Web dei dati alla Biblioteca nazionale centrale di Firenze[1]

Anna Lucarelli
Biblioteca nazionale centrale di Firenze

Le Biblioteche nazionali di tutto il mondo stanno lavorando per favorire l'accesso alle proprie risorse da parte degli utenti del Web, al di là dei confini ristretti dei cataloghi. L'obiettivo è mettere a disposizione non solo opere ed edizioni, ma anche dati di qualità strutturati e attribuiti in fase di catalogazione o di digitalizzazione, rendendoli riusabili e disponibili nei formati del Web semantico. I linked open data sono una nuova e straordinaria opportunità e potranno dare un significato nuovo al ruolo sociale e pubblico che le Nazionali rivestono. La Biblioteca nazionale centrale di Firenze ha inaugurato le sue esperienze di open data con il Thesaurus del Nuovo soggettario, strumento evoluto nella direzione del multilinguismo e interoperabile con data set di altre amministrazioni pubbliche e istituzioni della memoria.

Open data: esperienze di biblioteche nazionali

In ogni luogo del mondo le biblioteche nazionali sono istituzioni che conservano, tutelano e diffondono la memoria culturale e il patrimonio documentario del proprio Paese. Hanno il compito non solo di ospitare la produzione editoriale nazionale, ma anche di coordinarne l'informazione, giocando un ruolo importante anche a livello internazionale con il mantenimento di bibliografie nazionali, di authority files ecc.

Le loro funzioni si sono da un lato sempre più precisate, dall'altro si sono allargate per soddisfare, con prodotti nuovi, un'utenza più ampia e diversa rispetto a quella tradizionale, l'utenza del Web. E sappiamo che gli utenti del Web hanno nuove aspettative e nuove abitudini.

[1] Relazione presentata alla Tavola Rotonda *Biblioteche digitali verso il futuro: accesso ai linked open data-Book to the future* (coordinata da Anna Maria Tammaro).

Le biblioteche nazionali, grazie alla normativa sul deposito legale, sono a contatto con l'intero panorama editoriale di un Paese - libri, periodici, opere grafiche e altre risorse documentarie - e dunque si muovono in un contesto molto vicino a quello rappresentato dal Salone del libro di Torino del 2015, nell'ambito del quale si è svolta la Tavola rotonda su "Biblioteche digitali verso il futuro".
Sappiamo che l'universo digitale ha coinvolto le biblioteche per gradi. In principio, i formati elettronici hanno riguardato le schede dei cataloghi, le liste di autorità, alcuni strumenti; poi, nello spazio virtuale e fisico delle biblioteche, hanno iniziato a evolvere funzioni di sviluppo di alcuni servizi, si sono attivate procedure legate all'acquisizione e conservazione di documenti digitalizzati o nati in formato digitale, iniziative che hanno avuto una diffusione massiccia proprio negli ultimi anni anche grazie alle collaborazioni avviate con progetti ben noti come Google Books al quale, in Italia, hanno aderito le Biblioteche nazionali centrali di Roma (coordinatrice del progetto) e di Firenze, la Biblioteca nazionale Vittorio Emanuele III di Napoli e recentemente altre biblioteche[2].
A livello internazionale, ci si è chiesto come fornire maggior accesso e visibilità ai dati prodotti dalle biblioteche, al di là dei confini ristretti dei cataloghi. Il passo avanti più significativo è stato quello di comprendere che ciò che le biblioteche possono mettere a disposizione non sono soltanto le opere e i testi che le veicolano, ma anche i dati che le riguardano, recependo così appieno il passaggio di prospettiva, per certi aspetti rivoluzionario, del Web semantico, il "Web dei dati", come lo ha appunto notoriamente definito Tim Berners-Lee.
La diffusione dei metadati, intesi come dati strutturati che descrivono altri dati e consentono ai computer, tramite particolari applicazioni, di riconoscere relazioni tra entità, favorendo il reperimento di informazioni, si inserisce appieno nella logica degli open data, dati aperti collegati fra loro (linked open data), grazie a licenze che ne garantiscono accessibilità e riuso[3]. E d'altra parte, la consapevolezza che i linked data siano una nuova opportunità per istituzioni e utenti è sempre più condivisa ed è alla base dell'allestimento di aggregatori di risorse digitali e di ben noti progetti internazionali come Europeana[4].
Così le biblioteche nazionali di tutto il mondo si stanno facendo, in varia misura,

