Metadati e *open data*: Nuovi paradigmi per vecchie professioni

Giovanni Bergamin
Biblioteca Nazionale
Centrale di Firenze

L'uso del termine metadati si è affermato con il Web tra la fine e l'inizio di questo secolo[1]. Nel mondo delle biblioteche, anche se con qualche iniziale resistenza e sospetto, il termine metadati viene ormai usato quasi in maniera intercambiabile con il termine dati catalografici[2].

Per quanto riguarda la definizione del termine metadati può essere interessante rilevare che accanto a quella generica e da manuale - dati che si riferiscono ad altri dati[3] - si sono affermate definizioni che mettono l'accento sulla struttura e sul tipo di servizio che i metadati sono chiamati a svolgere. P. Caplan definisce i metadati come «informazione strutturata su risorse informative di qualsiasi tipologia o formato»[4]. K. Coyle propone invece, specialmente ai bibliotecari, una

1 Si veda la frequenza di uso del termine metadata in *Google books Ngram viewer*: <http://books.google.com/ngrams/graph?content=metadata&year_start=1900&year_end=2005>.

2 Una efficace sintesi sulla "apparente" distinzione tra dati catalografici e metadati si può trovare in P. Caplan, *Metadata fundamentals for all librarians*, Chicago, ALA, 2003, pp 1-11.

3 "data about data". Si veda p. es.: <http://en.wikipedia.org/wiki/Metadata>.

4 P. CAPLAN, Metadata fundamentals for all librarians, cit., p. 3: "structured information about an information resource of any media type or format" dove "structured information" significa che l'informazione "must be recorded in accordance with some documented metadata scheme".

definizione di metadati come «dati costruiti per uno scopo e con l'obiettivo di facilitare un'attività umana»[5].

Le espressioni *open data*, *linked data* o *linked open data*[6] si affermano a partire dal 2006[7] e sono la testimonianza di un forte impegno progettuale per affermare nella vita quotidiana le visioni del Web semantico, proposte agli inizi di questo secolo[8] con l'obiettivo di affiancare al Web dei documenti un Web dei dati.

Il Web dei documenti è – com'è noto e in estrema sintesi – basato sulla pubblicazione di documenti contenenti dati strutturati prevalentemente ai fini della loro presentazione e fruizione da parte degli utenti, ma non strutturati per il loro riuso e recupero da parte di applicazioni. Solo per fare un esempio: due documenti tradizionali in formato HTML e pubblicati su due negozi online contenenti informazioni sul costo del medesimo oggetto, possono essere confrontati tra loro solo da un essere umano che visualizza in sequenza le due pagine. Nessuna applicazione sarà in grado in maniera affidabile di mettere a confronto i prezzi e stabilire quale è il negozio online più conveniente (mancano in questi documenti metadati strutturati che attribuiscono ad una determinata stringa di caratteri il *significato di* prezzo).

È dal 1997 che il World Wide Web Consortium propone il linguaggio RDF (*Resource Description Framework*) per pubblicare sul Web dati interpretabili senza ambiguità da parte di macchine. Occorre tuttavia dire che se il Web dei documenti si è sicuramente affermato, il Web dei dati (soprattutto con tutte le iniziative coinvolte nei *linked data*) sembra essere presente in una sorta di "universo parallelo"[9].

In generale si può dire che oggi il punto di partenza per chi ricerca informazioni (per trovare un ristorante, per sapere come iscriversi all'università, per studiare o per fare una ricerca su un determinato argomento, ecc.) è un motore di ricerca. La conseguenza è che molti ritengono che se qualcosa non è recuperabile attraverso un motore di ricerca, questa cosa semplicemente "non esiste". Naturalmente si può ritenere questa conclusione sicuramente errata, ma occorre considerare che nei fatti questa è una possibile e diffusa conclusione[10]. D'altra parte

---

5 K. COYLE, Metadata: data with a purpose, 2004, <http://www.kcoyle.net/meta__purpose.html>: "Metadata is: constructed (metadata is wholly artificial, created by human beings.); for a purpose (there is no universal metadata: for metadata to be useful it has to serve a purpose); to facilitate an activity (there's something that you do with metadata)".

6 Termini quasi sempre usati - nei testi italiani - in inglese e raramente tradotti con dati (aperti e/o) collegati.

7 T. Berners-Lee, Linked data, 2006, <http://www.w3.org/DesignIssues/LinkedData.html>.

8 T. Berners-Lee, J. Hendler, O. Lassila, The semantic web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, 2001, "The scientific American", 2001, <http://www.scientificamerican.com/article.cfm?id=the-semantic-web>.

