

l'aspetto più rilevante, in quanto richiede di poter rappresentare, esportare e condividere la conoscenza.

In questo panorama si inserisce poi l'affermarsi di una nuova modalità di interazione, con la partecipazione attiva degli utenti alla creazione di contenuti sul web.

2.2 Il Web 2.0

Non esiste una definizione precisa e univoca di Web 2.0. Il termine è stato coniato da Tim O'Reilly [O'REILLY2005], il fondatore della nota casa editrice scientifica, che ha indicato alcune caratteristiche essenziali di una applicazione Web 2.0.

L'interesse, o l'avvento, del Web 2.0 rendono particolarmente importanti due fatti: il primo è che la fornitura di un elenco di risultati (siano essi libri, oggetti d' arte, hotel, o altro) non costituisce più l'obiettivo primario e unico del servizio, qualunque esso sia. Il secondo è che la filosofia del Web 2.0 enfatizza gli aspetti sociali dell'informazione, e quindi accresce l'importanza di commenti, valutazioni, suggerimenti, e tagging[3]. [COYLE2007]

Il Web 2.0 nasce quindi ufficialmente con la definizione di O'Reilly, ma il suo modello è sempre stato presente nel web fin dalla sua nascita. Due caratteristiche essenziali del Web 2.0 sono il collaborative tagging e le communities.

Per *collaborative tagging* (spesso denominato anche *folksonomy*, *social classification*, *social indexing*) si intende la prassi e il metodo di creare e gestire in maniera collaborativa i tag per annotare e caratterizzare i contenuti. Contrariamente a quanto avviene con la soggettazione tradizionale (assegnazione di parole chiave), la generazione dei metadati non è più un compito demandato esclusivamente ad un gruppo di esperti, ma si avvale del contributo degli utenti e dei fruitori dei contenuti. In questo processo vengono adottate parole chiave (o *keyword*) libere, invece di ricorrere ad un vocabolario controllato. L'obiettivo del *folksonomic tagging* è rendere sempre più facile ricercare, scoprire e navigare un insieme di informazioni. Una folksonomy ben sviluppata dovrebbe essere consultabile come un vocabolario condiviso sviluppato e ben conosciuto dai suoi utenti principali. L'aspetto più interessante delle folksonomy risiede forse proprio in questa loro inerente inversione di tendenza (o, potremmo dire, di radicale sovvertimento): rispetto ai meccanismi tradizionali di ricerca dei contenuti web mediante i search engine, in favore di strumenti creati dalla comunità.

[3] Un tag è una parola chiave (significativa) associata o assegnata ad un elemento informativo (una foto, una mappa geografica, una voce di un blog, un video clip, etc.), che descrive quell'elemento informativo e ne consente la classificazione e la ricerca.

Secondo la definizione del Devoto-Oli, una comunità è un "*insieme di persone unite tra di loro da rapporti sociali, linguistici e morali, vincoli organizzativi, interessi e consuetudini comun*i". Una comunità[4], quindi, è un gruppo sociale di organismi, che condividono un ambiente, e hanno normalmente interessi comuni.

2.3 Il Semantic Web e le ontologie

Il Web è nato per consentire la condivisione di documenti. Il Semantic Web si pone l'obiettivo di consentire agli utenti di rendere disponibili i loro dati agli altri, e di aggiungere dei link per renderli accessibili seguendo i link. Da questo punto di vista, il Semantic Web è un'estensione dei principi del Web dai documenti ai dati e fornisce una infrastruttura comune che consente la condivisione e il riutilizzo dei dati tra applicazioni, imprese e comunità. Il Semantic Web talvolta viene chiamato "*web of data*" proprio per sottolineare il fatto che sul web è disponibile una enorme quantità di dati, ma, dato che questi sono controllati dai programmi applicativi, e ogni applicazione li gestisce in maniera diversa, non è possibile combinare le varie informazioni, se non con un intervento manuale.

L'ipotesi di base è che le macchine possano accedere ad un *insieme strutturato di informazioni* e ad un *insieme di regole di inferenza* da utilizzare per il ragionamento automatico. Occorre quindi un linguaggio per esprimere *dati* e *regole* per ragionare sui dati, che consenta l'*esportazione* sul web delle regole da qualunque sistema di rappresentazione della conoscenza, con l'obiettivo di consentire alle macchine di estrarre la conoscenza disponibile sul Web, spesso disponibile in formati eterogenei, e combinarla per poter estrarre nuova conoscenza. Questo processo è possibile solo se si riesce a rappresentare, esportare e condividere la conoscenza, mediante una rappresentazione indipendente dallo specifico ambiente operativo. Le tecnologie del W3C (RDF, RDFS, OWL, SPARQL, etc.) consentono appunto di raggiungere questo obiettivo. È importante sottolineare che *condividere i dati e la conoscenza non comporta tradurre tutte le informazioni in RDF*.

Nell'architettura del Semantic Web[5] riveste un ruolo fondamentale il *Resource Description Framework* (RDF) che è lo strumento base per la codifica, lo scambio e il riutilizzo di metadati strutturati, e consente l'elaborazione automatica delle risorse reperibili sul Web, e

[4] Tutta la letteratura riguardo alle comunità fa riferimento ad un lavoro fondamentale scritto nel 1986 da McMillan and Chavis [MCMILLAN1986].

[5] Per una presentazione più ampia si rimanda a [SIGNORE2008] e bibliografia ivi citata. Un'ottima descrizione delle tecnologie e dei principi del Semantic Web è in [ANTONIOU2004].

quindi l'interoperabilità tra applicazioni che si scambiano sul Web informazioni *machine-understandable*.

Le asserzioni RDF sono espresse come triple (soggetto-predicato-oggetto) in cui le risorse (che possono comparire sia come soggetto che come oggetto) vengono identificate come nodi (graficamente delle ellissi), le proprietà come archi orientati etichettati, e i valori corrispondenti a sequenze di caratteri come rettangoli. Sia le risorse che le proprietà sono identificate univocamente da URI. Una asserzione RDF può far riferimento a proprietà e risorse che sono definite in un qualunque punto del Web, senza necessità di centralizzare le informazioni. In Figura 1 la rappresentazione grafica di una serie di proposizioni RDF, di ovvia comprensione, che fanno riferimento a risorse distribuite sul Web.

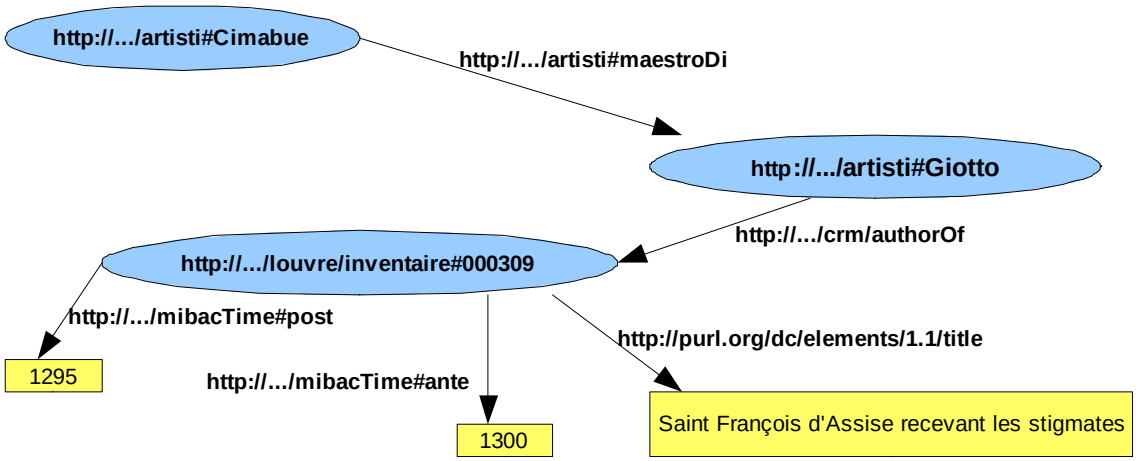

Figura 1: Rappresentazione grafica di asserzioni RDF.

Per esprimere le restrizioni sulle associazioni, quindi per evitare che possano essere formulate asserzioni sintatticamente corrette, ma prive di senso, è necessario un meccanismo per rappresentare "classi di oggetti". Da questa esigenza nasce "*RDF Vocabulary Description Language*", che mantiene anche, per ragioni storiche, il nome di "RDF Schema" (RDFS).

Per poter effettuare dei ragionamenti, per definire le classi, e per varie altre esigenze, però, RDFS non è sufficiente, e occorre un modo per rappresentare la conoscenza e le regole che permettono di dedurre ulteriore conoscenza: l'*ontologia*. Dato che il Web è intrinsecamente distribuito, occorre un linguaggio che non solo permetta di esprimere dati e regole sui dati, ma che consenta anche di esportare questa conoscenza (ontologia) per renderla disponibile a qualunque applicazione. Il W3C ha definito, per questa esigenza, il *Web Ontology Language* (OWL).

Il termine ontologia deriva dalla filosofia, dove viene inteso come una spiegazione sistematica dell'essere. Negli anni recenti il termine si è ampiamente diffuso nella comunità del Knowledge Engineering. Esistono diverse definizioni di ontologia[6], ognuna delle quali enfatizza qualche aspetto. Va posto l'accento su come un'ontologia includa non solo i termini che sono esplicitamente definiti in essa, ma anche la conoscenza che ne può essere *derivata* mediante un processo di inferenza. Inoltre, va sottolineato come le ontologie mirino a catturare la conoscenza *consensuale*, e possano essere condivise e riutilizzate tra applicazioni e gruppi di persone diversi.

Le ontologie possono essere *molto informali*, *semi-informali*, *semi-formali* o *rigorosamente formali* a seconda che siano espresse in linguaggio naturale, in linguaggio naturale ristretto, in un linguaggio artificiale e definito formalmente, o fornendo una descrizione meticolosa dei termini, utilizzando una semantica formale, teoremi e dimostrazioni di proprietà. Nella classificazione delle ontologie basata sulla ricchezza della loro struttura interna, i vocabolari controllati e i thesauri si collocano nella parte bassa e media delle ontologie informali, mentre le ontologie in cui vengono espressi dei vincoli sui possibili valori si collocano nella parte alta delle ontologie formali [DIGICULT 2003]. È opportuno sottolineare che i thesauri supportano alcuni requisiti di rappresentazione della conoscenza, ma non possono essere immediatamente e automaticamente trasformati in ontologie. Per esempio, nei thesauri la relazione NT in alcuni casi modella una relazione di sottoclasse come nel caso di "statue" e "korai (statue)", mentre in altri casi modella semplicemente una relazione tra istanze diverse della stessa classe, come nel caso in cui "Rinascimento" viene modellato come BT di "XV secolo", mentre è evidente che si tratta semplicemente di due periodi temporali (quindi istanze di una classe "periodo temporale", con proprietà definite come "data di inizio" e "data di fine") che presentano una certa sovrapposizione. L'ereditarietà multipla e le relazioni dipendenti dal tempo costituiscono anch'esse dei problemi non completamente risolti o risolubili.

3. Web 2.0 o Web 3.0?

6 Una delle più usate è quella data in [STUDER1998]:
"An ontology is a formal, explicit specification of a shared conceptualisation. A '*conceptualisation*' refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. '*Explicit*' means that the type of concepts used, and the constraints on their use are explicitly defined. For example, in medical domains, the concepts are diseases and symptoms, the relations between them are causal and a constraint is that a disease cannot cause itself. '*Formal*'refers to the fact that the ontology should be machine readable, which excludes natural language. '*Shared*' reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group."

La comunità interessata al patrimonio culturale (storico, artistico, paesaggistico, archivistico, etc.) non è ovviamente rimasta immune dal "contagio" del Web 2.0. La rapida diffusione di queste tecnologie nella comunità culturale mostra chiaramente che è proprio l'aspetto della socialità quello che è più rilevante. Va però tenuto presente che la comunità culturale presenta requisiti più stringenti, in cui il social software dovrebbe supportare in maniera più efficace i professionisti nella ricerca, collaborazione e comunicazione all'interno della comunità. In altri termini, contribuire più efficacemente alla condivisione della conoscenza[7]. [CAO2006]

[COYLE2007] descrive molto bene le caratteristiche dell'utente tipico nel mondo 2.0. Con la diffusione del Web 2.0 gli utenti sono ormai abituati a contribuire ai contenuti Web, con annotazioni, commenti, etc. Gli utenti si aspettano di trovare una comunità con cui interagire, e questo crea un grosso problema di cambio di mentalità nei gestori dell'informazione. In un mondo in cui le informazioni del catalogo erano inseribili e aggiornabili solo da un ristretto numero di specialisti, che seguivano regole talvolta anche molto complesse, e non era possibile neanche inserire commenti od osservazioni, questa possibilità costituisce un' autentica rivoluzione[8].