[2] Il sito della BNCR dedicato a Google Books è: <http://www.bncrm.librari.beniculturali.it/index.php?it/832/progetto-googlebooks>; notizie più recenti sul progetto si trovano anche nell'intervista rilasciata, a fine settembre 2015, da Andrea De Pasquale, direttore della BNCR:<http://fontegaia.hypotheses.org/1230>.

[3] Sul tema degli open data esiste una ricchissima bibliografia. Si segnala: Mirna Willer – Gordon Dunsire, *Bibliographic information organization in the semantic Web*, Oxford [etc.]: Chandos Publishing, 2013; fra i contributi italiani più recenti: 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: Bibliografica, 2015.

[4] http://www.europeana.eu/portal/.

sempre più promotrici di open data. Si sta iniziando anche in Italia a lavorare affinché i metadati da esse prodotti siano più visibili, più facili da ricercare, linkabili e riusabili direttamente. Se da un lato le biblioteche non sono più gli unici produttori di dati bibliografici (come dimostrano le esperienze che, ad esempio, nell'ambito di Wikidata, si stanno sperimentando grazie a Wikibase)[5], d'altro lato l'enorme massa di metadati, generalmente e tradizionalmente di qualità, che esse producono, se convertiti e messi a disposizione nei linguaggi del Web semantico, se resi interoperabili e aperti, potranno avere un'applicabilità ampia e diventare utili a un numero di cittadini ben superiore a quello attuale. Questo scenario offre alle biblioteche l'opportunità di svolgere nuove modalità di servizio pubblico (ad esempio, condividendo dati con altri *data set* della pubblica amministrazione) e di dare significato agli investimenti che impiegano nella catalogazione e nella digitalizzazione, evitando costi di duplicazione e generando forme di risparmio. D'altra parte, sul valore degli open data come servizio per i cittadini, ci sono ormai esperienze importanti come, ad esempio, quelle realizzate dalla Regione Piemonte, dalla Regione Emilia Romagna o dal Comune di Firenze.

Dunque le attività catalografiche svolte da sempre nell'ambito delle biblioteche sono tutt'altro che divenute inutili, ma gli investimenti impiegati avranno maggior valore se i dati relativi alle risorse descritte, oltre che affidabili, si baseranno su modelli concettuali condivisibili (come FRBR), adotteranno standard del *semantic Web*, ad esempio, rendendosi disponibili nelle triple del linguaggio RDF (*Resource Description Framework*)[6].

Come già accennato, sono numerose le esperienze e le sperimentazioni di *linked data service* avviate da biblioteche nazionali di tutto il mondo: fra i casi prototipali ed esemplari si citano solitamente quelli di Svezia, Germania, Francia, Regno Unito, USA ecc.[7]. E vanno senz'altro guardate con grande interesse le recenti iniziative dell'Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche (ICCU) che ha costituito un gruppo di lavoro per valutare la possibilità di strutturare set di dati prodotti nell'ambito di SBN come linked open data[8].

[5] Sull'argomento, Giovanni Bergamin, *La sfida di Wikidata alle biblioteche*, Convegno nazionale *Sfide e alleanze fra biblioteche e Wikipedia*, Biblioteca nazionale centrale di Firenze, 28 novembre 2014, <http://www.slideshare.net/GiovanniBergamin/la-sfida-di-wikidata-alla-biblioteche>.

