Dai metadati all'harvesting

la gestione di risorse informative attraverso repository interoperabili

di Gino Roncaglia[1]

Pubblicato su 'Culture del testo e del documento' 26/2008, pp. 109-122

1. Cosa sono i metadati

In questo articolo, cercherò di offrire una introduzione generale al concetto di metadati, discutendone la centralità per le prospettive del nuovo web e in particolare per la gestione organizzata dell'informazione attraverso sistemi di repository.

Il concetto di metadati ha in realtà un ruolo importantissimo nella gestione di qualunque tipo di contenuto informativo o documentale, soprattutto quando l'informazione disponibile è molta e deve essere dunque selezionata e organizzata per facilitarne il reperimento e l'uso. Non a caso dunque, nonostante si tratti di un concetto perfettamente applicabile in tante situazioni che non hanno nulla a che fare con il mondo della rete e delle nuove tecnologie, i metadati sono uno dei 'mattoni' fondamentali del nuovo web.

Ma cosa sono i metadati? Etimologicamente, il termine richiama l'idea di 'dati di secondo livello', dati utilizzati per descrivere e classificare altri dati. L'esempio tipico, e quello più comunemente utilizzato per spiegare il concetto, è il catalogo di una biblioteca. Gli scaffali di una biblioteca raccolgono libri (i nostri dati di primo livello, o 'informazione primaria'), mentre il catalogo raccoglie schede che *descrivono* i libri, sulla base di una serie di caratteristiche prefissate: il nome dell'autore, il titolo, il luogo e l'anno di pubblicazione, il numero di pagine, la collocazione negli scaffali della biblioteca (o in un sistema di classificazione astratto), e così via. Le schede di un catalogo contengono dunque dei dati di secondo livello, dei *metadati* appunto, che descrivono l'informazione primaria e aiutano a gestirla e reperirla. Dal canto suo, la biblioteca comprende *sia* i libri *sia* il catalogo: una buona raccolta di risorse informative, infatti, comprende sempre anche informazione secondaria, metadati, che permettono di descrivere, organizzare e reperire l'informazione primaria (una biblioteca, come sappiamo, offre al proprio utente anche molte altre cose, e in particolare un insieme assai ampio di strumenti e servizi che aiutano a lavorare sia sull'informazione primaria sia sui metadati).

Situazioni simili – in cui utilizziamo dei metadati per identificare e descrivere alcune caratteristiche del mondo informativo che ci circonda, in modo da muoverci al suo interno nel modo più efficace – sono molto comuni: decidiamo quale programma guardare in televisione o in quale cinema andare la sera utilizzando le guide alla programmazione pubblicate dal giornale (e cosa sono queste guide se non raccolte organizzate di metadati relativi ai programmi televisivi o cinematografici?), ordiniamo un pasto al ristorante dopo aver valutato sul menu i piatti disponibili e il loro prezzo, facciamo la spesa al supermercato consultando le

[1] Dipartimento di Scienze Umane, Università della Tuscia

etichette dei prodotti per verificarne gli ingredienti, il prezzo o la data di scadenza, decidiamo se comprare un libro dando un'occhiata all'indice e alla descrizione in quarta di copertina, e così via.

Tutto molto semplice, sembrerebbe. Ma in realtà anche i metadati nascondono le loro insidie (o il loro fascino, per chi ama l'avventura intellettuale!). Ad esempio: dove finiscono i dati e dove cominciano i metadati? Il titolo di un libro, stampato a chiare lettere sulla sua copertina (e ripreso nella scheda della nostra ipotetica biblioteca) è un dato (dopotutto è parte del testo del libro…) o un metadato? Il mio nome e il mio cognome servono a individuarmi… sono forse metadati anch'essi? E se lo sono, cosa descrivono? Il nostro linguaggio non è forse costituito in primo luogo di nomi utilizzati per riferirci a oggetti o concetti? Ma allora *tutto* il linguaggio è fatto di metadati?

È facilissimo, insomma, scivolare dai metadati alla filosofia. E non è certo un caso che, come vedremo, nel parlare di metadati intervenga spesso un termine nato proprio in ambito filosofico: il termine *ontologia*. Ma procediamo con ordine, cercando – a rischio di qualche semplificazione – di chiarirci un po' le idee.

Partiamo da quello che è probabilmente l'aspetto fondamentale dei metadati: abbiamo a che fare con metadati quando, volendo descrivere o gestire una risorsa informativa o un certo insieme di risorse informative, ne individuiamo e selezioniamo alcune caratteristiche al fine di costruirne un modello semplificato, funzionale ai nostri scopi e facilmente condivisibile con altri. Questo avviene, di norma, individuando all'interno dell'universo informativo di partenza (l'informazione primaria) proprietà, tratti comuni e differenze che ci aiutino a organizzarlo. Così, davanti ai libri della mia libreria, potrei scegliere come caratteristiche rilevanti per descriverli e gestirli l'autore e il titolo di ciascuno di essi, ma anche, ad esempio, la loro dimensione, il colore della copertina, il tipo di carattere tipografico utilizzato, il numero di pagine, l'anno di pubblicazione, e ancora la data in cui l'ho comprato, se l'ho prestato o no a qualcuno, se il libro ha delle sottolineature, e così via.