Le esperienze condotte finora, anche nell'ambito dei beni culturali, hanno evidenziato diversi vantaggi derivanti dall'adozione dell'approccio Web 2.0, essenzialmente legati all'aggiornamento continuo dei contenuti a cura degli utenti, al miglioramento della qualità dei tag e alla crescita di un senso "sociale".

Tuttavia, i critici del contenuto generato dagli utenti segnalano il venir meno delle fonti tradizionali del sapere, come l'autorevolezza e il contributo degli studiosi, e segnalano come il passaggio dal mondo della carta a quello assai più dinamico, ma anche volatile, dell'informazione elettronica, determini una perdita di credibilità e fiducia nelle informazioni. Un ulteriore elemento da non trascurare è il fatto che il contenuto viene creato dagli utenti utilizzando sistemi diversi (podcast, blog, wiki, sistemi di chat, e altro software per il social networking), rendendo difficile tener traccia di dove si trovi l'informazione, e problematico accedere ad essa, sia per utenti abituali che casuali.

7 Gli utenti trovano comodo condividere le loro risorse informative e combinarle con quelle di altri utenti, e si aspettano anche che le loro risorse informative interagiscano tra di loro. A titolo di esempio, i plug-in OpenURL (http://www.openly.com/openurlref/) permettono di muoversi, senza soluzione di continuità, da una citazione sul web a una copia dell'articolo, non disponibile liberamente sul web.

8 Si noti, tuttavia, che questa è una restrizione nata con il diffondersi della gestione elettronica dell'informazione. Molti studiosi possono confermare che spesso nel consultare fototeche o altri sistemi cartacei di documentazione il valore delle annotazioni a margine o sul retro delle fotografie era superiore a quello della documentazione "ufficiale".

Infine, [HARDMAN2008] il Web 2.0 presenta alcuni rischi intrinseci. Proprio il fatto di contribuire in maniera così consistente al contenuto di un sito, con un evidente impegno e dispendio di energie, rende l'utente "dipendente" dal sito medesimo. L'utente è legato definitivamente al formato dati adottato, esattamente come accade nel caso in cui si utilizzino programmi proprietari. Eventuali cambi di ambiente saranno inevitabilmente onerosi. Questo è un esempio tipico della legge di Metcalfe [METCALFE1995], secondo la quale il valore di una rete è proporzionale al quadrato del numero di nodi presenti. Ne consegue, evidentemente, che se si divide una rete in due parti uguali, il valore di ognuna di esse è un quarto dell'originale, e il valore totale la metà[9].
L'esistenza di un unico World Wide Web è un fattore positivo, mentre il Web 2.0 suddivide il Web in una serie di sub-Web dedicati ad aspetti specifici, con il risultato di legare gli utenti a formati particolari, e diminuendo il valore della rete nel suo complesso.

Web 2.0 e Semantic Web (o Web 3.0) sono da considerare due approcci complementari, piuttosto che alternativi. Il Web 2.0 ha un livello d' ingresso più basso (è molto facile utilizzarlo), ma anche orizzonti abbastanza limitati (in particolare, l'approccio delle folksonomy ha dei limiti intrinseci). D' altro canto, il Web 3.0 richiede investimenti iniziali più rilevanti, ma ha un potenziale nettamente superiore.
Se il Web 2.0 partiziona il Web in una serie di sub-Web, inficiando così il valore della rete nel suo complesso, il Semantic Web, al contrario, offre il vantaggio di poter avere i dati distribuiti sul Web con un approccio che esclude qualsiasi centralizzazione.
La semplice aggiunta di metadati ai dati consente la realizzazione di servizi di aggregazione dell'informazione che offrono gli stessi vantaggi del Web 2.0, senza incorrere nei rischi conseguenti all'adozione di formati proprietari o addirittura di perdita di dati nel caso in cui il servizio su cui si fa affidamento scompaia.

3.1 Metadati: vocabolari controllati, tag e folksonomy

In estrema sintesi i metadati vengono assegnati alle risorse da professionisti, dagli autori o dagli utenti. Tipicamente, nei primi due casi i metadati vengono estratti da "*vocabolari controllati*", e lo scopo è che gli utenti possano utilizzare questi termini per accedere al contenuto del sito, mentre i tag assegnati dagli utenti sono estratti da un vocabolario libero.

9 Non mancano tuttavia critiche alla validità di questa legge (vedi [Briscoe2006])

I vocabolari controllati offrono il vantaggio di poter utilizzare termini non ambigui, con un' organizzazione gerarchica dei concetti. Tuttavia accade frequentemente che, per supportare meglio la ricerca da parte degli utenti, che non necessariamente condividono il vocabolario messo a punto dagli esperti, si debba assegnare un termine a più categorie, con una forzatura delle regole tassonomiche. Inoltre i vocabolari controllati e le strutture gerarchiche vengono creati *ex ante*, sulla base di quello che gli "Information Architect" prevedono sarà più utile agli utenti. Quando si aggiungono contenuti al sito può accadere che non esista una categoria adeguata per contenenre i nuovi termini necessari per indicizzarli. È ovviamente sempre possibile aggiungere nuove categorie, con un processo generalmente costoso e fonte di forzature. La comparsa di nuovi termini non è un fatto raro, ed è evidente che il processo top-down nell'organizzazione dell'informazione sulla base di vocabolari controllati non è in grado di prevedere questi cambiamenti e far fronte alle mutate esigenze dell'utenza (che sono un fatto ineludibile) e che siti basati su vocabolari controllati tendono a non adeguarsi ai cambiamenti.
D'altro canto, indipendentemente da alcune limitazioni, derivanti più che altro dall'assenza delle caratteristiche dei vocabolari controllati, non si può ignorare l'importanza del "*social tagging*" ed è da prevedere la sua coesistenza con i vocabolari controllati.

Il social tagging è la base delle *folksonomy*, che sono semplicemente l'insieme dei termini che una comunità di utenti ha utilizzato per assegnare dei tag alle risorse. L'aspetto rilevante delle folksonomy consiste proprio nell'essere uno *spazio uniforme di nomi*, senza gerarchie o legami di qualunque genere tra di essi. In alcuni casi vengono generate delle relazioni tra i tag, sulla base del fatto che fanno riferimento allo stesso URI, ma siamo ben lontani dalle relazioni definite in una tassonomia formale o in thesaurus. Proprio per la loro natura, le folksonomy presentano alcuni punti di debolezza, che si possono identificare nell'*ambiguità* dei termini (utilizzo semanticamente diverso da parte degli utenti), *sinonimie*, mancata gestione degli *spazi* o delle *parole multiple* (i tag sono stati pensati come parole singole, e spesso non viene operata distinzione tra lettere maiuscole e minuscole). I vocabolari controllati sono nati proprio per risolvere questi problemi.
Nonostante una folksonomy non sia un vocabolario controllato, e quindi presenti senza dubbio diverse limitazioni, vi sono alcuni punti di forza importanti per comprendere l'utilità e il potere di attrazione di questi sistemi. Le folksonomy sono in continua evoluzione, e questo ne costituisce forse l'aspetto fondamentale. Quando un utente aggiunge un tag, l'effetto della

sua azione è immediatamente percepibile a tutti gli utenti che accedono a quel contenuto, rendendo loro possibile una nuova navigazione. A poco a poco emergono i tag più comuni, che creano ancora maggiori potenzialità di navigazione, favorendo il rinvenimento di contenuti che altrimenti non sarebbero mai stati trovati. Questo fenomeno è noto come "*serendipity*".

Il punto di forza più significativo di una folksonomy è che rispecchia direttamente il vocabolario degli utenti. In un sistema di information retrieval esistono vari vocabolari (utente, progettista del sistema, autore dei contenuti, creatori dello schema di classificazione) e la traduzione dall'uno all'altro è un compito difficile. Una folksonomy costituisce invece un vero cambio di prospettiva, perché il suo contenuto è determinato non dai professionisti o dai curatori del sistema, ma direttamente dagli utilizzatori, e ne riflette le preferenze in termini di dizione, terminologia e precisione, diventando così uno specchio di quanto la comunità ritiene rilevante.

Non ci sono costi significativi, né per l'utente né per il sistema, nell'aggiungere un termine ad una folksonomy. Il vero problema è che se è vero che la presenza di vocabolari e termini definiti dagli utenti favorisce il browsing e il reperimento di informazioni, proprio questa totale libertà può distruggere i contenuti per la presenza di metadati poco utili o non rilevanti per gli utenti.

3.2 Metadati: Dublin Core

Molte applicazioni nel settore del patrimonio culturale utilizzano come standard di metadati il Dublin Core. [BAKER2000]

Dublin Core è presentato spesso come una forma moderna di scheda di catalogo (catalog card), cioè un insieme di elementi (*element set*) eventualmente qualificati (*qualified element set*) che descrivono in maniera completa una risorsa. Talvolta viene anche proposto come formato di scambio per condividere le risorse tra varie collezioni. Il principio di base è: "ogni elemento è opzionale e ripetibile". In termini tecnici, un elemento o un qualificatore Dublin Core è un identificatore univoco formato da un nome (es. creator) con un prefisso formato dall' URI del namespace nel quale è stato definito, ad esempio: http://dublincore.org/documents/dces/#creator.

In questo contesto un namespace è un vocabolario pubblicato formalmente, tipicamente sul Web, che descrive gli elementi e i qualificatori con etichette in linguaggio naturale, definizioni e altra documentazione pertinente. I quindici elementi del Dublin Core element

set definiscono le caratteristiche di Dublin Core come linguaggio. Nella loro forma abbreviata gli elementi sono: dc:title, dc:creator, dc:subject, dc:description, dc:publisher, dc:contributor, dc:date, dc:type, dc:format, dc:identifier, dc:source, dc:language, dc:relation, dc:coverage, e dc:rights. Essi corrispondono a quindici proprietà molto generali, utili per ricercare documenti in archivi pertinenti a domini diversi. In effetti, Dublin Core è una classe di proposizioni (statement) del tipo "Resource has property X," dove "*resource*" è il soggetto implicito, seguito da un verbo implicito ("*has*"), seguito da una delle quindici proprietà appartenenti al Dublin Core element set, e seguita infine da un valore per la proprietà, specificato sotto forma di sequenza di caratteri come un nome di persona, una data, una serie di parole o un URI. Per esempio: Resource has dc:creator 'Oreste Signore', e Resource has dc:date '2009-04-01'. È possibile utilizzare i qualificatori per dare un significato più preciso alle proprietà.

3.3 Ontologie o folksonomy?

Una folksonomy rappresenta simultaneamente il meglio ed il peggio nell'organizzazione dell'informazione. Essendo per sua natura non controllata, è sostanzialmente caotica, e presenta notevoli problemi di ambiguità e scarsa precisione, che invece migliora sensibilmente se si utilizzano vocabolari controllati. D'altra parte, sistemi che utilizzano un tagging libero, che incoraggia gli utenti ad organizzare il materiale a loro piacimento, si adeguano molto rapidamente alle esigenze degli utenti e ai loro vocabolari, e favoriscono il loro coinvolgimento nell'organizzazione dell'informazione.

Si va diffondendo l'idea che le folksonomy siano da preferire rispetto all'uso di ontologie, perché queste richiedono investimenti significativi, mentre il tagging è informale e semplice.

Ontologie e folksonomy sono state spesso considerate due approcci opposti, mentre in realtà sono due concetti separati, anche se alcune funzionalità delle ontologie possono essere supportate, in molti contesti, dalle folksonomy. In sostanza, folksonomy e ontologie sono diverse perché le prime sono una variante della ricerca di informazioni basata su parole chiave, mentre le seconde mirano a disciplinare il mondo dei dati, e rendere possibile il mapping e l'interazione tra dati memorizzati in formati diversi e in punti diversi, o raccolti da organizzazioni diverse per le loro specifiche finalità. Si tratta quindi di due approcci differenti a due tipi di problemi, che in qualche caso possono presentare delle aree di sovrapposizione, rendendo così possibile optare per l'una o l'altra soluzione.

Le ontologie formali possono essere utilizzate per migliorare il social tagging, mettendo a disposizione insiemi di tag più significativi, e supportando il tagging a faccette ([XU2006], [QUINTARELLI2007]).
Va però sottolineato che *una folksonomy non diventerà mai un' ontologia*, perché nasce come spazio uniforme di nomi, e non come organizzazione di concetti.

