Linked data: il mondo di internet e il ruolo delle biblioteche, degli archivi e dei musei

Margherita Aste - Maria Cristina Mataloni - Luca Martinelli
ICCU

Introduzione

Il mondo dell'informazione in rete è sempre stato fortemente interessato al patrimonio di dati provenienti da biblioteche archivi e musei, una grande quantità di dati di qualità, fortemente strutturati e rispondenti a regole condivise. D'altro canto anche biblioteche, musei e archivi sono estremamente interessati all'integrazione dei propri dati in rete al fine di garantirne maggiore visibilità e riusabilità.
Numerose sono le esperienze di collaborazione tra biblioteche, musei e archivi a livello nazionale[1] e internazionale nell'ambito dell'International Federation of Library Associations (IFLA)[2] per il trattamento comune di patrimoni culturali diversi.
Le attività sono focalizzate alla definizione di standard e strumenti concettuali per la descrizione delle proprie risorse e a garantire l'interoperabilità e la navigazione tra le informazioni.
I modelli dei dati per la descrizione dei diversi tipi di materiali sono molto diversi tra loro.
In ambito bibliotecario è stato sviluppato il nuovo standard di descrizione delle risorse RDA[3] progettato per il mondo digitale e indirizzato non solo alle biblioteche ma anche ad altre istituzioni culturali, come musei e archivi.
Nel mondo archivistico la standardizzazione e l'uniformità del linguaggio sono caratteristiche piuttosto recenti.
Nell'ambito del settore museale è stato elaborato il modello concettuale per la descrizione degli "oggetti culturali" CIDOC-CRM[4].
Nel suo realizzarsi, il progetto che prevede la pubblicazione in Linked Open Data (LOD) degli archivi del Servizio Bibliotecario Nazionale (SBN) ha messo in luce le grandi potenzialità che tale formato può avere per la stessa base dati di SBN sia a

[1] MAB, acronimo per Musei Archivi e Biblioteche, coordinamento nazionale promosso nel 2011 da AIB, ANAI e ICOM Italia.
[2] http://www.ifla.org/node/8664.
[3] Resource Description & Access vedi http://www.rdatoolkit.org/.
[4] http://www.cidoc-crm.org.

livello di visibilità nei confronti degli utenti dei tradizionali motori di ricerca, ma anche per gli stessi bibliotecari, che avrebbero delle fonti autorevoli da utilizzare come utile ausilio alla catalogazione.
Il progetto ha altresì evidenziato alcuni aspetti problematici divenendo quindi spunto di ulteriore riflessione nel merito.
La pubblicazione in LOD dei dati SBN permetterebbe anche ad altre basi dati di potersi avvalere dell'autorevole lavoro svolto dai bibliotecari.
In questo senso, l'accordo sottoscritto nel marzo scorso con l'Associazione Wikimedia Italia e la presenza di un "wikipediano in residenza" presso l'ICCU stanno già delineando una collaborazione proficua fra l'Istituto, l'enciclopedia libera Wikipedia e gli altri progetti correlati, come il database di dati strutturati libero Wikidata.

1. La base dati SBN in Linked Open Data

Il catalogo SBN può prestarsi in maniera particolarmente efficace alla fruizione in modalità Linked Open Data. È per tale motivo che il Comitato Tecnico Scientifico per SBN, nel 2014, ha dato mandato all'ICCU di costituire un gruppo di lavoro per la definizione di un prototipo di sperimentazione dei dati del patrimonio informativo delle biblioteche della rete SBN in LOD, secondo le modalità del Web semantico, così come già sperimentato da analoghe istituzioni straniere come la BNF e il RISM.
Il Gruppo di lavoro, costituito da personale dell'ICCU e del VAST LAB PIN dell'Università di Prato, ha analizzato in prima battuta il set di dati SBN che maggiormente si prestava alla pubblicazione in LOD, decidendo di utilizzare i dati che non presentavano specificità inerenti la tipologia di materiale e rinviando la mappatura degli altri documenti (musicali, grafici, cartografici e audiovisivi).
L'architettura della base dati SBN è di tipo relazionale, ma l'archivio derivato che è visibile all'utenza dell'OPAC SBN perde la caratteristica relazionale, essendo esportato in formato UNIMARC.
Questa peculiarità, che potrebbe sembrare riduttiva, può rivelarsi invece utile all'approccio Web dal momento che il formato UNIMARC risponde ad uno standard internazionale (ISO2709) e viene utilizzato da diversi importanti Paesi (ad esempio la Francia).
La prima fase del progetto ha riguardato la mappatura dei campi UNIMARC utilizzati in SBN e, quindi, la scelta delle ontologie e del modello concettuale di riferimento.
Tale scelta è stata essenzialmente guidata dalla volontà di uniformarsi a modelli e standard internazionali che consentissero l'integrazione e lo scambio di informazioni bibliografiche tra i differenti settori della cultura e ne consentissero la strutturazione in RDF (Resource Description Framework)[5].

