JLIS.it

Linked open data per i nuovi servizi bibliotecari: l'esempio di data.bnf.fr

Romain Wenz

I cataloghi delle biblioteche sono stati progettati per localizzare i libri e per gestire le collezioni. Sono utilizzati dai bibliotecari per dare un ordine ai libri e dagli utenti per trovare le risorse. Tuttavia può essere difficile per un utente raggiungere le informazioni della biblioteca sul web, soprattutto perché possono coesistere numerosi cataloghi in una biblioteca. In sostanza, per diversi tipi di collezione sarebbero necessari diversi di strumenti. Per esempio, una raccolta d'archivio necessita di una struttura gerarchica, che possa descrivere i documenti nel loro insieme, così come sono stati prodotti e ricevuti durante le attività di una persona. Pertanto la gestione dei documenti può costituire un obiettivo diverso da quello di facilitarne l'accesso. Gli utenti del web hanno nuove aspettative e nuove abitudini in un ambiente mutevole. I dati della biblioteca dovrebbero rispondere a queste esigenze e appartenere realmente al web. Le biblioteche stanno cercando di rendere i loro dati veramente utili sul web. Ci concentreremo sul caso di data.bnf.fr,[1] un progetto della Bibliothèque nationale de France che si basa su link efficienti, tecniche automatiche e strumenti del web semantico.

[1] http://data.bnf.fr.

Traduzione italiana a cura di Silvia Delfitto.

Nuove aspettative

I cataloghi online rendono le cose differenti. Con il world wide web i ricercatori hanno accesso a moltissime risorse da un unico computer, anche da casa. Oggi c'è una certa concorrenza tra i fornitori di documenti, in quanto è molto facile passare da uno all'altro. Per esempio, la copia di un libro sarà meno necessaria nel momento in cui essa è digitalizzata e disponibile online. Le teorie di Walter Benjamin hanno dimostrato che il contenuto perde il suo valore nel momento in cui viene copiato con procedimenti di tipo industriale, cosa che è particolarmente vera nel mondo digitale. A sua volta, l'accesso online crea un altro tipo di valore, almeno per quanto riguarda le risorse di tipo culturale ed educativo, che sono destinate ad essere divulgate. Le risorse messe a disposizione dalle biblioteche devono essere facili da trovare, perché diventano parte di una più generale ricerca sul web. Ci sono sempre più documenti online. Molti siti web specifici forniscono informazioni che possono essere paragonate a ciò che può essere pubblicato nei libri. Inoltre collezioni digitali pubblicate dalle biblioteche diventano parte del web. Per i lettori che sono alla ricerca di risorse su internet, testi di libri digitalizzati forniscono informazione, come altre pagine web. Per questo motivo le collezioni digitali, ma anche riferimenti online di libri fisici devono essere facili da trovare e accessibili attraverso programmi automatici.

È possibile trovare documenti senza nemmeno sapere della loro esistenza. In genere, gli utenti, facendo oggi ricorso ai potenti motori di ricerca, effettuano comunemente le loro ricerche usando parole chiave associate al documento finale. Questo comportamento si è diffuso da una decina d'anni. Ha reso tutti i tipi di informazioni online sempre più facili da trovare, attraverso l'uso di motori di ricerca. Ciò implica che gli utenti tendono a cercare con parole chiave associate insieme, a differenza, ad esempio, dall'utilizzo di una serie