[6] Il tema dell'apporto che le biblioteche, specialmente Nazionali, possono dare a livello sociale tramite i propri dati aperti è stato già affrontato da bibliotecari ed esperti di vari paesi. Un riferimento a tali contributi si trova anche in Anna Lucarelli, *Nuove scommesse della BNCF: Wikipediani in residence, Wikisource e altro ancora,* «Dig*Italia*», 9 (2014), n. 2, p. 100-106. <http://digitalia.sbn.it/article/view/1292/849>. Sul valore sociale dei linked data, anche Francesca Di Donato, *Lo Stato trasparente. Linked open data e cittadinanza attiva*, Pisa : ETS, 2010.

[7] Interessante, per quanto riguarda la BNF, la rubrica aggiornata *Web de données, Web sémantique*, <http://www.bnf.fr/fr/professionnels/innov_num_web_donnees.html>.

[8] < http://www.iccu.sbn.it/opencms/opencms/it/main/attivita/gruppilav_commissioni/pagina_0005.html>. Sul tema, si segnala anche l'intervento di Patrizia Martini nella lista AIB-CUR in data 30 settembre 2015.

La Biblioteca nazionale centrale di Firenze (BNCF) ha inaugurato le sue esperienze di open data grazie a uno strumento che allestisce per l'intero sistema delle biblioteche italiane: il Thesaurus del *Nuovo soggettario*.

Nuovo soggettario e open data: da strumento catalografico a fornitore di dati aperti

Come la Library of Congress, la Bibliothèque nationale de France e molte altre grandi biblioteche, anche la BNCF cura il principale sistema nazionale impiegato nell'indicizzazione per soggetto, la cui componente centrale è un Thesaurus multidisciplinare online di oltre 56.000 termini usabili nella catalogazione di risorse varie e dunque in un'ottica fortemente orientata alla condivisione di linguaggi con musei, archivi, fototeche ecc.[9].

Lo strumento è disponibile online dal 2007 e dal 2013 con un'interfaccia anche in inglese[10].

È interessante raccontare di come uno strumento nato come un "attrezzo" del mestiere dei bibliotecari, dunque con fini esclusivamente catalografici e inserito pienamente in una prassi bibliotecaria di stampo tradizionale, possa in realtà evolvere con caratteristiche impreviste fino a qualche anno fa e con funzionalità che stimolano il collegamento non solo fra biblioteche e altre istituzioni della memoria, ma anche con altre amministrazioni pubbliche che stanno sviluppando progetti analoghi nella stessa direzione.

Il Thesaurus è integrato con il catalogo della BNCF ed è integrabile con qualsiasi Opac delle altre biblioteche che lo adottano, consentendo in questo modo navigabilità e collegamenti diretti con le notizie bibliografiche. Si alimenta accogliendo terminologia usabile nell'indicizzazione e proposta dalla rete di biblioteche, istituzioni e centri di ricerca che partecipano al progetto.

Rappresenta dunque un esempio di come si possa mettere a disposizione di biblioteche digitali, anche specialistiche, *tools* prodotti da una biblioteca nazionale e di come strumenti di accesso all'informazione si possano sviluppare sfruttando competenze e risorse professionali di varia natura, grazie alla cooperazione fra istituzioni diverse, generaliste e specialistiche, pubbliche e private[11].

Come è noto, i termini del Thesaurus sono organizzati secondo le categorie e le relazioni previste dagli standard internazionali (è in corso l'adeguamento all'ultimo standard ISO 25964 che ha sostituito i precedenti 2788 e 5964), sono corredati da

[9] Della possibile integrazione di metadati prodotti in contesti documentari diversi si sta occupando un gruppo di lavoro formatosi nell'ambito del coordinamento del MAB Toscana, < http://www.mab-italia.org/index.php/comitatati/mab-toscana>. Il lavoro del Gruppo è confluito nel report: *Verso l'integrazione tra archivi, biblioteche e musei. Alcune riflessioni*, a cura del MAB Toscana – Gruppo Linguaggi (in via di pubblicazione).

