Seth van Hooland
Ruben Verborgh

Linked Data for Libraries, Archives and Museums. How to clean, link and publish your metadata

London, Facet Publishing, 2014, 254 p.

Da qualche tempo nella letteratura professionale italiana si trovano saggi sul tema dei linked (open) data; questi lavori mettono in risalto l'importanza di queste novità per il futuro e il rilancio delle biblioteche e sottolineano la stretta affinità che, in questo ambito, lega le biblioteche alle altre istituzioni della memoria: archivi e musei.

Infatti "biblioteche, archivi e musei si stanno confrontando con la sfida di fornire accesso a collezioni in rapida crescita malgrado i continui tagli di bilancio". Secondo gli autori Seth van Hooland – professore associato presso l'Université libre de Bruxelles (ULB) – e Ruben Verborgh – ricercatore della Ghent University, Belgio – la soluzione al problema è "creare, collegare e pubblicare metadati di qualità in forma di linked data in modo che le collezioni siano individuate, viste e fruite in un modo sostenibile".

Van Hooland, che svolge consulenze per organizzazioni pubbliche e private, dirige il Master in Scienze dell'informazione e ha tenuto di recente un corso specifico sui linked data all'Information School of the University of Washington; mentre Verborgh si occupa delle relazioni tra le tecnologie del Web semantico e le caratteristiche architettoniche del web allo scopo di costruire macchine client "più intelligenti" ed è coautore di un manuale su OpenRefine, <http://openrefine.org/> (*Using Open Refine*, Birmingham, Packt Publishing, 2013), un programma client gratuito per la pulizia e la conversione di enormi quantità di dati (una via di mezzo tra un apparente foglio di calcolo e un potentissimo database, erede di Google Refine e che si autodefinisce "a power tool for working with messy data").

I due autori hanno unito le forze per scrivere un manuale di taglio pratico che insegna a moltiplicare il valore dei metadati già prodotti e disponibili nelle istituzioni culturali, pubblicandoli come linked data attraverso diversi passaggi: modellizzazione (capitolo 2); pulizia (o normalizzazione, capitolo 3), riconciliazione (riuso di vocabolari controllati, capitolo 4), arricchimento (riconoscimento delle entità, capitolo 5) e collegamento (tra insiemi di metadati eterogenei, capitolo 6).

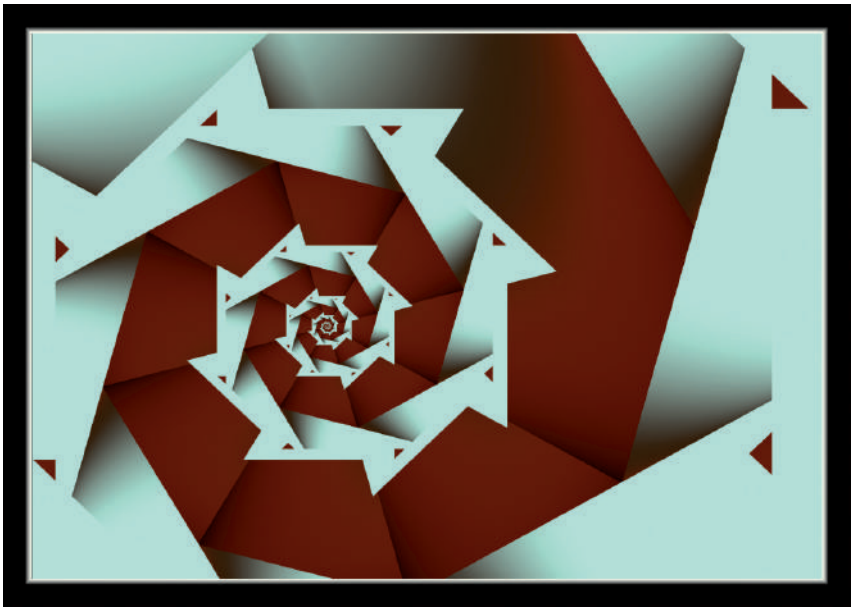
Come chiariscono gli autori, il libro ruota intorno a tre concetti essenziali: linked data, metadati e istituzioni culturali.