di campi come avviene nei cataloghi. Significa anche che risultati ordinati per algoritmi sono comunemente accettati. Abbiamo tutti familiarità con frasi del tipo risultati 1-10 di 120.000: il rumore non è una difficoltà, se viene dopo risultati pertinenti, trovato automaticamente e presentato prima. Come potrebbero le biblioteche trarre vantaggio da questi sviluppi? Varie fonti di informazione ci possono aiutare allo scopo di decidere come adattarsi. In primo luogo, le statistiche delle interfacce delle nostre ricerche locali forniscono un riscontro preciso e gratuito su ciò che i nostri utenti tradizionali cercano. Per esempio, alcuni anni fa le persone cercavano le opere complete degli scrittori, sapendo in anticipo cosa avrebbero trovato nel libro. Oggi, cerchiamo soprattutto i singoli libri, di solito attraverso il titolo e il nome dell'autore. Sondaggi pubblici di biblioteche o altre istituzioni mostrano che l'utilizzo dei motori di ricerca per navigare nel web è oggi diventato una pratica usuale. Gli utenti di internet di solito trovano i riferimenti bibliografici online prima di andare fisicamente nella biblioteca. Questo è confermato da tutte le indagini sugli utenti fatte negli ultimi anni, riferite sia a studenti che a ricercatori, come quelle realizzate da OCLC,[2] e dalla Bibliothèque nationale de France.[3] Dunque, i riferimenti bibliografici che non possono essere rintracciati online sono quasi inutili per la maggioranza delle persone. Se i bibliotecari vogliono che tali riferimenti siano trovati, devono metterli sul web. La maggior parte dei cataloghi è disponibile online all'interno di uno specifico portale. Essi non sono generalmente accessibili da link sul web, ed è impossibile che siano richiamati dai motori di ricerca. Queste nuove aspettative da parte del pubblico sono essenziali per le biblioteche, a causa delle dimensioni del contenuto di proprietà delle biblioteche. La quantità di contenuto e di informazione disponibile è così grande che le

[2]Per esempio http://www.oclc.org/reports/onlinecatalogs.

[3]Per esempio http://www.bnf.fr/documents/enquete_gallica_2011_rapport.pdf.

tecniche più recenti devono essere usate per gestirli. Per esempio, la Bibliothèque nationale de France mostra 1,5 milioni di oggetti in Gallica,[4] che è la più grande biblioteca digitale di lingua francese, e 12 milioni di record bibliografici, grazie al deposito legale dei documenti editi in Francia. Migliaia di manoscritti ed archivi sono anche disponibili, con tutti i tipi di risorse, dai manoscritti medievali a fondi archivistici di scrittori moderni. La gestione di questo tipo di risorse crea problemi di scala diversi, in quanto si tratta di milioni di documenti. Ci sono sempre duplicati, e la qualità dei dati è irregolare, conseguenza della lunga storia del nostro catalogo. Inoltre, libri a stampa e manoscritti sono generalmente descritti con logiche diverse, all'interno dei cataloghi. I record del catalogo principale descrivono un libro fisico, di solito in formato MARC. Esse sono strutturate volutamente attorno a una collezione che è stata costruita appositamente, con una serie di libri che sarebbero stati archiviati insieme per dare un senso all'utente finale. Invece gli archivi sono stati prodotti e ricevuti durante le attività di una persona, e ritenuti in un certo senso come derivati della vita e delle attività di una persona o di un'organizzazione. I documenti sono stati riuniti secondo questa logica, che non è sempre ovvia per l'utente finale oggi. Pertanto, i documenti non possono essere descritti con record semplici, ma con il modello di una struttura gerarchica, che permette di comprendere la logica originaria degli archivi. Il formato che viene comunemente utilizzato per questo tipo di risorse alla Bibliothèque nationale de France è XML-EAD (Encoding Archive Description). Le collezioni digitali, disponibili in Gallica, vengono descritte con un semplice formato: Dublin Core. Tutti gli oggetti digitali sono accessibili con un identificatore persistente (ARK), dato e mantenuto dalla Bibliothèque nationale de France. Tra questi cataloghi devono essere forniti link efficienti,

[4] http://gallica.bnf.fr.

in modo che gli utenti possano navigare in modo rapido e passare semplicemente da un documento all'altro. Le macchine non sono intelligenti, quindi è necessario fornire informazioni strutturate nei cataloghi, con link efficienti tra i documenti.