[5] Il Resource Description Framework (RDF) è uno standard flessibile proposto dal consorzio W3C per la codifica, lo scambio e il riutilizzo di metadati. Il data model è formato da risorse, presenti nel Web con un URI, proprietà ovvero relazioni e valori, anch'essi identificati da URI.

2. Analisi

Come accennato, l'analisi iniziale si è concentrata principalmente sulla scelta del modello concettuale e del tipo di ontologia da utilizzare, tenendo conto anche del fatto che una mappatura dei tag UNIMARC in RDF non è stata ancora mai realizzata. La scelta è caduta sui seguenti componenti:

- il CIDOC Conceptual Reference Model (CRM)[6], modello concettuale di riferimento, che costituisce la struttura formale per descrivere i concetti impliciti ed espliciti e le relazioni nella documentazione del patrimonio culturale;
- il Functional Requirements for Bibliographic Records object oriented (FRBRoo)[7], ontologia di riferimento che fornisce un sistema concettuale dotato di una complessa rete di classi e proprietà in grado di descrivere le principali entità e relazioni dell'ambito bibliografico, permettendo l'integrazione e lo scambio di dati di diversa tipologia e formato nell'ambito dei beni culturali.

Attraverso le classi e le relazioni di FRBRoo si arriva alla costruzione di *path* o "catene di relazioni" per mezzo delle quali è possibile esplicitare concetti anche complessi. Le operazioni di mappatura consistono essenzialmente nel creare di volta in volta corrispondenze fra ognuno dei campi UNIMARC (solitamente sintetici e semanticamente difficili da esportare al di fuori dell'ambito biblioteconomico) e uno o più *path* FRBRoo equivalenti. In questo modo si rende possibile la lettura semantica dell'oggetto bibliografico, che viene arricchito ulteriormente di relazioni a volte ricavate da altri dati standardizzati già esistenti sul Web (ad esempio Geonames, *authority file* online quali il VIAF, etc.).

La seconda fase di analisi ha riguardato la scelta della tipologia dei dati da trattare. Inizialmente si è posta attenzione alle notizie catalogate con un set minimo di dati comuni[8]; in una fase successiva sono stati mappati anche i dati specifici. Sulla base dei criteri citati sono stati selezionati 300 record relativi a documenti che descrivono monografie moderne, monografie antiche e periodici.

In relazione allo schema concettuale FRBR, va precisato che le informazioni presenti nello scarico UNIMARC SBN si riferiscono essenzialmente alla Manifestazione

[6] Il CIDOC CRM è il risultato dell'attività svolta per oltre 10 anni da parte dei gruppi di lavoro costituiti nel CIDOC in seno al CIDOC ICOM's International Committee for Documentation. Dal 12 settembre 2006, il CIDOC CRM è riconosciuto come standard ISO 21127:2006 e nel dicembre 2014 è stato aggiornato nella nuova versione ISO 21127:2014.

[7] L'ontologia FRBRoo è nata nel 2003 dal gruppo di lavoro FRBR/CIDOC CRM, che ha avuto lo scopo di allineare i due modelli di dati utilizzati in ambito bibliotecario (FRBR) e museale (CIDOC CRM) al fine di garantirne l'interoperabilità semantica.

