Data.bnf.fr

Presentato a Firenze il portale di rilevanza strategica dell'importante biblioteca francese

Molto partecipato è stato il seminario *Data.bnf.fr*, tenuto il 2 maggio 2017 presso la Sala Comparetti dell'Università di Firenze, organizzato da Mauro Guerrini, coordinatore del Master di II livello in archivistica, biblioteconomia e codicologia, in collaborazione con la Biblioteca nazionale centrale di Firenze, l'Associazione italiana biblioteche - Sezione Toscana e l'Institut français de Florence. All'inizio dell'incontro è stato fornito dalle relatrici ampio materiale didattico, tradotto in italiano nei giorni precedenti dal prof. Graziano Ruffini.

Le due relatrici – Aude Le Moullec-Rieu e Raphaëlle Lapôtre, della Bibliothèque nationale de France (BnF) – hanno presentato il portale data.bnf.fr, online dal 2011 e tuttora in fase di miglioramento, il cui scopo è rendere tutti i dati prodotti dalla BnF e dalla sua Biblioteca digitale Gallica utilizzabili in maniera più ampia sul web. Più precisamente, i principali obiettivi del portale sono:

- aumentare la visibilità dei dati della BnF, per una migliore esposizione sul web;
- federare i dati della BnF, dentro e fuori i cataloghi;
- contribuire alla cooperazione e allo scambio di metadati per la creazione di collegamenti fra risorse strutturate e affidabili;
- agevolare il riutilizzo dei metadati (sotto licenza aperta) da parte di terzi.[1]

Data.bnf.fr si basa su una constatazione: nel mondo di internet, dove proliferano i collegamenti ipertestuali e i link più o meno qualificati, i dati prodotti dalle biblioteche non sono "del web ma solo nel web".[2] Questo perché a oggi i cataloghi non sono costruiti in modo tale che le loro informazioni possano essere reperibili dai motori di ricerca come dati singoli collegati, ma risultano isolati fra loro e dal web nel suo complesso. In altre parole, essi sono strutturati come silos chiusi in loro stessi, con basi di dati non indicizzabili e, in quanto tali, irrecuperabili se non accedendo al sito del catalogo. Al contrario, il progetto data.bnf.fr è in grado di riunire e aggregare nel suo portale il catalogo della BnF e molte altre fonti informative, interne ed esterne alla BnF, altrimenti disconnesse fra loro, poco conosciute e scarsamente adoperate se non da utenti esperti e da professionisti (ad esempio, per le fonti interne: Archives et Manuscrits BnF, Gallica, Reliures per le legature antiche, Medailles et Antiques per le descrizioni di monete e di oggetti museografici ecc.; per le esterne: il catalogo collettivo delle biblioteche di ricerca francesi Sudoc, Wikipedia, VIAF, ecc.).

Data.bnf.fr. dimostra i vantaggi che il passaggio dal web ipertestuale al web semantico comporta in termini d'accesso e di diffusione dell'informazione bibliografica. Sono infatti la tecnologia *linked data* e il web semantico ciò che permetterà ai dati delle biblioteche e delle altre istituzioni della conoscenza di emergere dall'oblio di internet e di acquisire un ruolo di primo piano nell'informazione in rete. Occorre perciò fare evolvere la struttura degli OPAC tradizionali sulla base di criteri e di elementi utilizzati per lo sviluppo del web dei dati. Indicizzando i contenuti dei cataloghi su indici propri del web, e non più solo su thesauri e soggettari esclusivamente biblioteconomici, i cataloghi raggiungeranno così un livello di interoperabilità, di scambio e di condivisione pari a qualsiasi altra fonte di informazione in rete, e l'universo bibliografico diverrà recuperabile direttamente dal motore di ricerca generico.

Dopo l'avvento dell'informatizzazione nelle biblioteche, ci si è da subito posti il problema dello scambio e dell'interoperabilità dei dati, soprattutto con il fine di ottimizzare lo sforzo della catalogazione (derivazione di registrazioni già esistenti, catalogazione partecipata ecc.) e per permettere agli utenti l'interrogazione da remoto dei molteplici cataloghi disponibili. Uno scambio dei dati ottimale impone tre requisiti fondamentali: l'utilizzo dello stesso linguaggio d'interrogazione in tutti i sistemi informativi (nel caso specifico della BnF Z39.50); la medesima sintassi dei dati (MARC); e la presentazione coerente degli elementi MARC (con ISBD e AACR2).