Importanza di link efficienti: principi in data.bnf.fr

Se vogliamo che le risorse informative siano realmente parte del web, nel senso che gli utenti possano citarle nei siti, nei blog, nelle pagine web e nelle e-mail, e possano accedervi seguendo i link, dobbiamo assegnare loro degli identificativi appropriati e conformi agli standard del web. In questo modo è anche possibile collegare le risorse dei nostri diversi set di dati. Poiché le grandi biblioteche hanno il più delle volte diversi cataloghi, creare collegamenti fra di essi permette di trovare le risorse senza dover sapere come si usano i diversi strumenti, ma semplicemente andando a naso. Ciò rende possibile gestire su larga scala i dati delle biblioteche, con diversi tipi di documenti. Questo è assai importante poiché molte distinzioni tra le diverse tipologie dei documenti sono state fatte prima dell'avvento del web. Per esempio, per l'utente finale, un normale libro digitalizzato e un manoscritto medievale digitalizzato possono essere equivalenti, nel senso che lo stesso utente può accedervi allo stesso modo se essi sono online. La stessa nozione di collezioni speciali può cambiare se esse sono digitalizzate e disponibili sul web. Questa forma di apertura è ottenuta grazie ai processi di digitalizzazione. Nel contesto della digitalizzazione, molte risorse che erano interessanti solo per gli studiosi specializzati sono diventate rilevanti per un pubblico più ampio. Per esempio, miniature medievali sono sorprendentemente utilizzate da un pubblico molto ampio, una vol-

ta che sono online. La strada per la ricerca deve essere semplice per queste risorse. Devono essere facili da trovare all'interno di database, ma disponibili anche tramite link sul world wide web. In generale i cataloghi, così come le informazioni sulle collezioni digitali e i dati che descrivono i documenti, devono essere tecnicamente disponibili, ma anche legalmente riutilizzabili, se vogliamo che siano ampiamente diffusi. È per questo che molte biblioteche cominciano ad adottare le tecniche del semantic web, insieme alle licenze degli open data. In questo modo, alcune biblioteche diventano parte del movimento dei linked open data e sono coinvolte nello sviluppo del semantic web. La Bibliothèque nationale de France sta sviluppando un nuovo progetto, che unisce i dati dei cataloghi (MARC), degli archivi (EAD) e delle risorse digitali (DC). Tutti i dati sono estratti e raccolti automaticamente. Questo progetto, chiamato data.bnf.fr è un progetto ancora giovane, poiché è online da un anno. Data.bnf.fr raccoglie le informazioni descrittive, e collega direttamente ai cataloghi online e ai documenti digitali. Ci sono diversi aspetti: un primo obiettivo è quello di rendere l'informazione compatibile con gli standard del web semantico, fornendo gli identificativi per le risorse, con una prospettiva RDF sulle informazioni disponibili. Per la biblioteca, la raccolta di informazioni circa i concetti di opere, autori e soggetti implica anche di lavorare su questioni di modelli. Infatti, si tratta di una prima occasione per implementare il modello FRBR, e per utilizzarlo con corrispondenze e allineamenti automatici. Per fare questo usiamo un software gratuito, chiamato CubicWeb.[5]

Al fine di rispettare i principi del web semantico, è necessario fornire informazioni descritte con vocabolari comuni, con una struttura rigorosa, come sarebbe all'interno di un database.

Questo può essere fatto solo con identificativi assegnati a tutti

[5]Sito web e documentazione ahttp://www.cubicweb.org e http://docs.cubicweb.org.

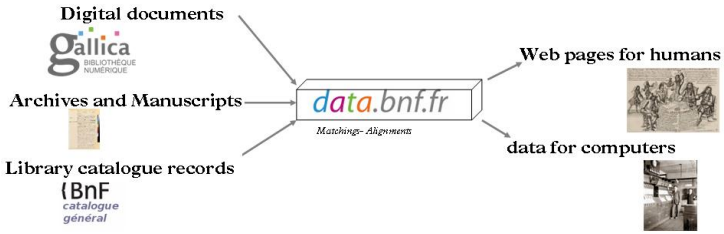

Figura 1: data.bnf.fr.

i concetti che devono essere gestiti. La Bibliothèque nationale de France usa identificativi persistenti per web URIs: gli identificativi ARK,[6] che vengono usati per identificare record del catalogo, risorse archivistiche, oggetti digitali da Gallica, e authority record. Questi identificativi ARK sono anche utili per citare queste risorse, con un resolver comune: ad esempio, un oggetto digitale[7] sarà accessibile anche con un collegamento permanente.[8] Per raccogliere informazioni su opere, scrittori e soggetti, data.bnf.fr si basa sugli authority files. Gli identificativi ARK che sono stati dati agli authority record sono utilizzati anche per le pagine in data.bnf.fr, al fine di costruire URIs affidabili e link efficienti tra i dati.