[8] Per set minimo di dati comuni in SBN si intende la descrizione di una notizia bibliografica con i soli dati essenziali per l'identificazione del documento senza entrare nel merito delle caratteristiche catalografiche più specifiche di ciascuna tipologia di materiale. Ad esempio, si può descrivere una diapositiva riportando il titolo, l'autore, i dati di pubblicazione senza dare indicazioni sul colore, dimensione, ecc.

del documento, essendo praticamente prive di riferimento all'Opera, all'Espressione e ai dati gestionali, tipici questi ultimi del livello Item e presenti nelle basi dati locali.
Per quanto riguarda gli archivi di autorità, al momento viene esportato in UNIMARC solo quello relativo agli autori, l'unico che è stato quindi mappato.

3. Mappatura

Il lavoro di mappatura è stato realizzato a partire dal tracciato del record UNIMARC che si presenta strutturato in campi, sottocampi e indicatori, con l'esclusione, in un primo momento, dei tag relativi alle specificità di Grafica, Cartografia, Musica e Audiovisivi e di quelli non utilizzati in SBN.
Nelle operazioni di mappatura, effettuate utilizzando il documento "FRBR object-oriented definition and mapping to FRBR-ER" (version 0.9 draft)[9], sono state create le corrispondenze fra campi, sottocampi e indicatori UNIMARC e uno o più *path* FRBRoo.
Ciò ha permesso di sviluppare le procedure di codifica in linguaggio RDF delle informazioni presenti nel catalogo SBN, rendendo i dati in formato *machine readable* e subito condivisibili e consultabili sul Web.
Tra l'altro, diversi campi mappati hanno il loro corrispondente valore codificato in tabelle, alcune ad uso interno, altre compatibili con gli standard UNIMARC.

4. Normalizzazione, arricchimento e conversione dei dati

4.1. Normalizzazione dei dati

La conversione dei dati UNIMARC in triple RDF ha richiesto una fase di normalizzazione mirata all'analisi della coerenza e correttezza delle informazioni.
La qualità del dato in entrata rappresenta un aspetto fondamentale durante la fase di conversione, perché un dataset con valori non corretti può rendere inefficienti o addirittura impossibili alcune operazioni di confronto, di similitudine e di allineamento dei dati, anche al fine di possibili link con database esterni.
La base dati SBN presenta un numero elevato di record in cui molti campi sono inseriti in modalità testo libero che mal si prestano ad una normalizzazione automatizzata; va considerato, inoltre, che le numerose importazioni di dati avvenute in maniera automatica provenienti da archivi non SBN hanno a volte compromesso l'uniformità dei dati. Questi problemi andrebbero analizzati e approfonditi ulteriormente, sia ai fini dei collegamenti con risorse esterne, sia per consentire un miglior risultato nelle ricerche mirate dell'utente finale. In particolare, va sottolineato che la produzione di LOD di qualità dipende da un buon livello di pulizia dei dati del catalogo SBN.

[9] Il documento è disponibile al link http://www.ifla.org/files/assets/cataloguing/frbrrg/frbr-oo-v9.1_pr.pdf.

4.2. Arricchimento dei dati

Una delle maggiori potenzialità offerte dai Linked Open Data è la possibilità di ampliare i propri dati in ricchezza di contenuto, quantità e qualità grazie alle informazioni che possono essere associate, attraverso un collegamento stabile, a fonti esterne autorevoli che possiedono già una descrizione accurata di numerose entità. Per tale motivo è indispensabile scegliere con particolare attenzione un *Uniform Resource Identifier* (URI), ossia una stringa alfanumerica *stabile* che identifichi *univocamente* un determinato oggetto della base dati, in maniera tale da renderlo facilmente identificabile e collegabile.
Nel nostro caso, l' URI è così strutturato:

– il dominio http://id.sbn.it/ come radice dell'URI;
– l'identificativo univoco del record.