Né va dimenticato l'apporto delle linee guida della famiglia FR (FRBR, FRAD, FRSAD e la consolidata IFLA LRM, in corso di approvazione), le quali mirano anch'esse allo scambio di "cose" diverse (dati parcellizzati, edizione, opera, versione in una data lingua ecc.) dentro e fuori le biblioteche.

Visibilità e interoperabilità dei dati

Le relatrici sono quindi entrate nel merito delle attività di lavoro finalizzate alla realizzazione dei quattro obiettivi di data.bnf.fr, iniziando da quelle destinate ad accrescere la visibilità e l'interoperabilità dei dati aggregati e descritti sul portale. Rendere visibili i dati sul web richiede che essi siano anzitutto validati (cioè verificati da catalogatori professionisti) e contestualizzati (le informazioni devono essere sempre accompagnate da indicazione della fonte e da una data), che gli si possano associare accessi e informazioni perenni e sempre funzionanti, e che ogni informazione sia liberamente riutilizzabile da terzi (a patto che venga sempre citata la fonte). L'ultima condizione, quella della riutilizzabilità, ci porta alla seconda questione. La presenza di dati in formati differenti ha imposto alla BnF una decisione in direzione dell'interoperabilità: il passaggio dal formato MARC a quello XML, essendo quest'ultimo dotato di una sintassi più flessibile, nonché più vantaggioso in termini di granularità dei dati, di strutturazione adattata al contesto, e di facilitazione nella conversione.

Lapôtre e Le Moullec-Rieu proseguono con il terzo obiettivo di data.bnf.fr: creare collegamenti fra le risorse strutturate. Rispetto al web ipertestuale, la peculiarità del web dei dati è rappresentata dalla capacità di costruire e di ordinare le informazioni contenute nelle pagine secondo delle precise regole semantiche. La semantica permette di specificare il significato dei dati, dando a essi un senso compiuto che va oltre la parola scritta, di definirne le relazioni reciproche e di esprimere i vincoli e le gerarchie esistenti fra classi di dati diverse. Grazie alla componente semantica, gli elaboratori elettronici sono quindi in grado di comprendere e di collegare autonomamente le risorse presenti nel web, di tradurre una stringa in un significato e, sulla base di questo, di trovare una quantità indefinita di risorse affini a quella ricercata.
Quella del web semantico è un'architettura gerarchica costituita da tre elementi fondamentali: i dati, i metadati (ovvero tutte le informazioni legate e ricavabili da una risorsa alle quali spetta il compito di far corrispondere i dati con i concetti di un determinato schema) e lo schema, nel quale si esprimono le relazioni fra i concetti che costituiscono le diverse classi di dati. Il W3C propone RDF (Resource Description Framework) come standard per la codifica e la descrizione di una qualsiasi risorsa univocamente identificabile nel web (e non solo). L'elemento di base della grammatica RDF è una tripla, ovvero una sequenza soggetto-predicato-oggetto con la quale si descrivono le relazioni che intercorrono fra un soggetto e un oggetto in modo serializzato; la serializzazione delle triple viene canonicamente eseguita con una sintassi XML. RDF si presenta come uno strumento semplice in grado di rappresentare l'insieme delle conoscenze del web in forma strutturata poiché riconosce come risorsa ogni elemento identificato da URI (Uniform Resource Identifier: ciò che identifica globalmente in modo unico e non ambiguo una risorsa e la rende dereferenziabile e facilmente accessibile).

Identificatori e ontologie

A proposito di URI, Le Moullec-Rieu ha sottolineato come sia di fondamentale importanza scegliere bene gli identificatori in funzione di cosa e di come si voglia individuare una risorsa, così da garantirne l'unicità. Con questo scopo, la BnF ha scelto l'ARK (Archive Resource Key), un identificatore persistente elaborato dalla Chicago Digital Library. La caratteristica principale di un ARK è la sua struttura: una sequenza di caratteri contenente l'etichetta "ark", solitamente preceduta dalla parte iniziale di un URL. L'URL composto con ARK contiene il Name Mapping Authority Hostport (NMAH), mutevole e sostituibile, ovvero l'indirizzo web dell'host atto a risolvere l'identificatore. L'identificatore vero e proprio è però il NAAN (Name Assigning Authority Number), una stringa numerica che identifica in modo univoco e immutabile il nome assegnato all'oggetto, che segue l'etichetta "ark". La specificità di un ARK consiste nella sua capacità di connettere