Casi d'uso per Linked Open Data (LOD)

I concetti provenienti da authority files sono essenzialmente utilizzati per l'identificazione e la descrizione affidabile di scrittori, libri e soggetti. Successivamente, la difficoltà è quella di fornire link utili a edizioni di libri, manoscritti, archivi, immagini. Alcuni esempi

[6] Archival Resource Key.
[7] http://gallica.bnf.fr/ark:/12148/bpt6k134521m.
[8] http://ark.bnf.fr/ark:/12148/bpt6k134521m.

online possono spiegare come data.bnf.fr fa fronte a questi problemi. Per esempio, per la ricerca di informazioni su uno scrittore come Goldoni,[9] l'utente troverà tutte le sue opere maggiori con l'indicazione di tutte le loro diverse edizioni. Ci sono link ai riferimenti delle edizioni presenti nei cataloghi, agli oggetti digitali, e ai manoscritti come anche le lettere che Goldoni ha ricevuto. Gli authority file sono stati estratti, e gli identificativi sono usati per fare link persistenti, e per evitare duplicati. Le pagine di Opere vengono create a livello di FRBR. Questo significa che il libro è descritto come un concetto e non come una particolare edizione di quest'opera. Tutte le edizioni vengono raccolte attorno a questo concetto. Per esempio, nella pagina sui Trionfi di Petrarca[10] è possibile trovare un elenco dei manoscritti e dei libri stampati, con i link a tutti i prodotti digitali se sono disponibili. Gli scrittori sono ovviamente collegati con le loro opere, con i documenti associati, e con altri scrittori. Questo tipo di informazione è estratta dai documenti connessi ad entrambi. Per uno scrittore come Leonardo Bruni, una sola pagina[11] contiene link a tutte le edizioni, ai manoscritti e agli oggetti digitali. I testi da lui scritti, tradotti, modificati o commentati sono disponibili separatamente, a seconda del ruolo da lui rivestito in ciascun documento. L'utente può cercare le traduzioni che ha fatto, per esempio. Ci sono anche link alle pagine di scrittori a lui associati, come Cicerone e Aristotele. Poiché era un editor di Cicerone, forniamo sia i riferimenti ai libri curati, sia un link alla pagina di Cicerone.[12] Gli strumenti del web semantico e i link affidabili ci permettono di creare pagine intorno a proprietà comuni e di dedurre nuove relazioni dal nostro grafo RDF. Per esempio si può trovare:

[9] http://data.bnf.fr/11905320/carlo_goldoni.
[10] http://data.bnf.fr/11953648/petrarque_les_triomphes.
[11] http://data.bnf.fr/12027636/bruni_leonardo.
[12] http://data.bnf.fr/11885977/ciceron.

- pagine data.bnf.fr associate alla data 1515;[13]
- pagine data.bnf.fr associate alla data 1789;[14]
- tutti gli autori che hanno coniato monete, come Luigi XIV.[15]

Data.bnf.fr crea link e pubblica pagine web con circa 2,5 milioni di risorse collegate. I dati completi sono presentati anche in RDF e disponibili cliccando sull'icona RDF nella parte inferiore delle pagine aggiungendo i seguenti suffissi all'URL: NT, N3, RDF-XML, secondo il formato necessario,[16] attraverso la negoziazione dei contenuti, utilizzando un browser web, dall' URL, o tramite downloads di massa.[17] Finora (estate 2012), il set di dati completo disponibile è di 6.3 milioni di triple, che non è di massa enorme, considerando i 2.5 milioni di risorse, grazie ai link adeguati che ci evitano un'eccessiva ridondanza. Tutti i dati non elaborati sono presentati anche in RDF e disponibili con una licenza open license. Consentendo tutti i tipi di usi, anche a fini commerciali, il che non era ovvio.