Ciascuna di queste caratteristiche – e potrei immaginarne un'infinità di altre – corrisponde a un possibile tipo di metadati, e ciascuna potrebbe essere utile in certi contesti: selezionare un certo insieme di metadati (e scartarne molti altri possibili) dipende sempre dai nostri scopi, e implica la scelta di un particolare modello, di una particolare rappresentazione dell'informazione primaria. In un certo senso, equivale a scegliere una particolare 'visione del mondo', definendo le categorie (e le relazioni) in base alle quali vogliamo organizzare l'universo informativo che ci interessa. Per questo,come si è accennato, a proposito di metadati si parla spesso di *ontologie*: un'ontologia ci dice 'cosa è reale' per noi, quali sono le caratteristiche che scegliamo di prendere in considerazione nel nostro dominio di riferimento.

Così, se il nostro scopo è organizzare i libri di una libreria in base al contenuto informativo del testo che contengono, il colore della copertina ci interesserà relativamente poco, mentre ci interesseranno di più il titolo, l'autore, gli argomenti trattati. Se prestiamo molti libri (e magari alcuni non tornano indietro) diventa rilevante tener traccia delle persone alle quali li abbiamo prestati; e se la libreria non è la nostra ma quella, poniamo, di un grande scrittore, può essere importante sapere se e quali libri hanno delle sottolineature o delle note manoscritte, che potrebbe essere interessante studiare. In ciascuno di questi casi, selezioniamo caratteristiche per noi rilevanti all'interno dell'universo informativo di riferimento, e le utilizziamo per organizzarlo in base a un modello.

Figura 1 - Metadati: esempio di metadati relativi a una curiosa scultura cinese (da http://www.lib.ku.edu/imagegateway/help/workspace.cfm)

Abbiamo detto che esistono molti tipi possibili di metadati. Non stupirà dunque che ne siano state tentate diverse classificazioni (il che equivale un po' a cercare dei meta-metadati…), anche con l'obiettivo di aiutare a strutturare l'insieme di metadati che utilizziamo, distinguendo al suo interno caratteristiche e tipologie diverse.

Una distinzione abbastanza diffusa è quella fra metadati *descrittivi*, *gestionali-amministrativi* e *strutturali*: come suggeriscono i nomi, i metadati descrittivi (ad esempio, il titolo di un libro) hanno lo scopo fondamentale di descrivere l'informazione primaria; quelli amministrativo-gestionali (ad esempio, il responsabile dell'acquisto di un libro) hanno lo scopo principale di semplificarne la gestione, e quelli strutturali (ad esempio, la suddivisione in sezioni di un testo) aiutano a strutturarla. Ma questa classificazione non è né l'unica possibile, né necessariamente esaustiva o sempre soddisfacente: proprio come la scelta dei metadati stessi, la scelta di particolari classificazioni o organizzazioni dei metadati è legata al tipo di informazione primaria di cui ci stiamo occupando, ai nostri obiettivi, alla nostra 'visione del mondo'.

Così, ad esempio, in certi contesti possono avere rilievo (mentre in altri contesti può essere preferibile evitarli) metadati *valutativi,* come le 'stelle' attribuite ad un film dalla recensione di un giornale. E in effetti fra le varie distinzioni proposte in quest'ambito non manca quella fra metadati *oggettivi* (ricavabili direttamente dall'analisi dell'informazione primaria, in maniera tendenzialmente indipendente dalle opinioni o valutazioni di chi classifica) e metadati *soggettivi* (come appunto i metadati valutativi, in cui l'elemento di valutazione soggettiva è preponderante). Così come non manca una distinzione relativa alla responsabilità della metadatazione, fra metadati *primari* (forniti dall'autore della risorsa informativa), *secondari* (forniti dagli utenti) e *terziari* (forniti da servizi o organizzazioni indipendenti). O ancora la distinzione fra metadati inclusi nella risorsa informativa (*embedded metadata*) e metadati esterni alla risorsa informativa (*external metadata*). Insomma, così come nel descrivere, strutturare o gestire una

risorsa informativa possiamo scegliere metadati diversi, anche nel guardare ai metadati (che sono, ricordiamolo sempre, a loro volta informazione, dunque anch'essi dei dati) possiamo adottare prospettive diverse, e proporre diverse distinzioni e classificazioni.