[10] Per la descrizione del sistema di indicizzazione: <http://thes.bncf.firenze.sbn.it/index.html>; per l'accesso al Thesaurus: <http://thes.bncf.firenze.sbn.it/ricerca.php>.

[11] Per l'elenco dei partner del progetto: <http://thes.bncf.firenze.sbn.it/enti.htm>.

note, da collegamenti con sinonimi e forme non più preferite (varianti storiche), dall'indicazione di corrispettivi numeri della *Classificazione decimale Dewey*, nella versione aggiornata della *WebDewey italiana*[12], e inoltre dalla segnalazione dei repertori, enciclopedie, database, usati come fonti per il controllo di morfologie e significati (Fig. 1).

Figura 1: *Esempio di un termine del Thesaurus del Nuovo soggettario*

Dati aperti e interoperabilità del *Nuovo soggettario*

Come già accennato, il Thesaurus negli ultimi anni si è sempre più evoluto nella direzione, oltre che del multilinguismo, anche dell'interoperabilità con altri strumenti e sistemi di organizzazione della conoscenza. Colloquia, attraverso link, con gli equivalenti in inglese di *Library of Congress Subject Headings* – LCSH (oltre 10.000 collegamenti reciproci), con migliaia di equivalenti in francese previsti dal sistema *Rameau* della Bibliothèque nationale de France (quasi 8.000), con circa 13.000 lemmi della versione in italiano di *Wikipedia*, con thesauri online (AAT, *Agrovoc*, *Eurovoc*, *Polimoda*) e con altre risorse digitali come l'*Enciclopedia Italiana* Treccani, sistemi di classificazione (BIAnet, DoGI). I dati numerici indicati, aggiornati al mese di dicembre 2015, sono tutti in costante accrescimento. Nel corso del 2016 sarà implementata anche la navigabilità con *WebDewey italiana*, le cui notazioni compaiono già nei record di circa 5.000 termini del Thesaurus (Fig. 2).

[12] http://www.aib.it/pubblicazioni/webdewey-italiana/.