Un esempio di URI di documento è il seguente: http://id.sbn.it/BVE0391786.
Per quanto riguarda i dati SBN, è necessario tener presente la possibilità che un record possa essere cancellato o fuso con un altro. Allo scopo di non vanificare la funzione dell'URI, sarà dunque necessario prevedere un sistema di re-indirizzamento che permetta, nel caso di fusione fra due record, di essere automaticamente reindirizzati al nuovo URI, tramite il vecchio collegamento, mantenendo così il riferimento all'oggetto originario.
La base dati SBN potrà essere collegata facilmente ad altre fonti esterne (ad es. il Nuovo Soggettario di Firenze, la Classificazione Decimale Dewey, l'Anagrafe delle biblioteche italiane, ISBN, etc.). Oltre a questi archivi, saranno aggiunti collegamenti con i *repository* di Geonames, Edit16, VIAF e OPAC SBN (per quanto concerne le schede di autorità), al fine anche di aumentarne la visibilità all'esterno. I legami alle entità dei suddetti *repository* sono stati introdotti in modo puntuale all'interno del prototipo a titolo esemplificativo e sono stati tradotti in triple RDF, memorizzati e consultabili all'interno del *triple store*.

4.3. Conversione dei dati

La fase successiva ha riguardato la conversione dei dati dal formato UNIMARC al formato RDF.
Il subset di record bibliografici, una volta normalizzato e arricchito nei campi dimostrativi per il prototipo, è stato convertito in triple RDF attraverso la mappatura FRBRoo, applicando strumenti e procedure sviluppati *ad hoc*, per essere poi caricato all'interno di un prototipo di *triple store* e sottoposto a test per verificare la coerenza e correttezza della mappatura utilizzata.

4.4. Pubblicazione

La fase successiva della sperimentazione ha riguardato la pubblicazione dei dati convertiti in RDF, che sono stati riversati all'interno di un *triple store* per renderli

accessibili all'utenza finale. Per la pubblicazione degli stessi sono state prese in considerazione piattaforme che, oltre al *repository* RDF e l'*endpoint* di interrogazione SPARQL, avessero a disposizione interfacce aggiuntive per la costruzione di servizi. La scelta è caduta sulla piattaforma *open source* Aduna Sesame che rispetto all'altra esaminata, OpenLink Virtuoso, ha evidenziato un'ottima efficienza nella gestione di database *in-memory*, considerando anche il numero limitato di record processati dal prototipo.
Le triple RDF disponibili nel triple store Sesame sono accessibili mediante:
– consultazione e download delle risorse sotto forma di file RDF/XML;
– *endpoint* SPARQL, mediante il quale è possibile ottenere informazioni a seguito dell'inserimento di una *query*;
– API che permettono la creazione di servizi aggiuntivi per facilitare le *query* sul *triple store*.

A questa prima fase sperimentale ne seguirà un'altra che prevede l'utilizzo di una piattaforma di creazione delle triple, *storage* e la pubblicazione dei risultati del tutto nuova.
L'ICCU infatti collabora al progetto del Polo digitale che interessa 5 Istituti Culturali di Napoli (Società Napoletana di Storia Patria, Pio Monte della Misericordia, Istituto Italiano per gli Studi Storici, Fondazione Biblioteca Benedetto Croce, Cappella del Tesoro di San Gennaro) che ha tra i suoi obiettivi la realizzazione di una piattaforma *open source* che consentirà la gestione dei complessi processi di lavoro finalizzati alla digitalizzazione, conservazione e divulgazione di beni culturali documentali archivistici, bibliografici e museali (dipinti, oggetti d'arte, disegni). La piattaforma prevede la conversione dei metadati bibliografici digitali in LOD utilizzando la mappatura realizzata dall'ICCU.

4.5. Aspetti da definire

Sono state analizzate infine una serie di criticità emerse nel corso delle attività, che riguardano sostanzialmente i dati del catalogo e il modello prescelto del CIDOC CRM/FRBRoo. Di seguito si riassumono le criticità inerenti i dati e le relazioni SBN:
– presenza di dati non normalizzati: si tratta di dati inseriti dai catalogatori in forma testuale e non codificata che inevitabilmente non risultano indicizzabili e di conseguenza di difficile ricerca;
– l'archivio d'autorità Autori presenta un numero limitato di schede di autorità rispetto al numero complessivo di autori presenti nell'Indice SBN. Basti pensare che, a fronte di circa 4 milioni di record autori in Indice sono presenti soltanto 204.000 schede di autorità controllate. Inoltre, in diversi casi, tali schede sono carenti delle informazioni necessarie all'identificazione univoca dell'autore (ad esempio, data di nascita e morte);