Come si sarà capito già da queste prime considerazioni, i metadati non nascono, come Minerva, perfettamente compiuti e indipendenti dalla mente di Giove: sono invece il risultato di un lavoro di elaborazione, che richiede siano presi in considerazione tre elementi, tutti e tre indispensabili: 1) la risorsa informativa che rappresenta la nostra informazione primaria (ad esempio, una copia della Divina Commedia); 2) le proprietà, categorie o relazioni attraverso cui abbiamo deciso, per i nostri scopi specifici e in accordo con la nostra 'visione del mondo', di organizzare l'universo informativo al quale quella risorsa appartiene (ad esempio 'autore', 'titolo', ecc.) e 3) i valori che a tali proprietà vengono attribuiti nel caso della particolare risorsa informativa che stiamo considerando (ad esempio, la proprietà 'autore' ha, nel caso di quel libro, il valore 'Dante Alighieri').

Volendo adottare un linguaggio appena un po' più preciso (ma senza inoltrarci in un formalismo veramente rigoroso, che richiederebbe ulteriori e ben più complesse precisazioni), possiamo dire che l'assegnazione di un metadato a una risorsa informativa corrisponde a una *tripla* (a_i , p_j , v_k) il cui primo elemento è appunto la risorsa informativa, cioè un oggetto del dominio composto dalle nostre risorse informative primarie ($a_i \in \mathcal{D}$, dove $\mathcal{D}$ = {a_1, a_2, a_3 ... }, e a_1, a_2 … sono le risorse informative che ci interessano, ad esempio i libri di una biblioteca); il cui secondo elemento è una particolare proprietà tratta dall'insieme di proprietà utilizzate per descrivere quel tipo di risorse primarie ($p_j \in \mathcal{P}$, dove $\mathcal{P}$ = {p_1, p_2, p_3 ... p_n} e p_1 … p_n sono le proprietà che ci interessano, ad esempio 'autore', 'titolo' ecc.); e il cui terzo elemento è un valore, scelto fra quelli ammissibili per la proprietà che stiamo considerando ($v_k \in \mathcal{V}^{pj}$, dove $\mathcal{V}^{pj}$ = {v_1, v_2, v_3 ... } e v_1, v_2 … sono i valori ammissibili per la proprietà p_j ; così, se ad esempio p_j è la proprietà 'autore', v_1, v_2 … potranno essere 'Dante Alighieri', 'Alessandro Manzoni' ecc.).

Abbiamo detto che questa descrizione, pur costituendo una approssimazione accettabile per i nostri scopi[2], è in realtà ancora tutt'altro che rigorosa. In effetti, per renderla tale dovremmo considerare il fatto che oltre a proprietà di una singola risorsa informativa ci può interessare anche prendere in considerazione *relazioni* fra risorse informative diverse; dovremmo inoltre disporre di strumenti per esprimere in maniera chiara quali sono i vincoli posti sui valori che una certa proprietà può assumere; potremmo poi voler definire all'interno del dominio di riferimento dei sottoinsiemi che ammettono proprietà diverse, fare scelte migliori per quel che riguarda il formalismo utilizzato (ad esempio distinguendone esplicitamente gli aspetti sintattici e gli aspetti semantici), introdurre regole e magari addirittura assiomatizzazioni che permettano di operare deduzioni (ad esempio, dedurre l'assegnazione di particolari valori a determinate proprietà in conseguenza di assegnazioni o scelte fatte in precedenza), e così via. La definizione rigorosa di tutte queste caratteristiche 'astratte' del nostro sistema di dati e di metadati corrisponde alla individuazione di un'*ontologia*: una specificazione formale, esplicita e condivisibile con altri (siano essi persone o agenti software) di quali siano le tipologie di oggetti che compongono il nostro dominio di riferimento, di quali siano le proprietà e relazioni che possono essere stabilite fra di essi, e di quali siano le regole che ne governano il funzionamento.

[2] I lettori che hanno qualche familiarità con i formalismi del web semantico avranno notato che la nostra descrizione si avvicina molto al modo in cui il data model RDF (Resource Description Framework) considera gli *statements*, che corrispondono appunto a triple composte, fondamentalmente, da risorse, proprietà e valori. Non a caso, RDF è uno degli strumenti essenziali per esprimere in maniera formalmente rigorosa e per condividere metadati strutturati

Torneremo fra breve su alcuni di questi temi: per concludere questa discussione introduttiva, e alla luce delle considerazioni fin qui esposte, limitiamoci per ora a dire, in maniera abbastanza generale, *che i metadati sono informazione, di norma altamente strutturata, utilizzata per descrivere, strutturare o gestire una risorsa informativa o un insieme di risorse informative (la nostra informazione primaria) attraverso l'identificazione di alcune sue proprietà e l'assegnazione ad esse di specifici valori*.