in modo indissolubile l'oggetto identificato, i suoi metadati e la responsabilità del gestore di garantirne la persistenza.
La questione degli identificatori conduce la riflessione sul piano delle ontologie formali. Predisporre e rendere comprensibile dall'uomo e dalla macchina un quadro che ospiti delle entità soggette a svariati tipi di interazione presuppone infatti la definizione univoca delle entità pertinenti e delle loro possibilità di relazione reciproca. La stessa condizione si dà anche nella progettazione di un qualsiasi database, tanto più se si vuole che i dati da esso contenuti siano scambiabili e interoperabili. Occorre, anzitutto, definire le tipologie di risorsa che ci interessano e i termini da utilizzare per descriverle e per collegarle. Dopo ciò, ogni risorsa sarà concepita come un'istanza di determinate classi (cioè di categorie astratte che organizzano la porzione di realtà in oggetto) e dotata di determinate proprietà (cioè di caratteristiche ritenute significative ai fini della descrizione).

Il portale data.bnf.fr: il valore dell'open

Data.bnf.fr è un portale che nasce su queste basi; fermo restando, però, che delineare un quadro di senso univoco non implica né il riferimento a un modello unico di realtà né l'utilizzo di un unico vocabolario. Posto infatti che in rete non solo a ogni risorsa ma anche a ogni proprietà corrispondono URI individuali; e posto che nel web semantico le informazioni si esprimono sotto forma di triple RDF, data.bnf.fr intende sì presentare le risorse che aggrega con un linguaggio franco rispetto a tutte le fonti interpellate, ma si avvale di più ontologie differenziate. Ciò significa che, di ogni dato che va a recuperare, data.bnf.fr è in grado d'interpretarne i metadati strutturati e formulati seguendo regimi semantici eterogenei, individuati ciascuno da un *namespace* specifico. Sono i *linked open vocabularies* a provvedere alla definizione delle molteplici cornici concettuali e delle varie terminologie disponibili; in un contesto dove l'interoperabilità rappresenta il valore cardine, il requisito *open* è cruciale. FOAF, SKOS, Geonames, Dublin Core, ad esempio, sono solo quattro delle ontologie specializzate e aperte che data.bnf.fr può interpretare in modo integrato – e può farlo proprio perché ciascuna di esse costituisce un modello di metadatazione aperto. Cosa succederebbe, invece, se, rispondendo a una richiesta di un utente, data.bnf.fr andasse ad attingere a una fonte che ha codificato i metadati delle proprie risorse in un formato chiuso? L'informazione rimarrebbe incomprensibile e la *query* non produrrebbe risultati.
Il rapporto d'indipendenza fra le ontologie esistenti, quindi, non ostacola affatto lo scambio dei dati, né esclude che alcune di esse possano essere allineate fra loro sulla base di determinate corrispondenze reciproche. Si tratta del processo di *ontology matching*, il quale non consiste solamente nella scoperta di eventuali relazioni di equivalenza fra *dataset* differenti, ma prevede che si dichiarino le proprietà per descrivere i diversi gradi di equivalenza possibili (ad esempio, *owl:sameAs* per equivalenza stretta, e *rdfs:seeAlso* per relazioni di similarità più generiche) e le proprietà dedicate all'interconnessione di referenziali diversi nell'ontologia SKOS (ad esempio, *skos:exactMatch* per legami d'equivalenza concettuale esatta,

La sede principale della Bibliothèque national de France, nel quartiere di Tolbiac a Parigi. Foto: Manu Dreuil (CC BY-NC-ND 2.0)

e *skos:closeMatch* per equivalenza approssimativa).