– per quanto riguarda gli altri archivi d'autorità (titolo uniforme, luoghi, marche, ecc.) le relative schede non vengono attualmente esportate in OPAC, in quanto non ne è ancora stata definita la modalità di compilazione;
– le entità FRBR non sono tutte rappresentate nell'architettura SBN, che prevede la descrizione dell'entità Manifestazione (*Manifestation*), ancora parzialmente quella relativa all'Opera (*Work*), mentre non sono rappresentate quelle riguardanti l'Espressione (*Expression*) e l'Esemplare (*Item*);
– impossibilità di riferirsi a identificativi persistenti (ad esempio, BID e VID) per la natura stessa della catalogazione partecipata che prevede fusioni e cancellazioni dei record da parte delle biblioteche;
– la mole di dati contenuti nell'archivio SBN (circa 14 milioni di titoli) rende di difficile attuazione la pulizia e la normalizzazione dei dati stessi.

A proposito del modello FRBRoo le criticità riguardano essenzialmente la difficoltà di tradurre la totalità dei dati e dei legami presenti in SBN. Nel caso dei periodici, ad esempio, non è stato possibile completare la mappatura fra SBN e FRBRoo in quanto nel modello non sono rappresentati i dati relativi ad alcuni legami (tag 423; 431; 434; 447, rispettivamente "Pubblicato con", "Continuazione in parte di", "Ha assorbito", "Si fonde con").
Si attende il consolidamento del lavoro del gruppo costituto in ambito ISSN, che ha elaborato l'evoluzione e l'estensione di FRBRoo alle risorse in continuazione[10]. Il documento prodotto, PRESSoo (versione 1.0)[11], non è ancora uno standard IFLA, ma può essere considerato un valido strumento per la mappatura delle relazioni tra risorse in continuazione.
Il Gruppo LOD SBN sta inoltre seguendo a livello internazionale altre iniziative, quale il progetto di sviluppo della rappresentazione di UNIMARC in RDF presentato nel 2013 al convegno IFLA con il documento "The UNIMARC in RDF project: namespaces and linked data"[12].

5. La collaborazione in corso con Wikimedia Italia

Come già accennato nell'introduzione, lo scorso 9 marzo l'ICCU ha sottoscritto un accordo triennale con l'Associazione Wikimedia Italia (WMI), che stabilisce una collaborazione volta al riutilizzo e all'integrazione dei dati e dei materiali dei progetti ICCU con i progetti Wikimedia e con il progetto OpenStreetMap[13].

[10] Il gruppo è costituito da: ISSN, International Centre ISSN IC, ISSN Review Group e Bibliotèque Nationale de France. Si veda anche http://www.issn.org/the-centre-and-the-network/our-partners-and-projects/pressoo/.
[11] Il documento è disponibile al link http://www.issn.org/wp-content/uploads/2014/02/PRESSoo_1-0.pdf.
[12] Il documento è disponibile al link http://library.ifla.org/156/1/222-willer-en.pdf.
[13] L'annuncio ufficiale è disponibile al link: http://www.iccu.sbn.it/opencms/opencms/it/archivionovita/2015/novita_0008.html.

La realizzazione pratica dell'accordo è affidata a un "wikipediano in residenza", ossia a un membro della comunità wikimediana che si occupa tanto di operare direttamente sui progetti Wikimedia, quanto di formare il personale dell'Istituto all'uso dei progetti[14].

In questi primi mesi[15], la collaborazione si è concentrata sull'Anagrafe delle Biblioteche e CulturaItalia, che sono stati considerati un buon punto di partenza per valutare la fattibilità di una collaborazione. Al 16 novembre 2015, sono stati generati 111 collegamenti da Wikipedia verso l'Anagrafe delle biblioteche[16] e 306 collegamenti verso CulturaItalia[17].

In aggiunta, si è approfondito il lavoro di sincronizzazione delle schede di autorità di SBN di livello AUF con il progetto Wikidata, database di conoscenza libera e multilingue appartenente alla famiglia dei progetti Wikimedia[18].