4. I vantaggi di un approccio ontologico

L'importanza dell'interoperabilità semantica è ampiamente riconosciuta dagli studiosi, e molti progetti internazionali hanno raggiunto un accordo su vocabolari comuni per i metadati, facendo quasi sempre riferimento allo schema di metadati Dublin Core. Questo costituisce un passo in avanti rispetto all'enfasi posta nel passato su XML come meccanismo di strutturazione dei dati. In effetti, XML è semanticamente povero, mentre le tecnologie del Semantic Web (in particolare RDF e OWL) possono fornire l'ambiente adatto per rappresentare, esportare e condividere la conoscenza, rendendo così possibile l'implementazione di sistemi intelligenti di reperimento e browsing, in grado anche di operare "ragionamenti". Nell'architettura peer-to-peer che è una delle caratteristiche del Web, le tecnologie del Semantic Web consentono la marcatura ("markup") dei contenuti, utilizzando classi e proprietà definite in una ontologia condivisa (es. CIDOC CRM), e gli agenti software possono utilizzare in modo appropriato questa conoscenza.
Considerando l'evoluzione della catalogazione e condivisione dell'informazione pertinente al patrimonio culturale, si può vedere chiaramente come si sia passati da uno stadio iniziale, in cui l'informazione veniva registrata in maniera informale, ad una maggiore strutturazione dell'informazione, mentre negli anni più recenti sono comparsi vari progetti che fanno riferimento a un insieme comune di metadati (Dublin Core o Qualified Dublin Core). Vi sono poi alcuni progetti più innovativi [HYVÖNEN 2004] che si basano su ontologie, essenzialmente come insieme di termini in relazione tra di loro, da utilizzare per formulare query più precise.
Considerato che, peraltro, esiste generalmente un accordo su un insieme comune di metadati, ci si potrebbe legittimamente chiedere: *perché adottare un approccio ontologico*?
Per prima cosa, come puntualizzato in [DOERR2003], è vero che sia dei metadati di base (come Dublin Core) che un'ontologia di base sono finalizzati all'integrazione

dell'informazione, ma differiscono in maniera sostanziale per l'importanza attribuita alla *comprensibilità da parte di esseri umani*. I metadati sono pensati per una elaborazione da parte di esseri umani, mentre un'ontologia è un modello formale utilizzabile da strumenti che integrano varie fonti di dati e svolgono varie funzioni. Su vocabolari basati su ontologie, che organizzano i termini in maniera che essi abbiano una semantica chiara ed esplicita, è possibile svolgere dei "ragionamenti", cioè svolgere un processo fondamentale di arricchimento della conoscenza, deducendo nuova informazione relativamente alle risorse.

In secondo luogo, esiste un preciso limite nella assunzione che è alla base dell'approccio basato su metadati. Aggiungere metadati alla descrizione di un oggetto significa assumere implicitamente una relazione uno-a-molti (o eventualmente molti-a-molti) tra l'oggetto e gli elementi identificati dai metadati.

Per esempio, codificando[10] alcuni metadati DC come:

dc:title=Pietà
dc:creator=Michelangelo
dc:date=1499
dc:subject=Madonna
dc:subject=Cristo

o

dc:title=Madonna del cardellino
dc:creator=Raffaello
dc:date=1505
dc:subject=Madonna
dc:subject=Bambino

intendiamo dire che un particolare oggetto (la Pietà) è stato realizzato da Michelangelo, è datato 1499, e ha come soggetto "Madonna" e "Cristo", mentre il secondo (il dipinto) è stato realizzato da Raffaello, è datato 1505, e ha come soggetto "Madonna" e "Bambino". Si possono aggiungere vocabolari controllati per essere sicuri che i termini specificati per "creator" o "subject" siano corretti, ma solo un essere umano può:

- verificare la coerenza tra dc:creator e dc:date, perché nessun artista può aver creato un'opera in data antecedente alla sua nascita (o prima che abbia raggiunto una certa età);
- una volta trovato un oggetto, trovare altri oggetti realizzati nello stesso periodo, o da artisti contemporanei;

10 Per semplicità non viene utilizzata una sintassi corretta, che richiederebbe espressioni del tipo:
<meta name="dc.creator" content="Michelangelo" /> o
<dc:creator>Michelangelo</dc:creator>

- avere informazioni sul contesto storico o politico (multidisciplinarità);
- trovare oggetti (ad esempio ritratti) immaginari, perché la scena è immaginaria, o i soggetti sono mitologici, oppure non erano in vita contemporaneamente o durante il periodo di attività dell'artista (ragionamento temporale).

Un passo importante per la rappresentazione della conoscenza è il *CIDOC object-oriented Conceptual Reference Model* ("CRM"), che è un'ontologia formale finalizzata a facilitare l'integrazione, la mediazione e lo scambio di informazioni tra fonti eterogenee di informazioni relative al patrimonio culturale. CRM è il risultato di oltre un decennio di lavoro per lo sviluppo di uno standard da parte dell'International Committee for Documentation (CIDOC) dell'International Council of Museums (ICOM), ed è ora un International Standard.

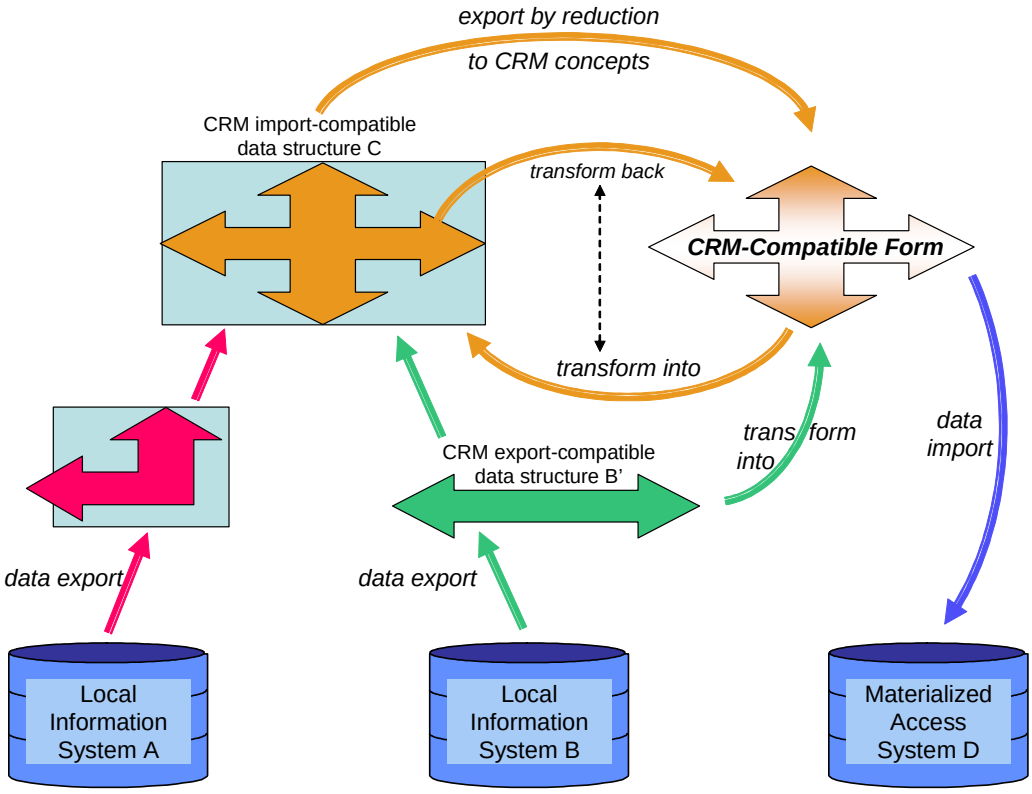

Figura 2: Possibili flussi tra diversi sistemi CRM-compatible
(da http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.2.doc).

Lo scopo primario di CRM è la specifica di definizioni semantiche che consentano di integrare archivi separati in un'unica risorsa informativa globale, distribuita su rete locale o geografica (Figura 2). CRM descrive la *semantica* su cui sono basati i database locali e le strutture dei documenti utilizzati nel settore della documentazione del patrimonio culturale,

ma *non definisce* uno *standard terminologico*, perché il suo obiettivo è consentire l'*interoperabilità semantica*.

CRM è un'ontologia costituita 81 classi e 132 proprietà, che descrive con un linguaggio formale i concetti e le relazioni rilevanti per la documentazione del patrimonio culturale. È stata concepita come ontologia formale con l'obiettivo specifico di coprire l'informazione contestuale, e può essere utilizzata per condurre ragionamenti spaziali e temporali. Per sua natura, e per il formalismo adottato, CRM è estensibile, per far fronte alle esigenze di comunità e applicazioni specializzate.

5. Un approccio nuovo alla catalogazione

L'approccio classico alla catalogazione presenta alcuni limiti. Per prima cosa, la catalogazione è stata sempre gestita come un'operazione condotta da esperti, e rivolta ad esperti. Mentre il primo aspetto ha delle giustificazioni reali, in quanto solo persone competenti possono portare un contributo credibile alla creazione della conoscenza che sottostà ad un catalogo inteso in senso compiuto, il risultato, cioè una serie di schede di catalogo compilate solo da persone autorizzate, e accettate solo dopo un processo di validazione complesso, se da un lato garantisce la qualità del risultato finale, dall'altro crea un corpus di informazioni spesso troppo sofisticato per il pubblico meno preparato. Inoltre, proprio per il tipo di approccio adottato, manca in generale un adeguato collegamento interdisciplinare, che renda l'informazione effettivamente fruibile e utile.

In un mondo pieno dei fermenti del Web 2.0, con tanta partecipazione sociale e produzione di contenuti da parte degli utenti, il catalogo tradizionale rischia di restare patrimonio di pochi eletti, e anche di perdere contributi significativi da parte della comunità. D'altra parte, un'apertura indiscriminata ai contributi degli utenti porterebbe fatalmente ad un abbassamento della qualità dell'informazione, anche se certamente arricchirebbe la valenza multidisciplinare del catalogo.

Per passare ad una catalogazione arricchita dai contributi degli utenti, senza perdere in qualità dell'informazione, occorre rendere la scheda di catalogo visibile e aperta ad eventuali annotazioni o aggiunte, e poi disciplinare in qualche modo i contributi degli utenti.

La scheda di catalogo può essere resa "aperta" grazie alle tecnologie del Semantic Web, senza rinunciare alle necessarie misure di protezione da accessi esterni e modifiche non

autorizzate. Il passo da compiere è semplicemente una sua trasposizione in RDF/OWL, o anche la sua esportazione mediante un'interfaccia in grado di "esporre" il suo contenuto in formato RDF. In ogni caso, è necessaria un'operazione di mapping del contenuto della scheda verso un'ontologia di riferimento. È evidente, in base a quanto detto precedentemente, che, una volta reso visibile il contenuto della scheda mediante il mapping, sarà possibile fare riferimento ad un qualunque suo elemento, che a questo punto risulta semanticamente definito, aggiungendo informazioni sotto forma di triple RDF. È interessante notare che questo approccio, basato sui concetti del Semantic Web, è molto meno invasivo di altri. Infatti, non viene in alcun modo modificato l'archivio delle schede, in quanto la tripla RDF è costituita da un soggetto, un predicato e un oggetto che possono risiedere in un qualunque punto del Web. L'unico vincolo è che le risorse individuate dagli URI che la costituiscono siano associate a URI permanenti[11]. In un approccio basato unicamente sull'aggiunta di metadati con lo standard Dublin Core, invece, la singola scheda dovrebbe essere modificata aggiungendo gli attributi, quindi con un intervento diretto sulla scheda stessa.

Appare quindi evidente la valenza e la ricchezza di un approccio ontologico: ove un utente riscontrasse l'esistenza di un collegamento tra informazioni reperibili su siti diversi (e individuabili univocamente mediante un URI) potrebbe inserire in modo formale questo collegamento, rendendo possibile l'ampliamento delle conoscenze. Un esempio banale potrebbe essere costituito dall'inserire un collegamento (con un *tipo* definito) tra un luogo citato nella scheda e le informazioni turistiche o storiche relative al luogo, o collegare il nome di un artista o di un personaggio citato nella scheda con archivi specializzati contenenti le loro biografie.