L'espressione *linked data* non fa riferimento a una tecnologia precisa o specifica, ma è da ritenersi un insieme di buone pratiche per la pubblicazione sul web di dati strutturati. Le tecnologie disponibili – anche gratuitamente – per raggiungere questo obiettivo sono in continua evoluzione e, di conseguenza, anche la pubblicazione di linked data è una meta in movimento, che deve essere continuamente superata e aggiornata. Pertanto, pur essendo presente alla fine di ogni capitolo una parte dedicata alle esercitazioni, il focus del libro non è su una specifica tecnologia, ma sul processo di pubblicazione, che infatti viene presentato in modo teorico, spiegato in dettaglio e poi "sperimentato" dal lettore attraverso gli esercizi proposti.

Che un libro sui linked data parli di metadati sembra tautologico, ma non lo è. Come fanno notare gli autori, i dati prodotti in forma linked data sono triple RDF (Resource Description Framework) che descrivono una risorsa e quindi, come tali, sono metadati. Se si "pone la domanda di quale sia il confine tra dati e metadati, la risposta è: non esiste. È il contesto d'uso che consente di stabilire se si è di fronte a dati o a metadati". Per esempio, in questa recensione i dati descrittivi della pubblicazione *Linked Data for Libraries, Archives and Museums*, posti in testa, sono metadati della pubblicazione originale; tuttavia, rispetto a questa recensione, essi sono i dati di partenza e la recensione costituisce un insieme di metadati rispetto alla descrizione bibliografica posta in testa. A sua volta, una futura citazione bibliografica di questa recensione si può considerare come un insieme di metadati della recensione e così via.

Il terzo concetto è quello delle istituzioni culturali; in Italia l'attenzione delle comunità di archivi e musei verso i linked data è stata molto scarsa, se comparata a quella prestata nell'ambito delle biblioteche. Anche nei diversi corsi organizzati dall'AIB, pur essendo stati pubblicizzati nelle liste specializzate e presentati come corsi aperti all'insieme delle tre comunità professionali e scientifiche, la partecipazione di archivisti e curatori di museo è stata molto ridotta. Nel contesto di questo manuale invece le istituzioni culturali sono viste nel loro complesso perché, anche se non si possono negare tradizioni culturali e professionali diverse nel trattamento dei dati, i principi e le buone pratiche per la gestione dei metadati e per la loro

pubblicazione sul web sono i medesimi; gli autori confermano in pieno l'approccio che aveva caratterizzato anche il rapporto finale del Library Linked Data Incubator Group pubblicato nel 2011. L'attenzione verso i metadati delle istituzioni culturali costituisce inoltre una caratteristica unica e specifica di questo volume rispetto ad altri manuali sui linked data pubblicati a livello internazionale (come per esempio il fortunato *Linked data: evolving the web into a global data space* di T. Heath e C. Bizer, del 2011, e il manuale pratico *Programming the Semantic Web* di T. Segaran, C. Evans e J. Taylor, del 2009). Per rispondere a questa carenza, a un'apparente generale indifferenza da parte del mondo degli archivi e dei musei e, soprattutto, per offrire al pubblico italiano un testo per conoscere un linguaggio contemporaneo che investe la comunicazione globale, in Italia è stato pubblicato *Linked data per biblioteche, archivi e musei* di Mauro Guerrini e Tiziana Possemato, volume 8 della collana "Biblioteconomia e scienza dell'informazione" dell'Editrice Bibliografica.