Perché usiamo una open license per data.bnf.fr. Requisiti legali e tecnici

La presentazione di informazioni sul web significa che l'istituzione è responsabile della pubblicazione di documenti. Dal punto di vista giuridico, la biblioteca diventa responsabile per il contenuto che viene visualizzato. Dal punto di vista del marketing, pubblicare le informazioni sul web è un incentivo a concentrarsi su ciò che si

[13] http://data.bnf.fr/what-happened/date-1515.
[14] http://data.bnf.fr/what-happened/date-1789.
[15] http://data.bnf.fr/vocabulary/roles/r370.
[16] Esempio http//data.bnf.fr/11928016/jules_verne/rdf.xml.
[17] Da http://data.bnf.fr/semanticweb-en.

può fare meglio, e di lasciare che altri si prendano cura di cose che fanno meglio di te, perché gli utenti preferiscono utilizzare le loro risorse comunque. Per esempio, i cataloghi delle biblioteche sono la descrizione di risorse, e la gestione di concetti che hanno a che fare con i documenti. I punti di forza e di debolezza non corrispondono, come, ad esempio, in un'enciclopedia. Sarebbe inutile cercare di inserire conoscenza universale enciclopedica all'interno dei cataloghi delle biblioteche così come sarebbe inutile fornire elenchi completi dei documenti all'interno di enciclopedie di informazione generale. Le biblioteche stanno portando sul web una prospettiva a lungo termine. Stanno raccogliendo libri da secoli e dati da decenni. La manodopera disponibile non ha eguali in termini di descrizione di libri. Inoltre i dati sono stati strutturati abbastanza presto, con standard internazionali dal 1960. Questi dati descrittivi non sono stati prodotti in una prospettiva di marketing: tutti gli elementi sono accurati, e intendono essere interoperabili, anche se in realtà esistono diversi formati. Le regole per produrre le descrizioni bibliografiche sono rimaste stabili, e sono state rigorosamente seguite da esperti catalogatori. Attraverso diversi formati per i vari tipi di documento, come libri (formato MARC), archivi (EAD), e risorse digitali (DC), gli standard sono stati rispettati. Pertanto le informazioni possono essere attendibili ed elaborate automaticamente, anche su un lungo periodo di tempo, e con grandi quantità di dati. I cataloghi delle biblioteche sono già leggibili dalla macchina, anche se non ancora necessariamente con gli standard web. La loro visualizzazione sul web con gli standard web implica l'uso di identificativi (URIs) in modo che le persone possano citare le risorse. Se vogliamo far utilizzare queste risorse web, occorre fornire link affidabili. Poiché si tratta una grande opportunità per condividere i nostri prodotti culturali, la Bibliothèque nationale de France ha deciso di rendere i dati strutturati disponibili gratuitamente, con una open licence.