9 R. Cyganiak, The Linking Open Data cloud diagram, 2011, <http://richard.cyganiak.de/2007/10/lod/>.

10 Riporto qui di seguito considerazioni già presentate in G. Bergamin, A. Lucarelli, The Nuovo soggettario as a service for the linked data world, in "Jlis", 4(2013), n. 1, <http://leo.cilea.it/index.php/jlis/article/view/5474>.

si può certamente sostenere che se i motori di ricerca non si interessano di Web semantico quest'ultimo ha scarse possibilità di affermarsi come infrastruttura diffusa per lo scambio dell'informazione.

In realtà i motori di ricerca sono interessati da molto tempo alla semanticità dei documenti (ai dati contenuti nei documenti strutturati secondo un determinato modello). Il recente accordo - noto come *schema.org* - tra i maggiori motori di ricerca (Google, Yahoo, Bing e Yandex) per la codifica nel linguaggio RDF dei dati all'interno delle normali pagine HTML (HTML5) può (o deve) essere anche per le biblioteche un'interessante opportunità[11].

Grazie a questa codifica – che si presenta come un'estensione molto semplice dei tag HTML delle pagine Web ma che si fonda sul linguaggio RDF – i motori di ricerca sono in grado di sfruttare i dati strutturati contenuti in un determinato documento: ad esempio posso precisare non solo la stringa di ricerca – es. fragole - ma anche *che cosa* cerco (un luogo, un film, un libro, una ricetta). Per questo motivo il Web dei dati è detto anche Web delle cose (o entità). *Schema.org* risponde insomma alla domanda: quali sono le *cose* che vengono cercate attraverso i motori di ricerca? Si tratta in altre parole di un'ontologia[12] che fornisce in maniera strutturata tutti i metadati che possono essere inseriti nelle pagine HTML.

Tra le *cose* che possono essere cercate attraverso un motore di ricerca, *schema.org* prevede anche il libro (lo troviamo tra le *CreativeWork* accanto a *Article, Blog, Diet, ExercisePlan, Movie, Painting, SoftwareApplication, TVSeries* ecc) e le biblioteche (le possiamo trovare tra i *LocalBusiness* accanto a *AnimalShelter, ChildCare, DryCleaningOrLaundry, HealthAndBeautyBusiness, HomeAndConstructionBusiness, InternetCafe).*

Naturalmente i cataloghi delle biblioteche (in gergo OPAC) sono presenti in rete e normalmente non sono indicizzati dai motori di ricerca (in alcuni casi lo sono solo parzialmente). Come sappiamo i motori di ricerca indicizzano il Web seguendo i link che trovano nelle pagine, mentre le pagine temporanee generate al volo a seguito di una ricerca bibliografica effettuata da un utente non possono essere accessibili a un motore di ricerca. Naturalmente se l'indirizzo (URL) che restituisce dinamicamente quella pagina è citato in una pagina HTML, allora il motore di ricerca può indicizzare anche quella pagina. Esistono delle tecnologie standard (tra queste la più usata è il protocollo *Sitemap*[13]) che permettono al gestore di un sito di comunicare ai motori di ricerca elenchi di indirizzi (URL) che attivano dinamicamente un risultato (per esempio tutti i prodotti di un negozio online, tutte le notizie bibliografiche presenti in un catalogo, ecc). L'uso combinato di *Sitemap* e *schema.org* consentirebbe ai cataloghi delle biblioteche di essere presenti *anche* nei risultati di una ricerca effettuata sui motori di ricerca. Si trat-

11 J. Ronallo, HTML5 Microdata and Schema.org, "Code4lib jurnal", 16(2012). <http://journal.code4lib.org/articles/6400 >.

12 Per il significato in questo contesto del termine ontologia si veda (al termine del testo) l'Allegato 1 Metadati: quadro terminologico.

13 Protocollo Sitemap <http://www.sitemaps.org/protocol.html>.

terebbe di una presenza "semantica": grazie alla strutturazione dei metadati conformi a *schema.org*: il motore di ricerca sarebbe in grado di estrarre e indicizzare i campi di una notizia bibliografica sulla base del loro "significato" (titolo, autore, soggetto ecc.).

Occorre dire che l'accordo *schema.org* viene oggi applicato gradualmente. Ad esempio Google (ma solo nella versione inglese) in questo momento sfrutta già la strutturazione dei tag HTML5 prevista da *schema org* per le ricette, ma non ancora per i libri[14]. In ogni caso la maggioranza dei motori di ricerca fornisce oggi agli sviluppatori, strumenti per verificare la conformità a *schema.org* delle pagine pubblicate[15] ed è ragionevole ritenere che in futuro l'ontologia prevista da *schema.org* sarà sempre più sfruttata dai motori di ricerca.