Il manuale di van Hooland e Verborgh è strutturato per consentire al lettore di ridurre eventuali difficoltà di natura tecnologica che gli impediscono di comprendere e sviluppare una visione critica sui linked data, dal momento che gli autori non nascondono le difficoltà e gli svantaggi che la pubblicazione di questi può comportare. Il percorso di lettura è costruito per combinare sempre teoria e pratica, secondo uno schema ripetuto in ogni capitolo: l'introduzione teorica alle diverse tecnologie – che si avvale di casi di studio reali che riprendono esperienze di istituzioni di tutto il mondo – è seguita da esercizi e prove da svolgere al PC; così i lettori poco pratici degli aspetti informatici hanno modo di valutare gli effetti concreti dei processi necessari per la pubblicazione dei linked data (normalizzazione, riuso di vocabolari, riconoscimento delle entità e collegamento con altri set di dati).

Ogni capitolo (dal 2 al 6) può essere letto e utilizzato in modo indipendente e tutti i capitoli sono stati perfezionati grazie alla sperimentazione e all'utilizzo da parte di archivisti, bibliotecari e curatori di musei nell'ambito di vari corsi di formazione sui linked data tenuti dai due autori. Come si è detto, ogni capitolo approfondisce una particolare attività necessaria per realizzare il processo di pubblicazione dei dati. Dopo il capitolo iniziale, che serve da introduzione, il capitolo 2 è dedicato alla modellizzazione dei dati. Ha lo scopo di fare comprendere la filosofia dei linked data attraverso una panoramica dei più importanti paradigmi tradizionali di modellizzazione dei dati; in questo capitolo cioè si mettono in evidenza vantaggi e svantaggi dei dati registrati in forma di semplice tabella, di database relazionale, di linguaggi di marcatura (XML) e in triple RDF. Oltre a fornire gli elementi per comprendere l'impatto della modellizzazione dei dati sui metadati prodotti, il capitolo insegna a costruire le stringhe di ricerca per interrogare i database a grafo (cioè costituiti di linked data) e ad acquisire familiarità con SPARQL attraverso l'esplorazione di DBpedia (<http://wiki.dbpedia.org/>).

Il terzo capitolo mostra perché la maggior parte dei metadati che produciamo deve essere "pulita", ovvero normalizzata: sembra un passaggio superfluo per istituti culturali come i nostri, dal momento che i dati vengono prodotti secondo standard internazionali e norme precise, e in misura sempre crescente negli ultimi decenni. In realtà, questo capitolo aiuta bene a cogliere la necessità di pulire i dati quando si legge, per esempio, che gli autori hanno identificato 52 modi diversi per codificare data. Non c'è da stupirsi: oltre ai diversi formati internazionali esistono tipi di data diversi anche nelle norme di catalogazione (data certa, data incerta, data presunta, data di cui si conoscono gli estremi, o uno soltanto ecc.). Nel momento in cui i