2. Interoperabilità, standardizzazione

L'insieme di metadati che si sceglie di utilizzare per descrivere un certo insieme di risorse informative (ad esempio i libri di una biblioteca, o i contenuti di un corso) corrisponde, come si è detto, a delle scelte consapevoli, che portano comunque a privilegiare alcune caratteristiche e proprietà dell'informazione primaria a scapito di altre. Questo spiega perché, in situazioni diverse, si possa scegliere di utilizzare, con riferimento alla stessa informazione primaria, insiemi diversi di metadati. Ma se organizzazioni che devono gestire in maniera fondamentalmente analoga lo stesso tipo di informazioni primarie facessero scelte troppo diverse relativamente ai metadati utilizzati, diventerebbe molto difficile, se non impossibile, rendere interoperabili i rispettivi sistemi di gestione informativa. Così, ad esempio, se due musei utilizzassero insiemi di metadati molto diversi per descrivere i quadri esposti, diventerebbe molto difficile realizzare cataloghi collettivi, o gestire in maniera efficace l'informazione connessa al prestito di un quadro da un museo all'altro. E se biblioteche diverse schedassero i propri libri in maniera completamente diversa, il povero utente che ne dovesse consultare i cataloghi, magari per trovare un libro di difficile reperibilità, dovrebbe ogni volta imparare e adottare nuove convenzioni. L'esigenza dell'*interoperabilità* porta dunque a una naturale richiesta di *standardizzazione*: adottare insiemi standardizzati di metadati permette non solo di semplificarsi la vita (e di semplificarla ai propri utenti), ma anche di scambiare informazioni e 'schede', di utilizzare gli stessi programmi per la loro gestione, di realizzare servizi di ricerca integrata, e così via. E come è facile capire queste esigenze diventano ancor più stringenti quando – come accade ormai quasi nella totalità dei casi – riguardano sistemi di gestione dell'informazione che prevedano l'uso di computer e agenti software.

La standardizzazione, naturalmente, è possibile finché gli scopi alla base della nostra scelta di metadati, e la 'visione del mondo' che adottiamo davanti a un certo tipo di informazione primaria, sono ragionevolmente uniformi. In caso contrario, può essere del tutto ragionevole utilizzare per la stessa informazione primaria sistemi diversi di metadati (e dunque ontologie diverse). Così, ad esempio, un ufficio postale smisterà la corrispondenza sulla base dell'indirizzo e del codice postale scritti sulla busta, senza occuparsi – di norma – del contenuto e del mittente delle lettere, mentre il destinatario della corrispondenza potrà organizzarla nel suo archivio proprio sulla base di questi ultimi elementi. Le due prospettive, quella dell'ufficio postale che deve smistare la corrispondenza indipendentemente dal suo contenuto informativo e quella del destinatario umano che deve leggerla, interpretarla e gestirla proprio in funzione del suo contenuto, sono così lontane fra loro da rendere difficile l'identificazione di una 'ontologia' comune, se non a un livello molto generale ed astratto (compito, quest'ultimo, che può comunque rivelarsi interessante: non a caso, chi si occupa di ontologie ha un'innata passione per l'astrazione!).

Comunque, in tutti i casi in cui è sensato farlo, è più che ragionevole cercare di utilizzare sistemi di metadati standardizzati, o almeno sistemi nei quali sia standardizzata la gestione dei metadati comuni: così, ad esempio, una libreria e una biblioteca hanno entrambi bisogno di sistemi di metadati per la gestione dei libri, e questi sistemi potranno coincidere per una larga parte, ma la biblioteca non avrà bisogno dei

metadati relativi alla vendita dei libri (anche se potranno servire metadati relativi al loro acquisto e alla loro eventuale dismissione), mentre la libreria non avrà bisogno dei metadati relativi al prestito.

La ricerca della standardizzazione è dunque una caratteristica essenziale nella costruzione di sistemi di metadati: non a caso, il lavoro di definizione di uno standard è in genere lungo e complesso, ed è spesso affidato a commissioni di esperti, non di rado con la supervisione di istituzioni sovranazionali. Ma la ricerca della standardizzazione ha un prezzo, per quanto necessario: per cercare di adattare un sistema di metadati al maggior numero possibile di situazioni, si finisce spesso per allargarlo e renderlo più complesso di quanto ciascuna delle singole situazioni richiederebbe. Capita dunque in molti casi e per molti progetti specifici di utilizzare *sottoinsiemi* di sistemi di metadati standard: il riferimento allo standard garantisce l'interoperabilità (anche se solo rispetto ai metadati effettivamente utilizzati), senza sacrificare eccessivamente la semplicità delle applicazioni.

3. Organizzare i metadati: dalla strutturazione al rigore formale

Fin qui, abbiamo parlato di metadati immaginandoli un po' come una successione lineare di descrittori. Ma è importante ricordare che, nella maggior parte dei casi, i sistemi di metadati non sono 'piatti' ma sono invece fortemente strutturati. Per capire questo concetto, torniamo per un istante all'esempio di un libro e a uno dei suoi metadati più tradizionali, l'autore. Chi è l'autore di un libro? La risposta è a prima vista assai semplice: chi lo ha scritto. Ma sappiamo bene che il mondo reale in questi casi ama tenderci dei trabocchetti. Chi è l'autore dei primi fumetti di Asterix: l'autore dei testi, Goscinny, o il disegnatore, Uderzo? Tutti e due, risponderemmo probabilmente, ma con ruoli diversi. E l'autore della prefazione di un libro, o il responsabile di una edizione critica, o il curatore di una antologia, che ruolo hanno, e come devono figurare nella nostra ipotetica 'scheda' dell'opera? Si tratta evidentemente di figure diverse con ruoli diversi, ma di figure che hanno tutte una parte nella costruzione dei contenuti del libro. Può essere utile, allora, prevedere dei metadati più generali (ad esempio 'creator' e 'contributor'), articolandoli poi in ruoli più specifici: autore principale, illustratore, prefatore, curatore, e così via. In una parola: i sistemi di metadati sono normalmente *strutturati*, non consistono in una serie piatta di descrittori ma in gerarchie organizzate.