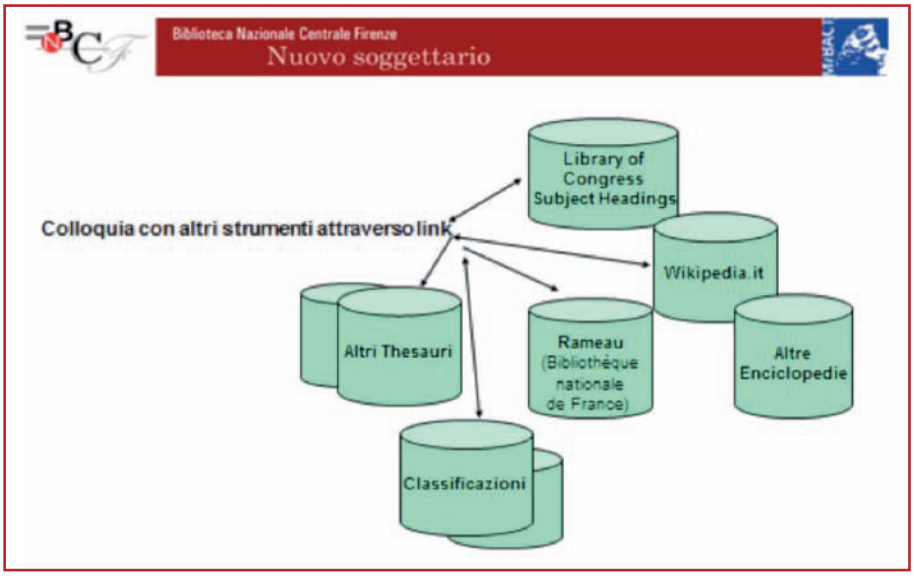

Figura 2: *I collegamenti del Nuovo soggettario con altri strumenti e data set*

La maggior parte dei collegamenti attivati dal *Nuovo soggettario* sono possibili perché i metadati del Thesaurus dal 2010 sono rilasciati nel particolare formato SKOS, sviluppato dal W3C (Semantic Web Deployment Working Group - SWDWG) e basato sul linguaggio RDF[13]. Come è noto, si tratta di uno standard molto efficace per far colloquiare sul Web i thesauri fra di loro o con altre strutture concettuali, come tassonomie o classificazioni. SKOS, infatti, può semplificare la comunicazione e collegare mondi che altrimenti rimarrebbero separati, grazie al fatto che la sua struttura si basa su concetti, indipendentemente dalla loro espressione linguistico-letterale e che tali concetti e le relazioni fra loro sono espressi con triple RDF. Il concetto viene associato (tramite il predicato *rdf:type)* alla classe specifica *skos:Concept* ed è identificato in modo univoco con un *Uniform Resource Identifier*. Così, l'assegnazione di un valore univoco ai concetti facilita l'interoperabilità tra *Knowledge Organization Systems* (KOS) differenti, cioè la possibilità di mappare entità semantiche di schemi concettuali diversi[14].
La conversione in SKOS del *Nuovo soggettario* ha seguito varie fasi, non è stata priva di problemi, alcuni tuttora aperti, ma il lavoro è comunque evoluto e sta facendo tesoro degli ultimi allineamenti stabiliti fra SKOS e il già citato nuovo standard ISO 25964[15].

[13] http://www.w3.org/2004/02/skos/.
[14] Sull'argomento già Giovanni Bergamin – Anna Lucarelli, *The Nuovo soggettario as a service for the linked data world*, «JLIS.it», 4 (2013), n. 1, <http://leo.cineca.it/index.php/jlis/article/view/5474/7903>.
[15] Sulle corrispondenze fra i due standard: <http://www.niso.org/apps/group_public/download.php/12351/Correspondence%20ISO25964-SKOSXL-MADS-2013-12-11.pdf>.

Il Thesaurus del *Nuovo soggettario*, insomma, punta a soddisfare quelli che vengono usualmente definiti come i requisiti di base dei dati aperti. È indicizzabile dai motori di ricerca, è disponibile in un formato aperto e standardizzato che facilita la sua consultazione, è rilasciato attraverso licenze libere che ne incentivano la diffusione e il riuso da parte di soggetti interessati[16].
Oltre ai fronti già citati, se ne dischiudono altri come, ad esempio, quello dell'impiego del Thesaurus come strumento di indicizzazione automatica di risorse digitali, versante su cui si sono già avviate sperimentazioni[17].

Implementazione dei linked data nel *Nuovo soggettario*

Le corrispondenze del *Nuovo soggettario* possono essere attivate mediante l'indicazione di altri strumenti nel campo "Fonte" previsto nel record di ogni termine. Se lo strumento citato è disponibile in SKOS, le relazioni si arricchiscono tramite *skos:closeMatch*, se invece non è disponibile in questo formato, l'informazione si perde, pur restando citata nel campo "Fonte" la sigla relativa allo strumento linkabile. Il collegamento può avvenire anche tramite l'indicazione di equivalenza esplicitata in altri campi, come avviene per i corrispettivi previsti da LCSH, con l'attivazione di *deep link* allo strumento citato.
È interessante rilevare che la collaborazione con la maggior parte delle istituzioni che producono dati collegati è precedente alla pubblicazione dei dati in formato SKOS, a riprova del fatto che il piano delle *policies* resta comunque fondamentale anche nel mondo degli open data.
E comunque non tutte queste equivalenze sono attive nell'implementazione SKOS, poiché non tutti gli strumenti collegati pubblicano dati in RDF. La BNCF scarica costantemente dati dagli strumenti collegati e disponibili nello SKOS da essi aggiornato; alcuni degli enti che li gestiscono, periodicamente, fanno l'inverso e ciò fa sì che si stabiliscano link reciproci, spesso esplicitati, come nel caso di LCSH in cui i link ai termini italiani del *Nuovo soggettario* sono segnalati anche con l'icona del nostro tricolore[18].
L'interoperablità con *Wikipedia* è evidentemente una delle più significative e si è sviluppata nell'ambito di un progetto di collaborazione con il capitolo italiano di