Attraverso i motori di ricerca aumenta quindi oggi la probabilità di far incontrare "domanda e offerta" di informazioni sul Web, con una forte valorizzazione dell'informazione prodotta anche dalle biblioteche. *Schema.org* non è (e non vuole essere) quindi una risposta alle necessità di evoluzione nella codifica del record bibliografico (andare oltre il MARC)[16], ma per il mondo delle biblioteche può essere un modo per valorizzare qui e ora l'informazione che oggi produciamo (con tutta la sua ricchezza e con tutti i suoi limiti). Più in generale grazie a *schema.org* tutto l'investimento di oggi nei *Linked* (e/o *Open*) *data* potrebbe finalmente diventare parte della vita quotidiana.

Allegato 1:
Metadati: quadro terminologico

È significativo che, nell'ambito del Web semantico, non si sia ancora raggiunto un consenso nell'uso della terminologia di settore. La tabella che segue mostra concordanze e differenze di significati e propone, nella colonna di destra, possibili usi italiani nella terminologia relativa ai metadati.

Nota: le prime tre colonne della tabella sono tratte dall'ultimo lavoro di K. Coyle. *Linked data tools: connecting on the Web*, "Technology Reports", May/June 2012, p.15. Ringrazio Anna Lucarelli per l'aiuto nella preparazione di questa tabella.

14 Si può provare a cercare il termine "pasta" nella versione inglese di Google. Si vedrà comparire nel menu dinamico prima dei risultati la categoria "Recipes". Se si sceglie quest'ultima (si limita il risultato alle sole ricette), sempre nello stesso menu dinamico comparirà l'opzione "Search tools" con la possibilità di categorizzare le ricette per "ingredienti", "tempo di cottura" e "apporto calorico". Si veda (al termine del testo) l'Allegato 2.

15 Google <http://www.google.com/webmasters/tools/richsnippets>; Bing <http://www.bing.com/toolbox/markup-validator>; Yandex <http://webmaster.yandex.ru/microtest.xml>.

16 Oggi il progetto più importante <http://bibframe.org/>.

Uso tradizionale	*Uso nel semantic web*	*Proposta K. Coyle*	*Discussione (aperta) per un uso italiano*
Data elements	Classes Properties	Elements (rifermento generale); Classes, Properties (formalmente definite in RDF o OWL)	Elementi *(in grado di ospitare dati)* P. es. (in grassetto gli elementi): <dc:title> I promessi sposi </dc:title>. Classi, Proprietà
Metadata schema	Vocabulary Ontology	Ontology	Modello di dati, Schema di metadati, Ontologia modelli/schemi funzionali di codifica e strutturazione dei dati (modelli di dati, schemi di metadati codificati e strutturati per scopi determinati). P. es.: *schema.org* si autodefinisce come data model e rientra qui: è una ontologia delle cose (entità) che possono essere messe in rilievo (possono essere indicizzate a livello di dati conformi a un modello) dai motori di ricerca; SKOS <http://www.w3.org/TR/skos-reference/> si autodefinisce come data model e il suo oggetto è costituito da entità di tipo particolare: concetti e gerarchie concettuali modellati secondo lo standard ISO per i thesauri (ora ISO 25964-1). Per SKOS l'oggetto primario è il significato di un concetto e non il significante (p. es. il termine che rende quel concetto in una determinata lingua). Un vocabolario in quanto contenitore conforme a un modello/schema funzionale di codifica dei dati rientra qui, così come le tassonomie (taxonomies) rientrano qui (in quanto contenitori).
Data	Values	Data	Dati *(ospitati dagli elementi)*. P. es. (in grassetto i dati): <dc:title> I promessi sposi </dc:title> In questo contesto è diffuso l'uso di data sets (datasets) come insieme di dati implementato secondo un modello/schema funzionale di codifica dei dati
Controlled list	Vocabulary	Vocabulary	Lista controllata, Vocabolario controllato L'insieme di dati contenuti in un vocabolario implementato secondo un modello/schema funzionale di codifica dei dati dove i requisiti per l'appartenenza dei dati all'insieme sono stabilititi da procedure formalmente definite (le liste controllate/ i vocabolari sono datasets, ma non tutti i datasets sono vocabolari controllati)

Allegato 2:
Un esempio di ricerca del termine "pasta" su Google (versione inglese)