Un'altra possibilità per arricchire i contenuti è, come abbiamo già detto in precedenza, l'implementazione di un paradigma Web 2.0, quindi il supporto di tag e folksonomy. Anche in questo caso l'utilizzo di tecnologie semantiche può contribuire ad un netto miglioramento della qualità dell'informazione. Oltre a consentire agli utenti di inserire tag in formato libero, si può implementare un meccanismo più sofisticato, ma che offre notevoli vantaggi. In questo caso lo strumento del Semantic Web che offre nuove e interessanti possibilità è SKOS (Simple Knowledge Organization System) che è un formalismo standard per la rappresentazione della conoscenza espressa in thesauri, tassonomie e vocabolari. Se i tag vengono formalizzati usando il formalismo di SKOS, possono essere successivamente, ad

11 Ricordiamo che l'oggetto della tripla RDF può essere anche un "literal", cioè una sequenza di caratteri. Questa esigenza si presenta evidentemente quando l'oggetto della tripla non fa riferimento ad un vocabolario controllato o comunque a una lista di termini definiti.

opera di un esperto del dominio, raggruppati in una tassonomia o diventare termini che riferiscono concetti in un thesaurus, con il vantaggio che la conoscenza apportata dai singoli utenti viene organizzata e può essere resa disponibile ad altre applicazioni, senza il rischio che modifiche dell'applicazione o la scomparsa del servizio vanifichino il lavoro svolto.
La formalizzazione in SKOS dei tag può avvenire sia a posteriori, prendendo l'insieme dei tag come vocabolario dei termini preferiti dagli utenti, che a priori, consentendo agli utenti di selezionare i tag dal thesaurus (implementando quindi una navigazione sui termini), o di aggiungere nuovi tag in una posizione precisa del thesaurus, in questo caso con un intervento successivo da parte di un esperto, che validi il termine e la sua collocazione. L'approccio più probabile e presumibilmente più efficace è quello misto, in cui alcuni dizionari e thesauri, che possono essere arricchiti dal contributo degli utenti, vengono definiti a priori, mentre altri vengono creati a posteriori sulla base dei tag inseriti dagli utenti.
La possibilità di aggiungere tag e collegamenti è la base per la formazione di comunità specializzate. Vanno presi in considerazione meccanismi adeguati perché i contributi siano affidabili, e cresca il senso di partecipazione alla comunità e il livello dei contributi. A tal fine si possono utilizzare meccanismi "premianti", per cui gli utenti che aggiungono informazioni che poi vengono riconosciute valide da parte degli esperti o dell'amministratore di comunità possano passare ad un livello superiore, ed essere riconosciuti essi stessi come esperti.

6. Dalla scheda ICCD all'ontologia: i problemi del mapping

L'approccio catalografico adottato dall'ICCD nella formulazione della struttura delle schede suddivide l'informazione in una serie di campi, eventualmente strutturati in sottocampi e raggruppati in insiemi di campi (denominati paragrafi). Paragrafi, campi e sottocampi possono essere ripetibili, rispettivamente nel contesto dell'intera scheda, o del paragrafo di appartenenza, o del campo di cui costituiscono la strutturazione.
Alcuni dei campi mettono in relazione una scheda con altre, mediante un insieme di riferimenti verticali (componenti di un oggetto), orizzontali (oggetti che compongono un aggregato di oggetti) e semantici (affinità definite mediante la specifica di proprietà). Per

ogni tipo di scheda è definita una struttura gerarchica, rappresentabile in XML e con uno schema XSD[12].

La migrazione dalle schede ad una rappresentazione ontologica non è banale e non è un processo automatico.

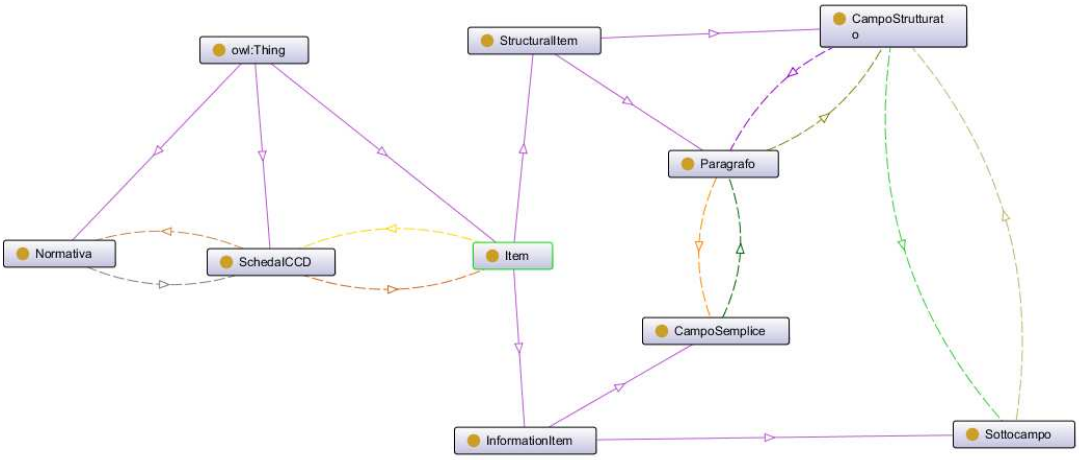

Figura 3: Lo schema astratto della scheda ICCD

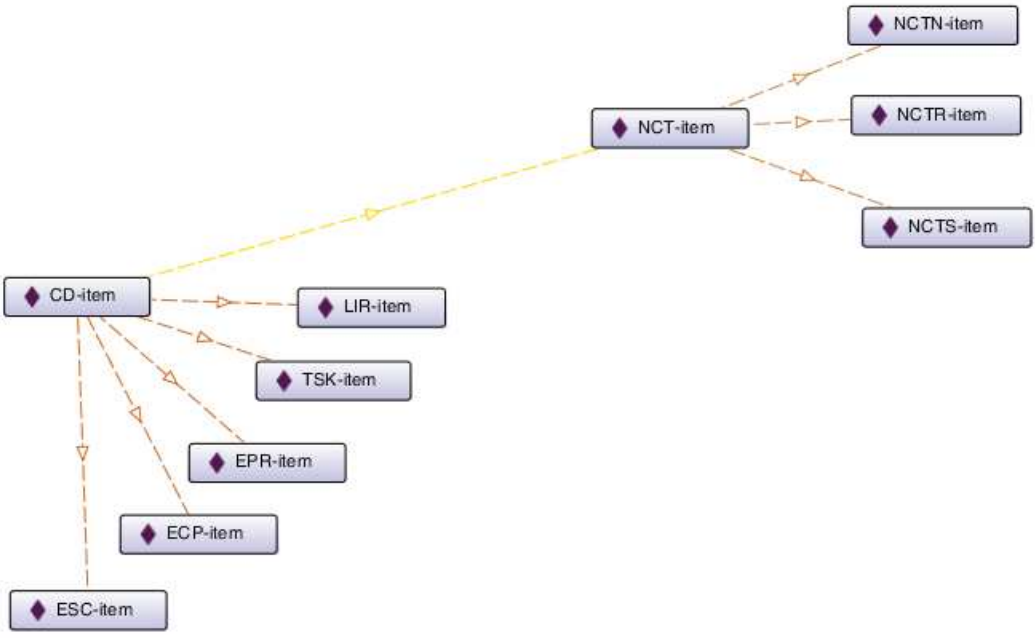

Figura 4: Le relazioni tra gli elementi costitutivi del paragrafo CD della scheda ICCD

Per prima cosa è opportuno rappresentare in termini ontologici la *struttura astratta* della scheda ICCD, descrivendone gli elementi strutturali e informativi, ed esplicitando il

12 Tuttavia, non sono stati pubblicati ufficialmente gli schemi XSD delle schede ICCD.

riferimento alla normativa di riferimento (vedi Figura 3). L'ontologia può essere individuata univocamente da un URI del tipo: http://.../strutturaSk-ICCD.owl.

La struttura di una qualunque scheda può poi essere rappresentata formalmente come ontologia, definita importando la struttura astratta e specificando campi e sottocampi, con le loro caratteristiche. Ogni elemento della scheda ICCD può essere modellato come classe, definendo poi per ogni classe un'istanza generica. In questo modo è possibile modellare esplicitamente come i singoli campi si mappano sulle istanze delle classi CRM, e definirne i tipi e le eventuali regole di validazione.

In Figura 4 sono riportate le relazioni tra i vari campi che costituiscono il paragrafo "CD" (Codici) della scheda ICCD, e nella Figura 5 la rappresentazione è arricchita con la visualizzazione delle classi a cui appartengono i singoli campi, che possono essere campi informativi o campi strutturali. L'ontologia è individuata univocamente da un URI del tipo: http://.../sk-BDI.owl.

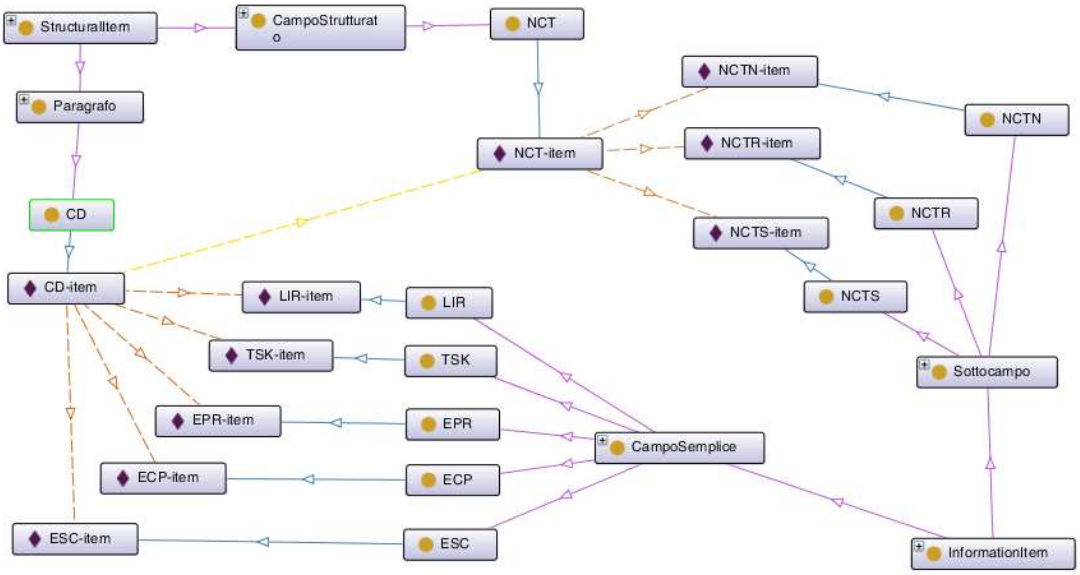

Figura 5: Gli elementi costitutivi del paragrafo CD della scheda ICCD

Va tenuto presente che l'approccio catalografico delle schede ICCD è profondamente differente dall'approccio ontologico adottato da CIDOC CRM e da qualsiasi struttura rappresentabile in RDF, dato che:

- la struttura ontologica esprimibile con la tripla "soggetto-predicato-oggetto" tende ad esplicitare una struttura logica dell'informazione che riproduce il linguaggio naturale ed esplicita relazioni e passaggi logici, "*predicando*" qualcosa di qualche altra cosa,

con un formalismo basato su RDF, per cui soggetto, predicato e oggetto possono essere una qualunque risorsa disponibile sul Web, quindi identificata da un URI, e l'oggetto può essere anche una stringa di caratteri ("literal").

- l'approccio catalografico tende a schematizzare e semplificare in campi e sottocampi i vari attributi e relazioni di una entità, per cui le relazioni semantiche rimangono del tutto implicite nel contesto di una scheda catalografica.

In alcuni casi l'insieme dei campi vuole modellare una serie di relazioni del tipo *partOf* o *belongsTo*, come nel caso delle localizzazioni. È evidente che la sequenza di Regione, Provincia, Comune e Località con i campi controllati da "vocabolari correlati" è rappresentata molto meglio e in forma più compatta da una relazione *partOf* tra istanze di una classe Luogo, o come istanze di classi diverse in relazione tra di loro.

Altre volte la strutturazione in campi e sottocampi esprime un unico concetto, suddiviso in più campi, come ad esempio un periodo cronologico, meglio esprimibile come istanza di una classe, con specifici valori per le proprietà.

Poiché la finalità del mapping dalla scheda ICCD al CRM è dettata dall'esigenza di strutturare semanticamente le informazioni ed i concetti che si estrapolano dai beni culturali, e poiché non tutti i campi delle schede catalografiche contengono valori significativi ai fini di una strutturazione semantica, la strategia di mapping può essere basata su queste ipotesi:

- la documentazione catalografica del bene rimane la scheda ICCD, pubblicata sul web in siti esterni a cui si potrà puntare, per cui il mapping tra la scheda ICCD e CRM non serve a produrre un file che si sostituisca alla documentazione catalografica esistente;
- non tutti i campi della scheda ICCD dovranno essere necessariamente mappati in CRM;
- informazioni tratte da campi soggetti a vocabolari di controllo e/o da serie di campi che contengono informazioni gerarchizzabili possono essere registrate in vocabolari e thesauri SKOS e collegate alla classe E 55 (Type) di CRM.