La sincronizzazione è iniziata il 6 aprile 2013 su input della comunità di Wikidata, attraverso la creazione della proprietà[19] P396, che identifica univocamente gli "identificatori SBN"[20] e genera un collegamento fra un elemento[21] e la relativa scheda su SBN. Il collegamento è gestito tramite il link di re-indirizzamento apposito http://id.sbn.it/af/ seguito dal codice della scheda[22]. Al 16 novembre 2015, sono già 13.471 i collegamenti confermati fra Wikidata e SBN sulle 64.502 schede di livello AUF finora rese disponibili.

Questo lavoro di sincronizzazione fra il lavoro svolto indipendentemente dall'ICCU e quello svolto dalla comunità wikimediana è ovviamente solo un primo passo, volto a preparare la base per la seconda fase di collaborazione in cui l'ICCU potrà sfruttare (previa validazione) i dati provenienti dai progetti Wikimedia e la comunità wikime-

[14] Il termine di "wikipediano in residenza" è derivato da quello di "artista in residenza", ossia dalla definizione di un artista contemporaneo che viene ospitato (spesso in forma retribuita) da un privato o da una istituzione e che "ricambia" l'ospitalità ricevuta tramite la creazione di opere d'arte che restano poi nella disponibilità del soggetto ospitante.

[15] Per motivi di sintesi, ci si concentrerà qui solo sui principali punti di collaborazione finora sviluppati. Una panoramica più completa dei risultati ottenuti è disponibile al link https://it.wikipedia.org/wiki/Progetto:GLAM/ICCU.

[16] La lista di collegamenti è disponibile al link https://it.wikipedia.org/wiki/Categoria:Codice_ISIL_letto_da_Wikidata.

[17] La lista di collegamenti è disponibile al link https://it.wikipedia.org/wiki/Categoria:Codice_CulturaItalia_letto_da_Wikidata.

[18] Il progetto è visibile all'indirizzo https://www.wikidata.org/.

[19] Su Wikidata, una "proprietà" è l'equivalente di un tag UNIMARC.

[20] La proprietà è visibile all'indirizzo https://www.wikidata.org/Property:P396.

[21] Un "elemento" è equivalente alla scheda di un database. Ciascun elemento riguarda un determinato soggetto del "mondo reale" (come un personaggio reale o fittizio, un oggetto, un concetto, un luogo, un evento, eccetera) e contiene al suo interno tanto i dati fondamentali riguardo quel determinato soggetto quanto i collegamenti ai singoli progetti Wikimedia finora integrati con Wikidata. Al 16 novembre, i progetti integrati sono 617 in più di 290 lingue differenti.

[22] Per esempio, il link http://id.sbn.it/af/IT\ICCU\CFIV\002049 corrisponde alla scheda di autorità in OPAC su Giacomo Leopardi (IT\ICCU\CFIV\002049).

diana potrà sfruttare le operazioni di controllo e validazione effettuate dall'ICCU, in una dinamica di interscambio dei dati che sia profittevole per entrambi.

6. Sviluppi futuri

L'Istituto, dopo la necessaria fase di test e la messa a punto della mappatura con i dati di SBN, intende proseguire il percorso della condivisione di altre basi dati in LOD sia gestite al suo interno, sia partecipando ad altri progetti di settore oltre quello con WMI. Nello specifico ha costituito un gruppo di lavoro finalizzato all'analisi dei LOD per il profilo della base dati EDIT16 (Censimento nazionale delle edizioni italiane del XVI secolo) per valorizzare le risorse di tale base dati con l'arricchimento del contenuto informativo e della rete di relazioni.
Nel lavoro di mappatura dei dati con il modello concettuale di riferimento, si partirà dalla documentazione prodotta per i LOD del Catalogo SBN e verrà sviluppata la mappatura degli elementi aggiuntivi e peculiari (marche, luoghi, editori) per la conversione in RDF del profilo dati di EDIT16.
L'ICCU inoltre sta collaborando con il progetto "Polo digitale degli Istituti culturali di Napoli" per il quale fornirà la documentazione prodotta dal gruppo di lavoro e avvierà quindi una fase di sperimentazione che prevede l'integrazione di basi dati bibliografiche archivistiche e museali che verranno pubblicate in LOD.

L'ultima consultazione dei siti Web è avvenuta nel mese di dicembre 2015.