Inoltre, anche se molti metadati riguardano quelle che sembrano essere intuitivamente delle *proprietà* di un oggetto (così il titolo o l'autore possono sembrare intuitivamente delle 'proprietà' di un libro), abbiamo già avuto occasione di accennare al fatto che molti metadati hanno una natura *relazionale*. Così, ad esempio, fra le informazioni relative a un libro possiamo trovare "traduzione inglese". Essere una traduzione può sembrare a prima vista una proprietà del libro (e in un certo senso lo è), ma… quel libro è una traduzione *di che cosa*? In effetti, per essere completa la nostra indicazione deve fare riferimento a un'altra risorsa informativa: l'informazione ha in questo caso anche una natura relazionale, ci dice che un certo libro è una traduzione di un altro libro, e nel farlo stabilisce una relazione di un certo tipo fra due oggetti del nostro dominio.

Insomma, i sistemi di metadati hanno di norma una natura complessa, altamente strutturata, spesso gerarchica, e comprendono non solo metadati diversi ma, come abbiamo visto, anche metadati *di tipo* diverso o con *funzioni* diverse. Una considerazione che del resto non dovrebbe sorprenderci troppo, dato che, per quanto semplificati, i sistemi di metadati si propongono di offrire modelli per la gestione di una informazione di primo livello che è a sua volta in genere complessa, variamente articolata e strutturata.

Nelle intenzioni, si è detto, i sistemi di metadati dovrebbero aiutarci a gestire l'informazione primaria in maniera funzionale e standardizzata. Ricordiamo di nuovo che sempre più spesso a gestire questa informazione non sono esseri umani ma programmi (quelli che tecnicamente si chiamano 'agenti software'). Ai programmi non basta la standardizzazione, serve anche una estrema attenzione al rigore formale. Ad esempio, l'anno di pubblicazione e il numero di pagine di un libro sono tutti e due espressi attraverso numeri, ma si tratta di numeri di tipo ben diverso! Per capirlo, basta pensare che in certe circostanze può aver senso sommare dei numeri di pagina (ad esempio per avere la paginazione complessiva di un libro in più tomi), mentre non ha senso sommare due anni di pubblicazione. L'anno di pubblicazione è una *data*, e le date vanno gestite in modo da riconoscerle e tenere conto della loro particolare natura. Un essere umano riconosce in genere a colpo d'occhio una data e capisce che si tratta di un dato numerico di tipo molto particolare, mentre il computer lo sa fare solo se abbiamo *esplicitato* le differenze esistenti fra i due tipi di dati.

Una buona formalizzazione di un sistema di metadati richiede che sia prestata specifica attenzione a tutti questi aspetti. Innanzitutto, dovrà essere esplicitato il dominio di riferimento: a quali classi di dati si applicano i nostri metadati? Definire il dominio di riferimento può sembrare banale, ma spesso non lo è affatto. Se vogliamo descrivere dei libri, dovremo capire (o stabilire attraverso qualche convenzione) *cosa* esattamente si intenda per libro; se vogliamo descrivere dei contenuti per la didattica a distanza, dovremo capire *quali tipi di contenuti* ci interessa considerare e classificare. Una operazione che spesso ci porta a individuare non solo una classe di dati, ma anche la presenza al suo interno di sottoclassi rilevanti ai fini della nostra classificazione. Ad esempio, se vogliamo un sistema di metadati che si applichi sia ai libri sia alle riviste, ma che distingua queste due classi di oggetti relativamente ad alcune proprietà, potremmo voler distinguere all'interno del nostro dominio di riferimento due sottoclassi specifiche.

Dovremo poi specificare quali proprietà vogliamo considerare, e a quali classi o sottoclassi di oggetti queste proprietà si possono applicare. Molto spesso, fra proprietà diverse esistono relazioni di dipendenza (ad esempio, se un libro ha la proprietà di essere diviso in più volumi, ciascuno di questi volumi avrà a sua volta la proprietà di essere parte del libro in questione): se vogliamo che il sistema sia capace di gestirle, magari ricavando autonomamente alcune delle conclusioni deducibili a partire dalla (meta)informazione già in nostro possesso, anche queste relazioni dovranno essere riconosciute ed espresse formalmente. Dovremo specificare quali tipi di valore possano assumere le proprietà che consideriamo (l'anno di pubblicazione sarà ad esempio una data, ma potrebbe anche essere espresso attraverso un punto interrogativo o attraverso un intervallo di date, se non siamo sicuri di quella giusta). Se i nostri metadati sono strutturati in forma gerarchica, dovremmo esprimere formalmente anche le relative dipendenze. Insomma, la costruzione di un buon sistema di metadati implica, come si era già accennato, la creazione di una vera e propria *ontologia*, descritta con il maggiore rigore formale possibile. E sono stati elaborati linguaggi specifici, come OWL (Web Ontology Language) che hanno proprio lo scopo di descrivere tali ontologie in modo rigoroso e formale.