L'approccio scelto per il mapping porta dunque a procedere estrapolando dalla scheda entità e concetti fondamentali e traducendoli secondo uno schema ontologico in classi e proprietà del CRM, riproducendone gli attributi e le relazioni utili ad esporre, attraverso una serie di triple (s-p-o), le informazioni più significative che caratterizzano la conoscenza e ad operare ragionamenti complessi in base a connessioni semantiche.

Un problema[13] da risolvere è quello delle *relazioni semantiche implicite*. In molti casi l'approccio delle schede è riconducibile a una serie di relazioni (1:1 o 1:N) tra l'oggetto descritto dalla scheda e istanze di altre entità/classi (es. autore, attori) con delle connotazioni semantiche implicite, e questo meccanismo può essere ripetuto su due livelli. Ad esempio:

manifestazione	adotta	comunicazione Musicale
comunicazione Musicale	impiega	strumento Musicale

è un modo diverso di rappresentare una particolare sequenza di paragrafi/campi di una scheda BDI.

Un altro aspetto interessante è l'identificazione e la risoluzione delle *relazioni implicite di thesaurus*. Talvolta una serie di campi rappresentano informazioni correlate, o, come viene spesso detto, controllate da "vocabolari collegati".

Un esempio è quello di Descrizione del Bene (paragrafo DB) articolato nei campi DBL (Denominazione locale), DBD (Denominazione del bene) e DBC (Categoria), ognuno ripetibile, che sembrano identificare una struttura di concetti per cui un esemplare di un bene (DBD) appartiene ad una o più categorie (DBC), e può avere una o più denominazioni locali (DBL). Questo fatto viene modellato consentendo la ripetitività, in maniera indipendente, dei singoli campi.

Un caso in cui invece la ripetizione dei campi implica una molteplicità, mentre i sottocampi rappresentano una gerarchia di thesaurus, è quello relativo agli strumenti musicali utilizzati.

Nell'esempio:

CUS
- CUSC cordofoni
- CUSS violino

CUS
- CUSL aerofoni
- CUSA fisarmonica

è evidente che la ripetizione del campo[14] CUS indica una ripetitività (più strumenti), mentre la coppia di valori CUSC/CUSS e CUSL/CUSA riporta lo strumento utilizzato con la sua classificazione. Peraltro, il vocabolario aperto definito per CUSC è lo stesso definito per CUSL, e analogamente coincidono i vocabolari definiti per i sottocampi CUSS e CUSA.

13 Nel seguito, per illustrare le problematiche, di valenza assolutamente generale, del mapping, vengono riportati alcuni esempi che fanno riferimento alla scheda BDI (Beni Demoetnoantropologici).

14 Nel paragrafo CU (Comunicazione) è definito il campo ripetibile CUS (Musicale strumentale) con sottocampi CUSC (Strumenti musicali solisti/classificazione), CUSS (Strumenti musicali solisti), CUSL (Strumenti musicali di accompagnamento/classificazione) e CUSA (Strumenti musicali di accompagnamento).

In questo esempio è evidente la presenza di una classificazione di thesaurus per gli strumenti, mentre si potrebbero unificare i nomi dei campi, utilizzando un ulteriore campo che specifichi il "ruolo" con il quale compare lo strumento (solista o di accompagnamento). La soluzione più ovvia, e semanticamente più corretta, è quindi la definizione di un thesaurus (in questo caso un "thesaurus Strumenti"), e la definizione di una proprietà che identifichi anche il ruolo ricoperto dallo strumento (quindi o una proprietà con attributi o una coppia di proprietà).

Come già detto, la scheda originale ICCD rimane il documento ufficiale che descrive il bene, ed è identificata univocamente da un URI. Le opportune classi e proprietà CRM permettono di rappresentare il collegamento tra il bene (crm:E73.Information_Object) e la scheda catalografica (crm:E31.Document) che lo descrive (P67.refers_to).
Per esempio[15],

- http://iccd.beniculturali.it/paci/iccd/cards/stampa/283
  è la scheda che descrive il "Carnevale di Viareggio (bene che non ha una denominazione locale);
- http://iccd.beniculturali.it/paci/iccd/cards/stampa/157
  è una risorsa Web che descrive la manifestazione "Composizione satirica dialettale recitata pubblicamente", con denominazione locale (DBL) "*Businà*";
- http://iccd.beniculturali.it/paci/iccd/cards/stampa/323
  è una risorsa Web che descrive la manifestazione "*La Sartiglia*", con denominazione locale (DBL) "*Sa Sartiglia*".

Si noti che tutte e tre sono manifestazioni caratteristiche di luoghi diversi (Viareggio, Castelletto Monferrato e Oristano) che si svolgono in un'occasione (CA nella terminologia BDI) che è di tipo civile (CAC=sì) nel periodo di Carnevale (CAA=Carnevale).
Con una rappresentazione ontologica sarebbe possibile condurre in modo automatico un ragionamento spazio-temporale per trovare altre risorse (manifestazioni, in questo caso) collegate, perché si svolgono nello stesso periodo o in luoghi vicini. Allargando il discorso, ogni manifestazione potrebbe essere associata, grazie ad un ragionamento "spaziale", ad altre risorse (turistiche, alberghiere, artistiche), e, grazie ad un ragionamento "temporale", ad altri eventi o cibi tradizionali dello stesso periodo.

[15] In questo esempio è stato preso come URI l'identificatore della risorsa corrispondente alla stampa della scheda. I casi in cui non sia disponibile la scheda ICCD, ma esista un altro ente che rende disponibile l'informazione, potrà essere utilizzato l'URI corrispondente.

7. Un caso di studio: il progetto LabC

Le idee espresse in questo lavoro sono state oggetto di un progetto, denominato LabC (Laboratori per la Cultura)[16], sviluppato dall'azienda ETCWare con la collaborazione del CNR-ISTI di Pisa e dell'ufficio italiano del W3C.

Il progetto nasce come un ambiente di aggregazione di informazioni sui beni culturali per la fruizione da parte del pubblico e del mondo scientifico, e si pone come obiettivo lo sviluppo di una piattaforma multimediale web innovativa di tipo semantico, che consenta l'acquisizione, la condivisione e la fruizione di contenuti culturali appartenenti al patrimonio storico, paesaggistico e artistico italiano raccogliendo l'informazione sui beni culturali a partire da una base di utenza generica molto diffusa, e di trasformarla in informazione scientificamente strutturata, a disposizione sia dell'utenza autorevole che di quella generica.

Facendo leva sull'approccio di sviluppo "social" della piattaforma e sulle tecnologie del semantic web, l'utente ricopre un ruolo fondamentale con la possibilità di interagire direttamente con il portale contribuendo ad aumentare la base della conoscenza seguendo gli schemi catalografici definiti dall'ICCD, utilizzando appositi moduli per l'inserimento delle informazioni, che garantiscono il rispetto di tutti i vincoli previsti dalla normativa. Gli utenti potranno anche contribuire con documenti meno articolati, utilizzando moduli di inserimento più agili, e caratterizzare semanticamente l'informazione fornita inserendo (secondo l'approccio noto come folksonomy) tag che potranno essere scelti in piena libertà, oppure estratti dai vocabolari e thesauri già esistenti, sui quali sarà possibile navigare e operare una selezione.

Un ultimo aspetto del progetto, ma non il meno importante, perché di valenza assolutamente generale, è rappresentato dallo sviluppo del modulo SKOSware; una libreria che consente la consultazione e la manipolazione dei vocabolari e thesauri definiti durante le fasi di mapping della scheda BDI che hanno caratterizzato lo sviluppo del progetto.

16 Sviluppato nell'ambito del finanziamento di un progetto di RSI (Ricerca di Sviluppo Industriale) POR FESR Lazio 2007/2013 – Asse I – Attività I.1.

8. Conclusioni e possibili sviluppi

La conoscenza gioca un ruolo fondamentale nella catalogazione. Nella catalogazione tradizionale non c'è comunicazione bidirezionale di conoscenza tra gli esperti e gli utenti generici, e il problema forse più rilevante è costituito dall'approccio centrato sull'oggetto, per cui non vengono espresse chiaramente le importanti correlazioni semantiche con altri oggetti e con elementi conoscitivi di altre discipline.

L'avvento del Web ha mutato radicalmente le modalità di accesso all'informazione, e ha portato a riconoscere l'importanza dell'interoperabilità, sia tecnologica che semantica. In molti casi gli studiosi hanno ritenuto opportuno convergere verso un insieme comune di metadati, e Dublin Core si è imposto come standard riconosciuto. Tuttavia anche Dublin Core è di fatto un approccio centrato sull'oggetto, e solo gli esseri umani possono combinare la conoscenza proveniente da fonti informative diverse.

Negli ultimi anni, poi, si è affermato un approccio all'informazione in cui gli utenti giocano un ruolo attivo, e sono essi stessi produttori di conoscenza (il Web 2.0).

Il Semantic Web apre uno nuovo e affascinante scenario, in cui il Web è un immenso deposito di conoscenza, in cui buona parte dell'informazione proviene dai legami (link) tra dati memorizzati in punti diversi del web. La ricerca e il browsing sul web possono trarre importanti vantaggi da una rappresentazione interoperabile della conoscenza, interagendo secondo la metafora d'interazione preferita (spaziale, temporale, tassonomica). Nel contesto del Semantic Web degli agenti intelligenti possono fare affidamento su ontologie di base, come CIDOC CRM, per adeguarsi al modello mentale che esprime gli interessi dell'utente, e implementare opportuni meccanismi di navigazione [SIGNORE 1995, 2004A, 2005A, 2006]. Facendo riferimento alle classi CIDOC-CRM, un utente interessato al contesto temporale verrebbe condotto a istanze di classi quali: E2.Temporal_Entity, E52.Time-span e le loro sottoclassi a vari livelli, come E3 Condition State, E4.Period, E5.Event. Si potrebbe avere una definizione ancora più precisa del contesto specificando le proprietà di interesse (es. P117.occurs during, P118.overlaps in time with, etc.).

Va sottolineato, peraltro, che il Semantic Web costituisce semplicemente il contesto tecnologico: le ontologie devono essere definite e popolate. Fortunatamente esistono decenni di attività degli studiosi, che hanno prodotto una conoscenza che può essere capitalizzata e rappresentata in maniera formale, condivisibile ed esportabile tra varie aree, in modo da poter essere utilizzata da esseri umani e da macchine.

In questo panorama generale la possibilità che gli utenti possano contribuire alla conoscenza in maniera attiva, secondo il paradigma Web 2.0, costituisce un ulteriore elemento di arricchimento, soprattutto se è possibile armonizzare in un contesto ontologico i contenuti prodotti dagli utenti.

In un approccio ontologico, quale quello postulato dal Semantic Web, la conoscenza è disponibile da molte fonti, distribuite sul Web. La sfida è poter rappresentare, esportare e condividere la conoscenza, in modo che la conoscenza dell'esperto sia disponibile per tutti gli utenti, che potranno così effettuare le loro ricerche in un universo online senza soluzione di continuità, come se le immagini e le informazioni testuali pertinenti al patrimonio culturale fossero disponibili in un'unica enorme base informativa [FINK 1997].

Ringraziamenti

Un vivo ringraziamento va a tutti i colleghi e amici che hanno partecipato al progetto LabC (Alessandra Donnini, Andrea Ciapetti, Maria De Vizia Guerriero, Matteo Lorenzini, Maria Emilia Masci, Davide Merlitti, Fabio Piro) per le interessanti discussioni e per lo sviluppo delle varie parti del progetto LabC, che ha portato alla luce le difficoltà, ma anche la possibilità, di realizzare davvero un ambiente innovativo basato su tecnologie semantiche.

BIBLIOGRAFIA

Tutti gli URI riportati nella seguente bibliografia sono stati verificati il 1° novembre 2011.

ANTONIOU2004 Grigoris Antoniou, Frank van Harmelen, *A Semantic Web Primer*, The MIT Press, Cambridge, Massachusetts, April 2004, ISBN 0-262-01210-3.