Data.bnf. il Giant Global Graph e il riutilizzo dei dati

Lo scenario che data.bnf.fr realizza non ha niente a che vedere con quello dei silos di dati non comunicanti che si è evocato in apertura del seminario. Al contrario, il portale "si integra nel web, proponendo dei collegamenti che permettono di reindirizzare l'utente verso dei link esterni a data.bnf.fr., che si tratti di siti della BnF o no".[3] Il modello entro il quale data.bnf.fr mira a inserirsi è il *Giant Global Graph* definito da Tim Berners-Lee nel 2007, cioè la rete globale dei dati collegati dall'uomo tramite link qualificati e comprensibili in quanto tali dalla macchina. Da ciò, fanno notare le due relatrici, derivano non pochi vantaggi: nessuna conversione né tavola di concordanza necessarie (ognuno utilizza il proprio formato); ridondanza limitata (ognuno crea i dati di cui ha bisogno e recupera le informazioni già esistenti); tracciabilità e apertura (ognuno può ritornare ai dati *source* grazie agli URI); e standard aperti (nessuno deve imparare linguaggi o protocolli diversi da RDF e SPARQL).

Un grande punto di forza della codifica in formato *open* dei dati e delle loro relazioni reciproche (in XML e RDF) è, come già detto, l'interoperabilità. Ogni informazione strutturata in ambiente *linked open data* è pertanto riutilizzabile da chiunque – e siamo giunti così alle modalità di realizzazione del quarto e ultimo obiettivo di data.bnf.fr. Oltre alla condizione dell'apertura tecnica, il riutilizzo dei dati impone infatti anche la loro apertura giuridica; ovvero, i dati devono essere sotto licenza aperta.

Data.bnf.fr soddisfa entrambe le condizioni. Della prima si è già parlato; a proposito della seconda Le Moullec-Rieu e Lapôtre ricostruiscono le tappe legislative che hanno portato la BnF a promuovere l'apertura giuridica dei suoi dati: a partire dalla l. **n.** 78/753 del 1978 sulla libertà d'accesso ai documenti amministrativi, all'allineamento con la Direttiva 2003/98/CE dell'UE (modificata nel 2013) sul riutilizzo dei dati del settore pubblico (resa gratuita con la l. n. 2015/1779), fino alla l. n. 2016/1321 "pour une République numérique", che sancisce l'accesso aperto, di default e non su richiesta, per ogni dato pubblico conservato.

Merita qui una menzione particolare la licenza elaborata dalla missione Etalab, "incaricata della creazione di un portale unico interministeriale dei dati pubblici" con il d. 2011/194 del Primo Ministro e dal 2012 alle dipendenze del Sécretariat général pour la modernisation de l'action publique. Si tratta della *Licence Ouverte*, aggiornata lo scorso aprile, la quale promuove la libera riutilizzazione di qualsiasi dato pubblico a patto di citarne la fonte; dal 1- gennaio 2014 essa è applicata a tutti i metadati descrittivi (dati bibliografici e d'autorità) della BnF.

Chiude il seminario e completa il quadro sulle iniziative pro *Open Data* in Francia, la presentazione di due progetti elaborati in seno al primo Hackathon della BnF, tenuto il 19 e il 20 novembre 2016 nel corso della settimana dell'innovazione pubblica: Gallicarte, il progetto vincitore, che tramite la geolocalizzazione dei dati associati ai documenti conservati dalla BnF mira a una navigazione cartografica su Gallica; e Monallica, con lo scopo di estrarre da Gallica gli indici e i testi acquisiti tramite OCR e di procedere al riconoscimento delle entità nominate, così da ritrovare i luoghi e le persone citati nei testi digitalizzati per allinearli con i dati di data.bnf.fr.

DESIRÈE MARIE KOEHRING
Università di Firenze
desikoe88@gmail.com

VALENTINA LEPORE
Università di Firenze
vtlepore@gmail.com

GIADA STIGLIANO
Università di Firenze
giadastigliano@gmail.com

NOTE

[1] http://data.bnf.fr/fr/about. Ultima consultazione 13 giugno 2017.

[2] Per la tradizione di ricerche che la formula porta con sé, si veda Mauro Guerrini - Tiziana Possemato, *Linked data per biblioteche, archivi e musei. Perché l'informazione sia del web e non solo nel web*, con un saggio di Carlo Bianchini e la consulenza di Rosa Maiello e Valdo Pasqui, prefazione di Roberto Delle Donne, Milano, Editrice Bibliografica, 2015.

[3] Vedi nota 1.

DOI: 10.3302/0392-8586-201706-054-1