4. Ontologie rigorose o classificazioni libere?

Tutto questo lavoro può certo intimidire. Pensavamo di dover descrivere la nostra informazione primaria usando un po' di parole chiave e di descrittori intuitivi, e ci troviamo impegnati in un lavoro di costruzione di complicati sistemi formali, che useremo poi magari per un decimo delle loro potenzialità ma che renderanno comunque il nostro lavoro di descrizione e classificazione assai più complesso e faticoso. È davvero necessaria tutta questa fatica?

Per un verso, la risposta è sicuramente affermativa. Anche se complesso e faticoso, il lavoro di costruzione di sistemi di metadati 'solidi', condivisi e ben strutturati garantisce la compatibilità fra le nostre descrizioni e quelle date da altri, permette la gestione automatica della nostra informazione primaria (e degli stessi metadati), risponde insomma all'esigenza già richiamata di *interoperabilità* dei sistemi di gestione informativa permettendoci in molti casi di automatizzare il loro funzionamento. Inoltre la discussione che accompagna la definizione di uno standard, e la sua sperimentazione in molte situazioni diverse, aiutano a migliorare e a *validare* il nostro sistema di metadati, e spesso anche a capire meglio la natura della nostra informazione primaria: un risultato importante, giacché – come abbiamo ricordato – un sistema di metadati rappresenta sempre una particolare 'visione del mondo' (anche se limitata alla specifica classe di dati che ci interessa descrivere) , con tutti i rischi di arbitrarietà che questo può comportare.

D'altro canto, la descrizione rigorosa dell'informazione ha un serio problema: è difficile! E la conseguenza di scelte sbagliate, o di applicazioni sbagliate di sistemi di classificazione magari validissimi, può essere in questo caso assai pesante, in termini di costi e di risultati.

Negli ultimi due o tre anni,si è parlato molto dell'uso di un paradigma del tutto diverso nel campo della descrizione dei contenuti informativi: la descrizione 'dal basso', da parte degli utenti, attraverso 'etichette' ('tag') libere. È la strada del cosiddetto 'social tagging', adottato da molti sistemi di gestione dell'informazione 'user generated' inserita in rete: meccanismi di social tagging sono presenti ad esempio in molti blog, in sistemi di gestione di immagini come Flickr, o di video come YouTube. In questi casi gli utenti descrivono i loro contenuti (e quelli degli altri) usando le etichette che preferiscono. Il sistema può poi usare meccanismi di filtraggio collaborativo ('collaborative filtering') per fare emergere le etichette e le classificazioni più usate o considerate più utili da parte degli utenti stessi.

In questa prospettiva, il social tagging nasce, se non in opposizione, almeno in una situazione di qualche tensione rispetto al progetto del cosiddetto 'semantic web' (web semantico) proposto da Tim Berners Lee[3]. Anche il semantic web nasce dalla necessità di trovare strumenti per migliorare la classificazione e la reperibilità dell'informazione; ma il progetto di semantic web è strettamente legato all'idea, che abbiamo discusso fin qui, secondo cui il lavoro di organizzazione e gestione dell'informazione deve essere in gran parte automatico e basato su descrizioni fortemente standardizzate e formalizzate, attraverso l'uso dei sistemi strutturati di gestione dei metadati. I sistemi di classificazione dell'informazione alla base del semantic web sono – o dovrebbero essere – *ontologie formali*, basate su griglie di classificazione elaborate da esperti del settore, espresse in maniera uniforme e rigorosa e associate all'informazione primaria attraverso l'uso di linguaggi e formalismi a loro volta rigidamente strutturali e ben definiti.

Al contrario, come si è visto, il social tagging e in generale molti strumenti del cosiddetto 'web 2.0'[4] mettono gli utenti al centro anche del processo di classificazione. Al posto di sistemi classificatori predefiniti, gli utenti che immettono informazione in rete e quelli che la usano sono invitati ad aggiungere

[3] T. Berners-Lee, J. Hendler, O. Lassila, "The Semantic Web", *Scientific American*, May 2001, pp. 34–43; per gli sviluppi più recenti delle posizioni di Berners Lee al riguardo cf. N. Shadbolt, W. Hall, T. Berners-Lee, "The Semantic Web Revisited", *IEEE Intelligent Systems* 21(3) pp. 96-101, May/June 2006, in rete all'indirizzo http://eprints.ecs.soton.ac.uk/12614/01/Semantic_Web_Revisted.pdf.