BAKER2000 Baker, T. 2000, *A Grammar of Dublin Core*, in *D-Lib Magazine*, October 2000 Volume 6 Number 10, http://www.dlib.org/dlib/october00/baker/10baker.html

BEARMAN2006 David. Bearman, Jennifer Trant: "*Museums and Web 2.0*", Museums and Web 2006 the international conference for culture and heritage on-line, Albuquerque, New Mexico (2006)

BERNERS-LEE1998 T. Berners-Lee, *Semantic Web Road map*. Accessible at: http://www.w3.org/DesignIssues/Semantic.html .

BERNERS-LEE 2001 T. Berners-Lee, J. Hendler, O. Lassila, *The semantic web,* in *Scientific American,*34-43

BERNERS-LEE 2007 T. Berners-Lee, *Linked Data*, http://www.w3.org/DesignIssues/LinkedData.html.

BRISCOE2006 B Briscoe, A Odlyzko, B Tilly: *Metcalfe's law is wrong,* IEEE Spectrum, July 2006, p. 34-39

CAO2006 Yiwei Cao, Satish Narayana Srirama, Mohamed Amine Chatti, and Ralf Klamma, Mobile Social Software for Cultural Heritage Management, in R. Meersman, Z. Tari, P. Herrero et al. (Eds.): OTM Workshops 2006, LNCS 4277, pp. 955 – 964, 2006, Springer-Verlag Berlin Heidelberg 2006

CHUN2006 S. Chun, R. Cherry, D. Hiwiller, J. Trant, B. Wyman: "*Steve.museum: An Ongoing Experiment in Social Tagging, Folksonomy, and Museums*", Museums and Web 2006 the international conference for culture and heritage on-line, Albuquerque, New Mexico (2006). http://www.archimuse.com/mw2006/papers/wyman/wyman.html

CIDOC-CRM http://cidoc.ics.forth.gr

COYLE2007 Karen Coyle: "*MANAGING TECHNOLOGY - The Library Catalog in a 2.0 World*", The Journal of Academic Librarianship, Volume 33, Number 2, pages 289–291 (March 2007), preprint at: http://www.kcoyle.net/jal_33_2.html

DAMADIO1989 D'Amadio M., Simeoni P.E. 1989 *Strutturazione dei dati delle schede di Catalogo.Oggetti di interesse demo-antropologico*, Istituto Centrale per il Catalogo e la Documentazione (Roma) - Museo Nazionale delle Arti e Tradizioni Popolari (Roma), Tipografia Città Nuova della P.A.M.O.M., Roma

DC The Dublin Core Home Page, URL: http://dublincore.org/

DIGICULT 2003 *Towards a Semantic Web for Heritage Resources,* Thematic Issue 3, May 2003, http://www.digicult.info/downloads/ti3_high.pdf

DOERR 2003A M. Doerr, *The CIDOC CRM—An ontological approach to semantic interoperability of metadata*. AI Magazine 24, 3, 75–92 (2003)

DOERR2003B Doerr M., Hunter J., Lagoze C., *Towards a Core Ontology for Information Integration*, Journal of Digital Information, Volume 4 Issue 1, Article No. 169, 2003-04-09, (April 2003) http://www.cs.cornell.edu/lagoze/papers/core_ontology.pdf

DONNINI 2011 Donnini Alessandra, Ciapetti Andrea, De Vizia Guerriero Maria, Lorenzini Matteo, Masci Maria Emilia, Merlitti Davide, Piro Fabio, Signore Oreste, *Lifting Communities towards Semantic Web*, 5th International Congress "Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin", Istanbul, Turkey, 22nd-25th November 2011

FINK1997 Fink E. 1997, *Sharing Cultural Entitlements in the Digital Age: Are we Building a Garden of Eden or a Patch of Weeds?*, Museums and the Web: An International Conference, Los Angeles, CA, March 16 - 19, 1997, http://www.archimuse.com/mw97/speak/fink.htm>

GÓMEZ-PÉREZ2004 Asunción Gómez-Pérez, Mariano Fernández-López, Oscar Corcho, Ontological Engineering, Springer-Verlag (2004), ISBN 1-85233-551-3

HARDMAN2008 Lynda Hardman and Steven Pemberton: "*The Path to Web n+1*", http://ercim-news.ercim.org/content/view/340/536/

HYVONEN2004 Hyvönen E., Junnila M., Kettula S., Mäkelä E., Saarela S., Salminen M., Syreeni A., Valo A., Viljanen K. 2004, *Finnish Museums on the Semantic Web. User's Perspective on Museum Finland*, Proceedings of Museums and the Web 2004

HYVONEN2009 E. Hyvonen, *Semantic portal for cultural heritage*, in *Handbook on ontologies*, Berlin, 2nd edition.

ISAAC2009 A. Isaac, E.Summers, *SKOS Simple Knowledge Organization System Primer. W3C Working Group Note, World Wide Web Consortium* . http://www.w3.org/TR/skos-primer/

MCMILLAN1986 McMillan, D. W., and Chavis, D. M. (1986). "Sense of community: A definition and theory," Journal of Community Psychology, Vol. 14(1), pp. 6--23

METCALFE1995 Metcalfe, R: "Metcalfe's law: A network becomes more valuable as it reaches more users", Infoworld, N. 17, 2nd October 1995

O'REILLY2005 O'Reilly, Tim. 2005. What Is Web 2.0. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

PAPALDO1986 Papaldo S., Ruggeri Giove M., Gagliardi R., Matteucci D.R., Romano G.A., Signore O. 1986 *Strutturazione dei dati delle schede di Catalogo: beni mobili*, Atti del Convegno sull' Automazione dei Dati

del Catalogo dei Beni Culturali, Roma, 18-20 giugno 1985 (pag. 39-42) Ministero per i Beni Culturali e Ambientali - Istituto Centrale per il Catalogo e la Documentazione (Roma, 1986)

QUINTARELLI2007 E Quintarelli, L Rosati, A Resmin: *Facetag: Integrating Bottom-up and Top-down Classification in a Social Tagging System* - IA Summit, 2007 - http://www.facetag.org/download/facetag-20070325.pdf

SIGNORE1986 Signore O. 1986, *Architettura di sistemi per la gestione dei dati catalografici* - Atti del Convegno sull' Automazione dei Dati del Catalogo dei Beni Culturali, Ministero per i Beni Culturali e Ambientali - Istituto Centrale per il Catalogo e la Documentazione (Roma,1986) - Roma, 18-20 giugno 1985 (pag. 51-58)

SIGNORE1991 Signore O. 1991, *Cataloguing Art Objects: A Comparison between French and Italian standards* - European Museum Documentation Strategies and Standards, Proceedings of an International Conference held in - Canterbury, England, 2-6 September 1991 (Museum Documentation Association, 1993), ISBN 0-905963-83-0, pp.138-14

SIGNORE1994 Signore O. 1994, *From data structuring to data exchange: a simple path* - in Yesterday, Proceedings from the 6th International Conference Art History and Computing - August 28-30, 1991, Odense, Denmark, Marker H. J. & Pagh K. editors, Odense University press, 1994, pp. 48-54, ISBN 87 7838 022 7

SIGNORE1995 Signore O., *Issues on Hypertext Design,* in DEXA'95 -Database and Expert Systems Application, Proceedings of the International Conference, Lecture Notes in Computer Science, N. 978, Springer Verlag , ISBN 3-540-60303-4, pp. 283-292 - London, United Kingdom 4-8 September 1995

SIGNORE2003 Signore Oreste: *Strutturare la conoscenza: XML, RDF, Semantic Web* - Clinical Knowledge 2003 (1st edition) - Udine, 20-21 September 2003 http://www.w3c.it/papers/ck2003.pdf, http://www.w3c.it/talks/ck2003/overview.htm

SIGNORE2004A Signore O., *Representing Knowledge in Semantic Cultural Web* - EVA 2004 Jerusalem Conference on the Digitisation of Cultural Heritage - Jerusalem, 11-12 October 2004, http://www.w3c.it/talks/eva2004Jerusalem/overview.htm

SIGNORE2004B Signore, Oreste: *Qualità dei siti web pubblici culturali: dal progetto Minerva all'interoperabilità semantica* - Bollettino d' Informazioni, Centro di Ricerche Informatiche per i Beni Culturali, Scuola Normale Superiore - Pisa - ISSN: 1126-6090 - XII n. 2 (2002), p. 9-23

SIGNORE2005A Signore, Oreste: *Ontology Driven Access to Museum Information* - CIDOC 2005 Annual Conference of the International Committee for Documentation of the International Council of Museums ICOM-CIDOC - May 24 -27, 2005 Zagreb, Croatia - ISBN 953-6942-15-1
document: http://www.w3c.it/papers/cidoc2005.pdf
slides: http://www.w3c.it/talks/2005/cidoc2005/overview.html

SIGNORE2005B	Signore, Oreste and Missikoff, Oleg and Moscati, Paola: *La gestione della conoscenza in archeologia: modelli, linguaggi e strumenti di modellazione concettuale dall'XML al Semantic Web* - Archeologia e Calcolatori - Vol 16, p. 291-319 (2005)
SIGNORE2005C	Signore, Oreste: *Verso l'interoperabilità semantica* - In: F. Filippi (a cura di), Ministero per i beni e le attività culturali - Progetto Minerva-Manuale per la qualità dei siti Web pubblici culturali (edizione 2005)
SIGNORE2006	O. Signore, *The Semantic Web and Cultural Heritage: Ontologies and Technologies help in Accessing Museum Information*, in Information Technology for the Virtual Museum (December 6-7, 2006 – Sønderborg, Denmark), p. 1.31 (published 2008), Accessible at: http://www.weblab.isti.cnr.it/papers/public/itvm2006.pdf
SIGNORE2008	Signore, Oreste: *Introduzione al Semantic Web* – Web Senza Barriere '08 – Roma, 7-9 maggio 2008, http://www.w3c.it/papers/wsb08.pdf
SIGNORE2009	Signore, Oreste: *Representing knowledge in archaeology: from cataloguing cards to semantic web*, Archeologia e calcolatori, Vol. 20 (2009), p. 111-128
STUDER1998	Rudi Studer and V. Richard Benjamins and Dieter Fensel in Knowledge Engineering: Principles and Methods, Data Knowl. Eng. 25(1-2): 161-197 (1998)
VANDERSLUIJS2008	Kees van der Sluijs and Geert-Jan Houben, *Tagging and the Semantic Web in Cultural Heritage*, ERCIM News n. 72 http://ercim-news.ercim.org/content/view/336/536/
VANHARMELEN2008	*Which Future Web?*, An interview with Frank van Harmelen http://ercim-news.ercim.org/content/view/339/536/
XU2006	Z Xu, Y Fu, J Mao, D Su - *Collaborative Web Tagging Workshop* at WWW2006, Edinburgh, 2006 - http://www.semanticmetadata.net/hosted/taggingws-www2006-files/13.pdf

APPENDICI

Appendice A: le tecnologie di base del Semantic Web

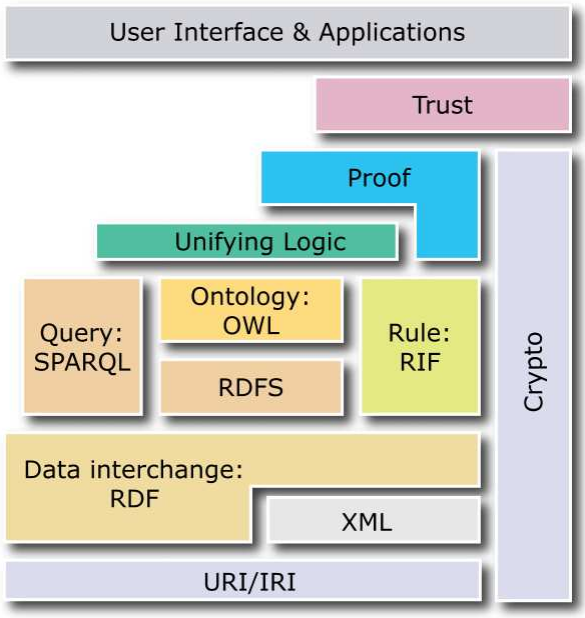

Figura 6: L'architettura del Semantic Web

URI

Un *Uniform Resource Identifier* (URI) consente di identificare in maniera semplice le risorse. La specifica della sintassi e della semantica degli URI deriva dai principi architetturali di base del World Wide Web, definiti già all'inizio degli anni 1990.

I concetti di base di un URI, che danno ragione del suo nome, sono:

Uniform È possibile utilizzare identificatori di tipo diverso nello stesso contesto, anche quando si usano meccanismi differenti per accedere alle risorse. È quindi anche possibile introdurre nuovi tipi di risorse, senza interferire sul modo in cui sono stati utilizzati gli identificatori già esistenti.