dati vengono pubblicati in forma di linked data, è necessario dare conto con precisione di tutte queste differenze, per un corretto trattamento da parte delle macchine.
Il capitolo presenta inoltre il concetto del *data profiling*, ovvero "l'uso di tecniche di analisi per scoprire la vera struttura, contenuto e qualità di una raccolta di dati", per il quale poi viene suggerito qualche esercizio con l'utilizzo di OpenRefine applicato ai metadati del Schoenberg Database of Manuscripts (che il lettore può scaricare sul proprio PC all'indirizzo indicato nel manuale).
Il quarto capitolo ha lo scopo di aiutare il lettore a capire le potenzialità e i limiti del riuso (*reconciliation*) dei vocabolari controllati. Dal momento che gli autori si rivolgono a un'audience più ampia dei soli bibliotecari, vengono spiegate le differenze tra schemi di classificazione, intestazioni per soggetto e thesauri e viene mostrato come, con l'aiuto di SKOS (Simple Knowledge Organization System), i vocabolari controllati possono essere rappresentati in un formato adatto al web. Attraverso un semplice editor di testo, il lettore si può esercitare a codificare manualmente un mini-thesauro e avere una dimostrazione dell'uso dell'estensione RDF di OpenRefine, mentre il caso di studio è dedicato alla riconciliazione di un insieme di metadati del Powerhouse Museum con le intestazioni tratte dalle LCSH (Library of Congress Subject Headings).
Il quinto capitolo mostra le potenzialità e i limiti dell'applicazione del riconoscimento delle entità (NER – Named Entity Recognition) ai metadati. Nel passaggio degli strumenti descrittivi (cataloghi e inventari) da cartacei a elettronici, il corpo di testo unico costituito dalla scheda nel suo complesso è stato ridotto in unità informative sempre più piccole (come, per esempio, le aree ISBD nella descrizione di oggetti bibliografici). In alcuni casi comunque gli elementi della descrizione di un libro, un documento, un manufatto, contengono testi che, per la loro natura, non possono essere ulteriormente ridotti o normalizzati (per esempio alcuni elementi dell'area delle note in ISBD, gli elementi "storia istituzionale/amministrativa" e "ambito e contenuto" di ISAD o l'elemento "descrizione del soggetto rappresentato" nelle linee guida CIDOC). Nel processo di pubblicazione dei dati sul Web semantico, è possibile anche in questi casi arricchire i metadati attraverso il riconoscimento automatico delle entità nominate nei campi descrittivi e rendere possibile una ricerca più semanticamente rilevante da parte delle macchine. L'esercitazione di fine capitolo avviene con l'utilizzo di un'estensione NER di OpenRefine, e tre diversi servizi vengono testati su un insieme di metadati prodotti dalla British Library per Europeana.
Il capitolo 6 presenta la parte finale del processo: la pubblicazione vera e propria. Ha soprattutto lo scopo di far comprendere l'approccio tecnico più sostenibile – dal punto di vista tecnico, economico e organizzativo – per pubblicare i dati delle proprie raccolte.
Il percorso di lettura proposto da questo libro è quindi tracciato in modo razionale e coerente e consente di avere una panoramica completa di tutti gli aspetti che il responsabile della pubblicazione dei linked data di un'istituzione culturale deve conoscere. Tuttavia, per la particolare impostazione data, è possibile anche che il lettore si dedichi allo studio e all'approfondimento di singoli capitoli, che sono infatti pensati ciascuno come un saggio breve e completo, teorico e pratico. Il manuale ha uno stile curato e chiaro e il testo realizza pienamente l'intento degli autori di adottare un linguaggio che consenta la lettura da parte di chiunque. Il lettore è continuamente aiutato con la ricchezza degli esempi e le molte illustrazioni, tabelle, schemi che permettono di seguire facilmente i contenuti teorici o le indicazioni pratiche per le esercitazioni proposte al termine del capitolo. La sperimentazione delle tecniche illustrate con il proprio PC costituisce senza dubbio un valore aggiunto che rende preziosa quest'opera; infatti gli esempi forniti sull'uso di SPARQL sono simili a tanti altri presenti in Rete, ma il manuale, grazie all'uso di uno specifico insieme di dati forniti per l'esercitazione, consente di verificare se il risultato è effettivamente quello che ci si aspettava.
Le introduzioni teoriche, mai lunghe né astratte, sono sempre ben documentate e l'elenco delle letture suggerite al termine di ogni capitolo (con riferimenti bibliografici completi e precisi) consente di approfondire autonomamente i singoli aspetti trattati, senza eccedere dal punto di vista quantitativo.
In conclusione, questo manuale pubblicato dalla nota casa editrice Facet Publishing costituisce uno strumento di studio e di lavoro completo, solido, efficace e concreto, ed è adatto a un pubblico molto ampio che va dallo studente universitario allo specialista dei metadati di un'istituzione culturale. Il prezzo non particolarmente competitivo (circa 70 euro) limita certamente il numero dei potenziali acquirenti personali; se ne consiglia l'acquisto da parte delle biblioteche.

CARLO BIANCHINI
Università degli studi di Pavia
carlo.bianchini@unipv.it
DOI: 10.3302/0392-8586-201506-066-1