[4] Si veda al riguardo G. Roncaglia, " Gli strumenti del nuovo web e l'organizzazione della ricerca in campo umanistico", in corso di pubblicazione negli atti del convegno "Le opere filosofiche e scientifiche. Filosofia e scienza tra testo, libro e biblioteca", Lecce, 7-8 febbraio 2008, a cura di Franco A. Meschini. Il preprint è disponibile in rete all'indirizzo http://dspace.unitus.it/handle/2067/433.

all'informazione primaria delle etichette descrittive totalmente libere, sulla base della loro percezione della natura dell'informazione stessa e dei contesti di sua potenziale utilità.

Ci occupiamo più a fondo di questi sistemi (spesso chiamati anche 'folksonomy') in altra sede, anche perché il social tagging rappresenta una delle tendenze distintive del cosiddetto 'Web 2.0'. Qui può essere il caso di ricordare che, se il social tagging può essere effettivamente utile a garantire un primo livello di gestione 'facile' e immediata di grandi quantità di contenuti generati dagli utenti, funziona assai meno bene quando l'obiettivo è garantire una buona reperibilità e interoperabilità di informazione validata e dalle caratteristiche un po' più uniformi. Va infatti considerato da un lato che le 'etichette' attribuite dagli utenti sono di norma piatte, non strutturate (e questo impedisce di rendere conto di molte forme di dipendenza e strutturazione della nostra informazione primaria), dall'altro che l'assenza di normalizzazione, anche solo nel lessico utilizzato, impedisce in molti casi il reperimento affidabile e funzionale dell'informazione che ci interessa.

Questo non toglie, però, che l'uso di sistemi di social tagging possa spesso utilmente *accompagnare* (anziché sostituire) una descrizione più rigorosa e standardizzata dei contenuti. In questo caso, la sua funzione diventa quella di esprimere descrizioni alternative, percorsi trasversali, permettendo di dar voce anche al punto di vista dell'utente (o meglio, ad alcuni fra i molti punti di vista possibili di molti tipi diversi di utenza). E indubbiamente il social tagging può rappresentare l'unica strada sensatamente percorribile quando l'informazione da descrivere è in quantità o di tipologia tale da non permettere, o da rendere troppo oneroso, l'intervento di classificatori esperti. Come accade spesso, appunto, nel mondo dell'informazione 'user generated' resa disponibile attraverso la rete. In questi casi, però, è bene che anche i sistemi di social tagging messi a disposizione degli utenti siano 'pensati' in maniera attenta e consapevole, in modo da cercare di guidare gli utenti (ad esempio – come già accennato – attraverso l'impiego di strumenti di collaborative filtering) a una scelta di etichette il meno arbitraria e il più informativa e sensata possibile. Questo potrà permettere, in alcuni casi, di strutturare e organizzare a posteriori, anche in maniera automatica, i metadati immessi (quasi) liberamente dagli utenti.

5. Metadati e repository; l'harvesting

Abbiamo visto che lo scopo principale dei metadati è di norma descrivere l'informazione primaria, organizzarla, semplificarne e automatizzarne la gestione. Non stupirà dunque che uno dei più importanti contesti d'uso di sistemi di metadati sia rappresentato dai depositi organizzati (*repository*) di contenuti, siano essi digitali o no. Anzi, possiamo dire che per poter parlare di deposito *organizzato* di contenuti, di repository appunto, la presenza di metadati rappresenta una condizione imprescindibile. Un repository è dunque uno strumento di raccolta, conservazione e (in genere) distribuzione di contenuti informativi descritti, gestiti e organizzati attraverso opportuni insiemi di metadati.

In questa sede, il nostro interesse si concentra soprattutto sui repository di contenuti in formato digitale, accessibili via rete. In genere, essi sono gestiti attraverso programmi che permettono l'archiviazione o l'autoarchiviazione di contenuti, la loro identificazione univoca, la loro metadatazione, la loro organizzazione in raccolte o collezioni, la loro conservazione anche di lungo termine, la loro ricerca e il loro recupero. Di norma, in particolare per l'accesso alle funzionalità di archiviazione dei contenuti, è prevista un'identificazione degli utenti attraverso sistemi di autenticazione. Talvolta i repository sono dotati anche

di strumenti aggiuntivi, ad esempio per la gestione dei diritti relativi ai contenuti archiviati, per la conversione automatica di formati, per la creazione di contenuti di pacchetto, per l'esposizione dei metadati, per l'aggiunta ai contenuti di commenti, giudizi o etichette da parte degli utenti, per la distribuzione di feed RSS relativi ai nuovi contenuti immessi o resi disponibili, ecc.: ci occuperemo in seguito di alcune di queste funzionalità, mentre di altre abbiamo parlato o parleremo in altre occasioni.