Resource: Una risorsa può essere qualunque cosa identificabile con un URI. Esempi tipici possono essere documenti, immagini, fonti di informazione, servizi, collezioni di risorse. Non è necessario che la risorsa sia disponibile e accessibile sul Web, anche se questo è il caso più frequente (si pensi per esempio a quando si fa riferimento a persone, oggetti fisici, tipi di associazioni, valori numerici, etc.).

Identifier Per definizione, un identificatore contiene le informazioni sufficienti per distinguere, nell'ambito del dominio a cui si riferisce, l'oggetto identificato

> rispetto agli altri. Non bisogna supporre che l'identificatore definisca o contenga l'identità della risorsa di cui è l'identificatore, anche se talvolta è così, e neppure ritenere che il sistema possa utilizzare l'URI per accedere alla risorsa specificata, o che la risorsa sia un oggetto "semplice".

In definitiva, un URI è una sequenza di caratteri che rispetta una precisa sintassi, e consente l'identificazione di una risorsa mediante un insieme estensibile, e definito a parte, di "*naming scheme*". Il modo in cui viene realizzata l'identificazione è delegato a ciascuna "*scheme specification*".

La specifica tecnica degli URI (http://www.ietf.org/rfc/rfc2396.txt) *non definisce* alcun limite sulla natura delle risorse, sulle ragioni per cui un'applicazione può richiedere una risorsa, o il tipo di sistemi che possono utilizzare URI per identificare risorse, e *non impone* che un URI identifichi permanentemente la stessa risorsa, anche se questo è un obiettivo comune di tutti gli *URI scheme*. Tuttavia la specifica consente alle applicazioni di limitarsi a particolari tipi di risorse, o a sottoinsiemi di URI che mantengono le caratteristiche desiderate dall'applicazione.

Ogni URI comincia con uno *scheme name*, che fa riferimento a una specifica per assegnare gli identificatori all'interno di quello schema. Di conseguenza, la sintassi degli URI costituisce un sistema federato ed estensibile per il *naming*, in cui la specifica di ogni *scheme name* può ulteriormente restringere la sintassi e la semantica degli identificatori che utilizzano quel particolare *scheme name*.

URI, URL, e URN

Inizialmente, agli albori del Web, i tecnici supposero che gli identificatori delle risorse potessero essere distinti in due o più classi. Un identificatore avrebbe potuto specificare la localizzazione di una risorsa (URL – *Uniform Resource Locator*) o il suo nome (*URN – Uniform Resource Name*) indipendente dalla localizzazione. Dopo varie e accese discussioni, e dopo aver verificato che altri potenziali sottospazi in cui suddividere gli URI non avevano di fatto avuto applicazione pratica, si convenne che, senza perdita di generalità, lo spazio degli URI poteva essere partizionato nei due sottospazi URL e URN. Quindi, per esempio, "http:" era un *URL scheme*, e "isbn:" potrebbe un domani essere un *URN scheme*. Qualunque nuovo *scheme* sarebbe poi ricaduto in una delle due classi.

Con il passar del tempo, l'importanza di questa suddivisione è andata scemando, e si è affermata l'idea che un particolare scheme non debba necessariamente ricadere in una

specifica classe come "URL", "URN", "URC" etc., mentre i Web-identifier scheme sono in generale degli URI-scheme, e un particolare URI-scheme può definire dei sottospazi, detti "namespace".

Allo stato attuale, il termine URL non fa riferimento ad una partizione formale dello spazio degli URI, ma è piuttosto un concetto utile ma informale: un URL è un tipo di URI che identifica una risorsa mediante la rappresentazione del suo metodo primario di accesso (la sua localizzazione nella rete), piuttosto che mediante altri suoi potenziali attributi. L'espressione "URL scheme" viene oggi utilizzata raramente, in genere per specificare una sottoclasse di URI scheme che esclude degli URN.

IRI

Un *IRI (Internationalized Resource Identifier)* è una sequenza di caratteri appartenenti all'Universal Character Set (ISO10646/Unicode). È stato definito un mapping dagli IRI agli URI, per cui è possibile utilizzare gli IRI invece degli URI quando ciò sia opportuno e appropriato per identificare le risorse.

XML

Extensible Markup Language (**XML**) è nato per far fronte alle limitazioni di HTML nella realizzazione delle nuove applicazioni Web, in cui i dati costituiscono un elemento essenziale (*data-centric Web applications*). XML è stato quindi il primo passo per assegnare una semantica ai tag e supportare le transazioni sul Web, permettendo lo scambio di informazioni tra database diversi. Ulteriori e significativi vantaggi sono costituiti dalla possibilità di avere viste diverse degli stessi dati, e la possibilità di personalizzare le informazioni mediante opportuni agenti. L'adozione di XML agevola la gestione di collezioni di documenti, e costituisce un supporto fondamentale per la pubblicazione di informazioni a livello internazionale, con il non piccolo vantaggio di essere indipendente dalla piattaforma e dal linguaggio.

```
01  <?xml version="1.0"?>
02  <!DOCTYPE ordine [
03  <!ELEMENT ordine ( cliente, prodotto+ )>
04  <!ATTLIST ordine id ID #REQUIRED>
05  <!ELEMENT cliente EMPTY>
06  <!ATTLIST cliente db CDATA #REQUIRED>
07  <!ELEMENT prodotto  ( importo )>
08  <!ATTLIST prodotto  db CDATA #REQUIRED>
09  <!ELEMENT importo( #PCDATA )>
10  ]>
11  <ordine id="ord001">
12 <cliente db="codcli123"/>
13 <prodotto db="prod345">
14<importo>23.45</importo>
15 </prodotto>
16  </ordine>
```

Figura 7: Un documento XML (con la sua DTD in grassetto)

Le caratteristiche di XML possono essere illustrate con un esempio semplice, relativo alla gestione degli Ordini (Figura 7). La sintassi XML usa tag di inizio e fine, come per esempio <importo> e </importo>, per marcare i campi informativi. Un campo informativo racchiuso tra due marcatori viene detto elemento (*element*) e può essere ulteriormente arricchito dalla presenza di coppie nome/valore (nell'esempio, id="ord001") dette attributi (*attribute*). Come si può vedere, si tratta di una sintassi semplice, la cui elaborazione automatica è poco complessa, senza codifiche particolarmente criptiche, per cui resta comprensibile alla lettura diretta. I tag devono essere inseriti correttamente uno dentro l'altro, deve esistere una corrispondenza tra il tag di apertura e quello di chiusura, sono previsti elementi a campo informativo nullo e gli attributi dei tag devono essere racchiusi tra doppi apici.

La presenza di una struttura formale del documento, espressa nella **DTD** *(Document Type Definition*), non ha un impatto diretto sul modello strutturale implicito: nell'esempio di Figura 7Figura , in cui la DTD è inclusa nel documento (ma potrebbe anche essere referenziata come risorsa esterna) la riga 6 specifica che l'attributo *db* è obbligatorio. Un documento XML si dice "*well formed*" quando rispetta le regole di scrittura; viene detto "*validato*" quando è coerente con la struttura definita nella DTD.

XML Schema Definition

La DTD presenta alcune limitazioni, riconducibili essenzialmente al fatto che viene espressa con una sintassi sua propria, e quindi richiede editor, parser e processor ad hoc. Inoltre, è

difficile estenderla, non contempla datatype e deve supportare tutti gli elementi e attributi descritti dai namespace[17] inclusi.

Gli schema hanno le stesse funzionalità delle DTD, ma offrono alcuni significativi vantaggi: sono espressi con la sintassi XML e includono datatype, inheritance, regole di combinazione degli schema. **XMLSchema** fornisce anche un miglior supporto dei namespace e offre la possibilità di agganciare documentazione e informazioni semantiche. XMLSchema permette di rappresentare vincoli sui possibili valori, tipi complessi e gerarchie di tipi.

RDF

Resource Description Framework (RDF) è lo strumento base per la codifica, lo scambio e il riutilizzo di metadati strutturati, e consente l'interoperabilità tra applicazioni che si scambiano sul Web informazioni machine-understandable. RDF consente l'elaborazione automatica delle risorse reperibili sul Web, e può essere utilizzato e portare vantaggi sono in molti settori.

RDF fornisce un modello per descrivere le risorse, che hanno delle proprietà (o anche attributi o caratteristiche). RDF definisce una risorsa come un qualsiasi oggetto che sia identificabile univocamente mediante un Uniform Resource Identifier (URI).

Il *data model* RDF, che consente di rappresentare statement RDF in modo sintatticamente neutro, è molto semplice ed è basato su tre tipi di oggetti:

Resources Qualunque cosa descritta da una espressione RDF viene detta risorsa (*resource*). Una risorsa può essere una pagina Web, o una sua parte, o un elemento XML all'interno del documento sorgente. Una risorsa può anche essere un'intera collezione di pagine Web, o anche un oggetto non direttamente accessibile via Web (per es. un libro, un dipinto, etc.). Le risorse sono sempre individuate da un URI, eventualmente con un *anchor id* (riferimento interno). Qualunque cosa può essere identificata da un URI.

Properties Una *property* (proprietà) è un aspetto specifico, una caratteristica, un attributo, o una relazione utilizzata per descrivere una risorsa. Ogni proprietà ha un significato specifico, definisce i valori ammissibili, i tipi di risorse che può descrivere, e le sue relazioni con altre proprietà. Le proprietà associate alle risorse sono identificate da un *nome*, e assumono dei *valori*.

[17] Un XML namespace è un insieme di nomi, caratterizzati da un URI di riferimento, utilizzati come element type e attribute name.

Statements Una risorsa, con una proprietà distinta da un nome, e un valore della proprietà per la specifica risorsa, costituisce un RDF *statement* (asserzione). Uno statement è quindi una *tripla* composta da un *soggetto* (risorsa), un *predicato* (proprietà) e un *oggetto* (valore). L'oggetto di uno statement (cioè il *property value*) può essere un'espressione (sequenza di caratteri o qualche altro tipo primitivo definito da XML) oppure un'altra risorsa.

Graficamente (Figura 1), le relazioni tra Resource, Property e Value vengono rappresentate mediante *grafi orientati etichettati*, in cui le risorse vengono identificate come nodi (graficamente delle ellissi), le proprietà come archi orientati etichettati, e i valori corrispondenti a sequenze di caratteri come rettangoli. Un insieme di proprietà che fanno riferimento alla stessa risorsa viene detto descrizione (*description*).

Uno statement può poi essere codificato in vari modi, a seconda della sintassi adottata.

È importante sottolineare che sia le risorse che le proprietà sono identificate univocamente da URI. La potenza del meccanismo di RDF risiede appunto nel fatto che un documento RDF può far riferimento a proprietà e risorse che sono definite in un qualunque punto del Web, senza necessità di centralizzare le informazioni.

RDFS

RDF è un linguaggio universale che consente agli utenti di usare il loro vocabolario per descrivere le risorse. Per questo motivo RDF non formula nessuna assunzione su qualunque dominio applicativo specifico, né ne definisce la semantica.

Per esprimere le restrizioni sulle associazioni, in altri termini per evitare che possano essere codificati degli "statement" sintatticamente corretti, ma privi di senso, è necessario un meccanismo per rappresentare "classi di oggetti". Da questa esigenza nasce "*RDF Vocabulary Description Language*", che mantiene anche, per ragioni storiche, il nome di "RDF Schema" (RDFS).

L'aspetto fondamentale è costituito dalla necessità di esprimere fatti e condizioni non solo sui singoli oggetti, ma anche sulle classi che definiscono i tipi di oggetto. Usualmente, si dice che una classe può essere vista come un insieme di elementi, che vengono indicati come istanze della classe. In RDF possiamo usare la proprietà rdf:type per specificare la relazione tra istanze e classi. Il meccanismo delle classi può essere utilizzato per imporre restrizioni sulle proposizioni che possono essere enunciate in un documento RDF che usi quello schema. Per esempio, potremmo dire che hanno senso proposizioni (*triple s-p-o*) come:

Leonardo	autoreDi	Gioconda.
Cimabue	maestroDi	Giotto.

mentre altre proposizioni quali:

Michelangelo	autoreDi	Leonardo.
ritrattoDiGiulioII	autoreDi	Gioconda.

sono entrambe prive di senso, perché un artista non può essere autoreDi un altro artista, e un'opera non può essere autoreDi un'altra opera. Invece, un artista può essere maestroDi un altro artista, e un'opera può essere in relazione con un'altra opera perché ne è una versione diversa o una copia. In maniera più formale, possiamo dire che è necessario poter restringere *dominio* e *codominio* (*domain* e *range*) delle proprietà.