[16] Il riferimento è alla licenza Creative Commons: <http://thes.bncf.firenze.sbn.it/thes-dati.htm>.
[17] La sperimentazione è stata avviata (con la collaborazione di @CULT) tramite la creazione di un modello di apprendimento multidisciplinare a partire dalle tesi di dottorato di ricerca in versione elettronica depositate in BNCF. I risultati ottenuti sono tuttora in fase di analisi; margini di miglioramento potranno essere legati al raffinamento delle procedure di attribuzione di metadati sui documenti che compongono questo modello di apprendimento e dall'uso di algoritmi più sofisticati rispetto a quelli basati unicamente su leggi statistiche e frequenza delle parole.
[18] Sui collegamenti reciproci fra *Nuovo soggettario* e LCSH: Anna Lucarelli – Elisabetta Viti, *Florence–Washington Round Trip: Ways and Intersections between Semantic Indexing Tools in Different Languages*, «Cataloging & Classification Quarterly», 53 (2015), n. 3-4, p. 414-429 (pubblicato online il 19 Marzo 2015, <http://www.tandfonline.com/toc/wccq20/53/3-4>).

Wikimedia Foundation. Anche grazie all'impiego sperimentale di automatismi, i link con *Wikipedia* sono notevolmente aumentati; questa interoperabilità ha accresciuto le visite del Thesaurus e gli accessi all'Opac della BNCF. La biblioteca non ha imposto alcun obbligo di registrazione a chi scarica dati tramite il formato SKOS/RDF del Thesaurus, al fine di renderlo usabile il più facilmente possibile, una politica che si è confermata con la possibilità recentemente offerta di interrogare i dati tramite un endpoint SPARQL su Datahub che utilizza CKAN come piattaforma pubblica[19] (Fig. 3).