L'apertura dei dati della biblioteca sul web è un modo per prendere parte al movimento open data, e dare accesso alle informazioni al grande pubblico, utilizzando le tecnologie più recenti. É anche un incentivo per altri a utilizzare questo materiale e per dare accesso alla cultura. Rendendo gratuiti i dati RDF, questo progetto prende parte anche alle sperimentazioni internazionali di Linked Open Data (LOD) che sono spuntate tra le biblioteche nazionali. Come dice Gildas Illien, «Trasformare preesistenti record MARC e vocabolari di autorità in triple RDF, iniziare a implementare il modello FRBR, rappresentare con gli standard del web semantico, creare applicazioni e set di dati di un nuovo tipo Linked Data-friendly: questo è ciò che guardando ai LOD significa per loro in questa fase («Are you ready to dive in? A case for open data in national libraries»). Poiché le biblioteche stanno lavorando su un prospettiva di lungo periodo data.bnf.fr cerca anche di sperimentare le soluzioni che possono essere utilizzate nei cataloghi della biblioteca originale. In primo luogo, quando sviluppando corrispondenze e algoritmi per la raccolta di dati intorno alle Opere, cerchiamo di fornire informazioni corrette per la visualizzazione dei dati sul web, ad esempio, per evitare duplicati, per evitare la visualizzazione di parole chiave che non corrispondono al contenuto dei documenti, o qualsiasi altra informazione che non sarebbe infatti utile per l'utente finale. Per questo motivo manteniamo tanti link a tutte le risorse nei cataloghi originali, all'interno delle pagine del data.bnf.fr. Cerchiamo anche di costruire procedure e meccanismi che possono essere usati all'interno dei cataloghi originali, a lungo termine, ad esempio per la generazione automatica di pagine di Opere all'interno dei nostri authority file, secondo FRBR. Inoltre, dopo un primo anno di presenza online, possiamo già avere un feedback da alcuni utenti, sul tipo di contenuto che viene utilizzato. Alcuni di essi stanno ridistribuendo il set di dati e fanno riferimento a esso per riutilizzarlo per altri a

partire da data.gouv.fr,[18] il portale ufficiale di dati aperti dello Stato francese, ma anche altri siti come CKAN,[19] OKF[20] e directory di Open data.[21] Altri utenti sono specialisti di dati del settore culturale, che utilizzano una parte dei dati per scopi specifici nelle loro applicazioni locali, come l'*Institut Français*.[22]. Alcuni sono sviluppatori che vogliono costruire linee cronologiche per scopi di ricerca, come ad esempio Yokafun,[23] o per applicazioni Smartphone.[24] Questa vasta gamma di utilizzi di dati non elaborati ci mostra che le informazioni della biblioteca possono essere utili per comunità più ampie, anche se il primo obiettivo rimane descrivere collezioni e dare accesso ad esse. Anche se raccoglie tutte le risorse disponibili a livello di opere intellettuali, il set di dati non è un catalogo nel senso tradizionale del termine, perché non è necessariamente utilizzato per identificare un documento o gestire una collezione. Diventa parte del web, in modo nuovo. Gli authority file e gli identificativi sono più importanti che mai per costruire questo tipo di servizio, ma il set di dati è qualcosa di diverso da un catalogo tradizionale. Inoltre, gli strumenti del web ci permettono di tenere traccia del comportamento degli utenti. Ovviamente possiamo raccogliere il feedback diretto su come le persone hanno reagito a esso, che tipo di contenuto viene utilizzato e ciò che deve essere migliorato; tutto ciò apre un ampio ventaglio di possibili miglioramenti per il futuro.

[18] http://www.data.gouv.fr/donnees/view/Donn%C3%A9es-compl%C3%A8tes-du-contenu-de-la-BNF-30383137.

[19] http://thedatahub.org/dataset/data-bnf-fr.

[20] http://en.wikibooks.org/wiki/Open_Metadata_Handbook/Technical_Overview#Biblioth.C3.A8que_Nationale_de_France_.28BnF.29.

[21] http://open.mflask.com/dataset/data-bnf-fr-bibliotheque-nationale-de-france.

[22] http://ifverso.com.

[23] http://plindenbaum.blogspot.fr/2011/07/drawing-svg-timeline-with-httpdatabnffr.html;https://gist.github.com/1093853.

[24] Per esempio https://sites.google.com/site/catbnf;http://www.appforcash.com/section/item/id/41491.

Works cited

Illien, Gildas. «Are you ready to dive in? A case for open data in national libraries». *Libraries now! Inspiring, surprising, empowering. IFLA World Library and Information Congress. 78th IFLA General Conference and Assembly*. 2012. http://conference.ifla.org/sites/default/files/files/papers/wlic2012/181-illien-en.pdf. (Cit. a p. 11).

Wenz, R. "Linked open data for new library services: the example of data.bnf.fr". *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art. #5509. DOI: 10.4403/jlis.it-5509. Web.