Una volta definite le classi, può risultare utile definire delle relazioni tra di esse. Per esempio, potremmo aver definito le classi:

- cane
- mammifero

e aver asserito che: "Attila" è un cane (cioè è un'istanza di cane).

Per un essere umano è evidente a tutti che "Attila" è un mammifero, ma, se vogliamo che le macchine siano in grado di comprendere questi fatti, e operare dei ragionamenti, dobbiamo essere in grado di asserire in modo formale la conoscenza che "ogni cane è un mammifero".

Questo genere di relazioni tra classi è conosciuto come "*gerarchia di classi*", che non è necessariamente una gerarchia ad albero semplice, nel senso che una classe può essere una sottoclasse di più classi.

È compito dell'utente definire i vincoli sulle classi e sulle proprietà, e le eventuali gerarchie di sottoclasse.

OWL

RDF e RDF Schema presentano dei limiti di espressività, in quanto RDF consente unicamente di specificare predicati binari, e RDF Schema consente unicamente di definire gerarchie di classi e proprietà, e di imporre vincoli per dominio e condominio. Applicazioni sofisticate richiedono di poter "ragionare" sui dati. Il Semantic Web deve quindi essere supportato da ontologie, e disporre di un linguaggio che consenta di definire la terminologia

usata, le caratteristiche logiche e i vincoli delle proprietà, l'equivalenza dei termini, le cardinalità delle associazioni, etc. Un'ulteriore complessità deriva dal fatto che il Web è intrinsecamente distribuito, e di conseguenza applicazioni diverse possono usare ontologie diverse, o le stesse ontologie, ma espresse in lingue diverse. Il W3C, sfruttando anche i risultati di altri progetti, quali DAML e OIL, ha definito un linguaggio, denominato OWL, che permette di esportare le ontologie in modo interoperabile.

Un buon ontology language dovrebbe avere un certo numero di caratteristiche sofisticate, ma quanto più è espressivo il linguaggio, tanto meno efficiente è il ragionamento, e, in particolar modo nel definire un linguaggio da utilizzare sul Web, occorre trovare un compromesso tra espressività e computabilità. In particolare, OWL avrebbe potuto essere un'estensione di RDFS, utilizzando il significato delle classi e proprietà definite in RDF (rdfs:Class, rdfs:subclassOf, etc.), aggiungendo le primitive necessarie per supportare la maggior ricchezza espressiva richiesta. In tal modo sarebbe stata mantenuta anche la coerenza con l'architettura a strati del Semantic Web (Figura 6). Purtroppo, questo avrebbe cozzato contro l'esigenza di contenere la complessità computazionale, e assicurare una adeguata efficienza nel ragionamento. In particolare, costrutti come rdfs:Class e rdf:Property (la classe di tutte le classi e la classe di tutte le proprietà), pur essendo molto espressivi, avrebbero portato a una complessità computazionale fuori controllo.

Per questo motivo OWL offre tre sottolinguaggi, di crescente potere espressivo:

OWL Lite indicato principalmente per utenti che hanno bisogno di rappresentare classificazioni gerarchiche e vincoli semplici. Consente una migrazione agevole per thesauri e altre tassonomie. Ha una complessità formale inferiore a OWL DL, e non consente alcuni costrutti, come cardinalità arbitrarie o statement di disgiunzione.

OWL DL (OWL Description Logic) indicato per gli utenti che desiderano la massima potenza espressiva garantendo comunque che tutte le conclusioni siano computabili (*computational completeness*) e concluse in un tempo finito (*decidability*). OWL DL offre un ragionevole supporto per il ragionamento (*reasoning*), ma non è perfettamente compatibile con RDF. Per ottenere un documento OWL DL corretto da un documento RDF, è necessario ricorrere ad alcune restrizioni e ad alcune estensioni. Tuttavia, qualunque documento OWL DL corretto è un documento RDF corretto.

OWL Full indicato per gli utenti che desiderano la massima potenza espressiva e la libertà sintattica di RDF, senza garanzie sui tempi di computazione. Qualunque documento RDF corretto è un documento OWL Full corretto, perché OWL Full è perfettamente compatibile con RDF, sia sotto l'aspetto sintattico che sotto quello semantico. Tuttavia, il linguaggio è indecidibile, non esistono strumenti che supportino il ragionamento in maniera efficiente o completa, e difficilmente sarà supportato nella sua interezza da software che implementano il ragionamento.

Ognuno di questi linguaggi è un'estensione del precedente, sia in termini di ciò che può essere espresso che in termini della validità delle conclusioni.

Appendice B: SKOS

Il Simple Knowledge Organization System (SKOS) sviluppato dal consorzio internazionale W3C è un formalismo standard per la rappresentazione della conoscenza espressa in thesauri, tassonomie e vocabolari.

Usando SKOS, la conoscenza può essere rappresentata come concetti *machine readable* che possono essere scambiati tra due o più piattaforme in quanto interoperabili e pubblicati sul WEB in modo standard.

Lo SKOS viene formalmente definito come un'ontologia OWL full i cui dati possono essere codificati secondo la sintassi RDF prevedendo la formalizzazione dei concetti in triple (RDF:Concepts).

Lo SKOS struttura la conoscenza in uno *schema concettuale* comprendente una serie di *concetti*. Sia lo schema che i concetti sono identificati da URI. Tali concetti sono relazionati tra di loro con relazioni *gerarchiche* o *associative*.I concetti possono essere etichettati con *n* stringhe; poi, all'interno della struttura, verrà definita una pref:Label; le altre verranno categorizzate come alt:Label. Ai concetti possono essere assegnate una o più annotazioni (*annotations*) che identificano unicamente il concetto.

I concetti SKOS possono essere documentati con note di vario tipo (skos:note, skos:scopeNote etc), raggruppati in collezioni (*collections*) che possono essere etichettate e ordinate (ex alfabeticamente) e, infine, possono essere mappati (*mapped*) su altri SKOS concept e concept schema.

Il concetto SKOS può essere visto come un'idea, una nozione o un'unità di pensiero. I concetti sono collegati a un URI o a un *RDF identifier*. Tale caratteristica permette allo SKOS di essere riutilizzato e collegato ad altri thesauri o ontologie. I concetti possono essere inoltre collegati a schemi concettuali caratterizzati dall'aggregazione di uno o più skos:concept. Uno skos:conceptScheme può avere uno o più *top concepts* che rappresentano il punto di partenza della struttura gerarchica.

In SKOS, le *relazioni* sono espresse tramite collegamenti tra concetti ed il loro significato. SKOS distingue due tipologie di relazioni: *gerarchiche e associative.* Una relazione gerarchica tra due concetti indica che uno è più generale (*broader*) dell'altro (*narrower*), mentre nella relazione associativa troviamo che due concetti sono relazionati in quanto *concettualmente coerenti*; ciò non significa che un concetto è più generale dell'altro. Si noti che le relazioni skos:broader e skos:narrower non sono transitive. Per definire gerarchie transitive di termini vanno usate le relazioni skos:broaderTransitive e skos:narrowerTransitive.

Appendice C: CIDOC CRM

Vengono riportati in questa appendice alcuni dei concetti fondamentali necessari per comprendere la struttura dell'ontologia e le regole di naming, utili per leggere correttamente i concetti espressi seguendo il formalismo di questa ontologia.

Classi, sottoclassi e proprietà

I concetti fondamentali del modello CIDOC CRM sono classi, sottoclassi e proprietà.

Classe Una classe è una categoria di elementi che condividono un insieme di caratteristiche utilizzabili come criteri per identificare gli elementi della classe. L'insieme di queste caratteristiche viene detto *intension* della classe. Una classe può essere *dominio* o *codominio* di una o più proprietà definite nel modello. Un elemento appartenente ad una classe viene detto istanza della classe. Una classe è quindi associata ad un insieme di oggetti del mondo reale, che sono detti *extension* della classe. Il numero di istanze della classe non è determinabile a priori ("*Open World*") e quindi una classe non può essere definita per enumerazione dei suoi elementi.

Sottoclasse Una sottoclasse è una specializzazione di un'altra classe (la sua *superclasse*). La specializzazione è una relazione IsA, per cui:

1. tutte le *istanze* della sottoclasse sono anche istanze della sua superclasse;
2. l'*intension* della sottoclasse *estende* l'intension della superclasse, cioè le sue caratteristiche sono più restrittive di quelle della sua superclasse;
3. la sottoclasse eredita, senza eccezioni, tutte le proprietà della sua superclasse (*strict inheritance*), e possiede zero o più proprietà addizionali.

Una sottoclasse può avere più di una superclasse (*multiple inheritance*), ereditando quindi tutte le loro proprietà. La gerarchia IsA (o gerarchia di classe) è *transitiva* e non può essere ciclica.

Proprietà Una proprietà consente di definire una *relazione* tra due classi (dette *dominio* e *codominio*). È possibile definire una *proprietà inversa*, in cui le classi partecipanti si scambiano il ruolo di dominio e codominio. Le proprietà possono anche essere *specializzate* in una gerarchia IsA, con conseguente definizione di sottoproprietà e superproprietà (*subproperties* e *superproperties*).

Convenzioni di Naming

Le convenzioni di naming adottate in CRM sono le seguenti:

- Le classi sono identificate da numeri preceduti dalla lettera "E" (storicamente, le classi erano denominate Entità), e hanno un nome costituito da un gruppo nominale scritto utilizzando le iniziali maiuscole. Per esempio: E63 Beginning of Existence.
- Le proprietà sono identificate da numeri preceduti dalla lettera "P", e hanno un nome (sia la diretta che l'inversa) costituito da un gruppo verbale scritto in lettere minuscole. Viene usato il tempo presente se le proprietà hanno un carattere di "stato" (es. "has type"), e il tempo passato se si riferiscono a eventi (es. "carried out"). Un esempio di proprietà è: P126 employed (was employed in).
- I nomi delle proprietà vanno letti nella forma non parentetica nella direzione dominio-codominio (*domain-range*) e nella forma parentetica per la direzione codominio-dominio (*range-domain*).

- Le proprietà il cui codominio è una sottoclasse di E59 Primitive Value (per esempio E1 CRM Entity. P3 has note: E62 String) non hanno una forma parentetica, perché la proprietà letta nel verso codominio-dominio non è significativa.

Le proprietà per le quali dominio e codominio coincidono sono o simmetriche o transitive. Istanziare una proprietà simmetrica significa affermare che la relazione vale in entrambi i versi. Un esempio è: E53 Place. P122 borders with: E53 Place. È evidente a priori che le proprietà simmetriche non hanno una forma parentetica, in quanto si leggono nello stesso modo nelle due direzioni. Le proprietà transitive asimmetriche, come per esempio E4 Period. P9 consist of (forms part of): E4 Period, hanno invece una forma parentetica che rende ragione del significato della lettura della relazione nella direzione inversa.

I tipi

Tutte le descrizioni strutturate di oggetti museali e di oggetti catalogati in generale hanno normalmente un identificatore univoco dell'oggetto ("*unique object identifier*") e riportano informazioni relative al "tipo" di oggetto, spesso articolato in una serie di campi (Classificazione, tipo, categoria, etc.).

In CRM la classe E55 Type comprende questi termini, estratti da thesauri e dizionari controllati, utilizzati per classificare le istanze delle classi CRM. Le istanze di E55 Type rappresentano concetti (universali) in contrasto con le istanze di E41 Appellation, che vengono invece utilizzate per dare un nome alle istanze delle classi CRM. E55 Type costituisce quindi l'interfaccia di CRM verso thesauri e ontologie di dominio, che possono essere rappresentati in CRM come sottoclassi di E55 Type, formando gerarchie di termini, cioè *istanze* di E55 Type collegate tra di loro dalla proprietà P127 has broader term (has narrower term). Le gerarchie possono poi essere arricchite con ulteriori proprietà.

CRM fornisce due proprietà di base che descrivono la classificazione mediante la terminologia, uniformandosi così alla pratica diffusa nella maggior parte dei sistemi informativi.

La classe E1 CRM Entity è il dominio della property P2 has type (is type of), che ha come codominio E55 Type. Di conseguenza ogni classe definita in CRM, con l'unica *eccezione* di E59 Primitive Value, *eredita* la *property* P2 has type (is type of). Questo è un meccanismo generale per simulare una specializzazione della classificazione delle istanze CRM a qualunque livello di dettaglio, con un link a fonti esterne come vocabolari, thesauri, schemi di classificazione, ontologie.