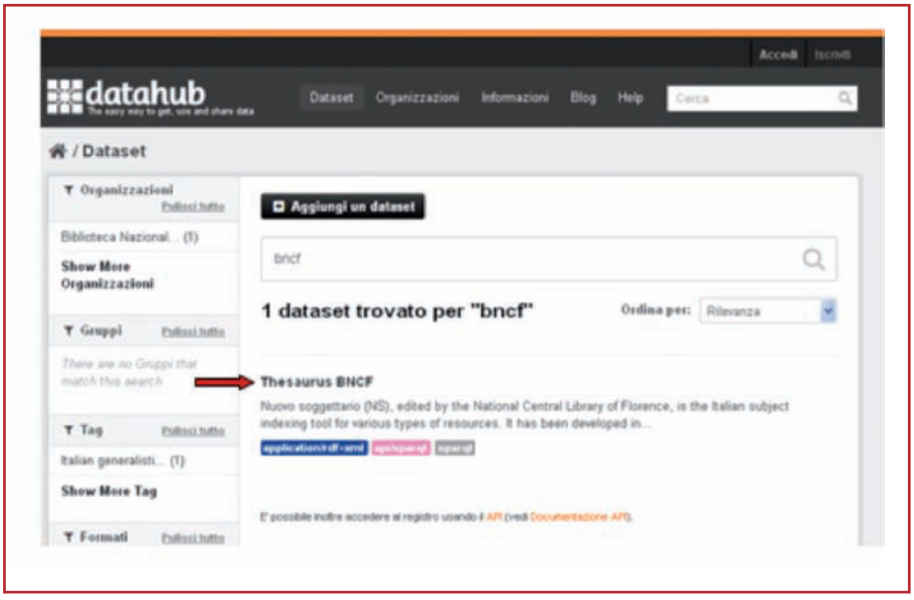

Figura 3: *Il data set del Thesaurus tramite la pagina di Datahub*

Pubblicazione dei dati e trasparenza informativa

I dati aperti del *Nuovo soggettario*, dunque, attivano interoperabilità e favoriscono forme varie di riuso. È interessante analizzare gli accessi al server, sapere quanti e quali utenti utilizzano i dati in formato SKOS, anche mese per mese, anno per anno. Si scopre, ad esempio, che viene scaricato da biblioteche di tutti tipi (dalla Nazionale Marciana, alla Library of Congress, dalla Biblioteca del Congresso del Cile al Sistema bibliotecario del Canton Ticino), da università italiane e straniere, da archivi di Stato (ad esempio, quello di Milano), da soprintendenze archivistiche, da musei (ad esempio, dal Museo Galileo di Firenze, oppure dai Musées royaux d'art et d'histoire di Bruxelles), da commissioni dell'Unione europea, da enti pubblici di vario genere (il Ministero dell'interno, la Regione Emilia Romagna, la

[19] Knowledge Foundation, è un sistema molto usato da amministrazioni pubbliche (sia straniere che italiane) per la pubblicazione dei propri open data; per il suo sito ufficiale:< http://ckan.org/>.

Provincia di Ravenna, il Comune di Firenze ecc.), da enti, fondazioni, aziende (come CNR, Fondazione Bruno Kessler, Cineca, CSI Piemonte, Cedoc, Data-Management[20]), da aziende sanitarie locali. Il riutilizzo dei dati è spesso testimoniato da esplicite clausole di attribuzione. Ma per la BNCF è interessante analizzare anche le applicazioni che ne vengono fatte, avere riscontri concreti e avviare canali informativi e comunicativi, poiché anche nell'epoca degli open data le biblioteche devono continuare a lavorare, oltre che sul piano tecnico, anche su quello delle politiche. Le esperienze realizzate dovrebbero comunque essere maggiormente pubblicizzate. È sempre interessante entrare in contatto con la riutilizzazione fuori contesto di dati aperti prodotti da biblioteche. Quando si scopre che possono trovare una loro utilità in altre amministrazioni pubbliche e in ambienti anche molto diversi da quelli di origine, se ne resta sorprendentemente colpiti, come quando qualche anno fa su «Che futuro!», periodico online dedicato all'innovazione, Giovanni Menduni, allora dirigente del Comune di Firenze e coordinatore dei progetti di open data di quell'amministrazione, riferendosi all'integrazione dei dati del *Nuovo soggettario* con altri *data set* del Comune, raccontava *Come abbiamo ordinato gli open data di Firenze con l'aiuto di una biblioteca* (Fig. 4)[21].

Figura 4: *Articolo sull'esperienza di riuso degli open data del Nuovo soggettario da parte del Comune di Firenze*

[20] L'uso dei dati del *Nuovo soggettario* da parte di Data Management, nell'ambito di Sebina, è stato dichiarato in modo formale e ha dato luogo ad uno scambio con la BNCF, utile per alcuni test e verifiche.

[21] http://www.chefuturo.it/2012/07/come-abbiamo-ordinato-gli-open-data-di-firenze-con-laiuto-di-una-biblioteca/.

National Libraries around the world are working to make their resources available on the Web, beyond the frontiers of their OPACs. The aim is to offer not only works and editions online but also structured data quality, assigned during the cataloging and digitalizing stages, so that they can be interlinked and made even more useful via the semantic Web. The Linked Open Data are a new and extraordinary opportunity, and give a new relevance to the social and public roles of the National Libraries. The National Central Library of Florence has just begun its Open Data experience with the Nuovo Soggettario Thesaurus, a tool developed towards multilingualism and interoperable with the data of other public administrations and memory institutions.

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