Fra le tipologie di repository più diffuse e importanti (a loro volta variamente articolate) ricordiamo le Digital Library (biblioteche digitali), gli Open Archive (archivi aperti), le piattaforme per la raccolta, gestione e condivisione di contenuti sonori o visivi (immagini e filmati), i sistemi per la gestione e l'accesso a basi dati di riviste e libri elettronici realizzati da molte case editrici, i LCMS (Learning Content Management System), destinati alla gestione e distribuzione di contenuti e risorse di apprendimento. Spesso i confini fra queste diverse tipologie sono labili e dipendono, più che dallo specifico strumento informatico utilizzato, dall'uso che ne viene fatto, o da considerazioni di mercato. In tutti questi casi, comunque, la gestione dei metadati gioca un ruolo fondamentale, e una delle caratteristiche distintive dei vari repository è proprio legata ai sistemi di metadati che essi riconoscono e consentono di utilizzare.

Per rendere disponibili e ricercabili i propri contenuti, un repository deve dunque innanzitutto rendere disponibili e ricercabili i propri metadati. Ed è bene ricordare che, anche se frequentemente pensiamo ai metadati come a una sorta di 'involucro' che circonda l'informazione primaria, non è affatto detto che i metadati debbano sempre essere utilizzati 'insieme' ai relativi contenuti. Proprio come le schede del catalogo di una biblioteca possono essere utilizzate anche indipendentemente dai libri che descrivono (e magari essere raccolte in pubblicazioni autonome, utili a controllare anche a distanza il patrimonio di una biblioteca o a recuperare specifiche informazioni bibliografiche), anche i metadati in formato elettronico possono avere una vita autonoma, circolare ed essere distribuiti per disseminare le informazioni relative ai contenuti di un certo repository e facilitarne il reperimento.

Una delle modalità più interessanti di questa circolazione 'autonoma' dei metadati è rappresentata dal concetto di *harvesting*. Vediamo di capire di cosa si tratta.

Intuitivamente, saremmo portati a pensare che il modo migliore per semplificare il reperimento e il recupero di grandi quantità di contenuti informativi sia quello di non disperdere i contenuti stessi e realizzare invece repository il più larghi e centralizzati possibile, abbastanza ricchi e completi da non richiedere all'utente eccessive 'peregrinazioni' da un repository all'altro per trovare il contenuto che lo interessa.

A ben vedere, però, per garantire semplicità ed efficienza non occorre necessariamente utilizzare pochi grandi repository: è anche possibile (e in molti casi è più semplice ed efficace) realizzare più depositi, magari settoriali o corrispondenti a istituzioni diverse, che però condividano l'uso di sistemi standard di metadati e di recupero dei contenuti, e rendano disponibili i loro metadati per forme di indicizzazione integrata e cooperativa. In sostanza, non occorre centralizzare i contenuti: basta standardizzare i metadati e centralizzare i servizi per la loro ricerca e indicizzazione (tenendo anche presente che in genere i metadati 'pesano' in termini di bit molto meno delle relative risorse informative primarie, e possono dunque viaggiare con estrema facilità).

È chiaro che questo modello richiede l'adozione generalizzata di standard e protocolli condivisi, sia per quello che riguarda i sistemi di metadati, sia per quello che riguarda le procedure per facilitarne la circolazione e la consultazione da parte di servizi esterni, sia per quello che riguarda la richiesta da parte dell'utente e l'invio dei contenuti da parte del repository.

Un meccanismo di questo tipo si baserà dunque su una pluralità di repository che fungano da *data provider*, che cioè rendano disponibili i contenuti, e che *espongano* i relativi metadati in un formato standard e riconoscibile (basato in genere sull'uso di file XML). Alla pluralità di data provider corrisponderanno pochi *service provider*, che saranno in grado di raccogliere automaticamente (*harvesting*) i metadati esposti dai vari data provider, di integrarli e di offrire servizi centralizzati di interrogazione e ricerca (ma anche, ad esempio, servizi statistici o di altro tipo).

Questo modello, nato nel mondo degli Open Archive (la tipologia di repository utilizzata sempre più spesso da istituzioni accademiche e di ricerca per raccogliere e distribuire in forma aperta i prodotti delle attività di ricerca: articoli, pre-print, post-print ecc.), è evidentemente applicabile a moltissimi altri tipi di repository, e l'uso del protocollo OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting), un insieme di convenzioni che disciplinano l'esposizione dei metadati e il loro harvesting, si sta estendendo dal mondo Open Archive a repository di altro genere.

La diffusione del protocollo OAI-PMH; quella ancor più generalizzata del (meta)linguaggio XML, che proprio nella gestione dei metadati ha uno dei propri principali punti di forza; il dibattito sul web semantico; il ruolo del social tagging nel web 2.0: in tutti questi sviluppi, la gestione dei metadati ha un ruolo centrale, e tutti segnalano con chiarezza la centralità che i metadati hanno assunto per la gestione dell'informazione in rete. Sempre più chiaramente, la capacità di fare del web e delle risorse di rete strumenti efficienti di gestione informativa dipende dalla capacità di creare buoni sistemi di metadati, e di utilizzarli in maniera generalizzata ed efficiente.