RIFERIMENTI BIBLIOGRAFICI

[1] Melissa Gymrek - Yossi Farjoun, *Recommendations for open data sciencd*, "Gigascience" 2016; 5(1): 1-3.

[2] Patty Kostkova et al., *Who owns the data? Open data for healthcare*, "Front Public Health" 2016; 4:7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4756607

[3] *Transparency and Open Government*, 1/21/2009, President Barack Obama. https://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment

[4] *Il piano d'azione europeo per l'eGovernment 2011-2015. Valorizzare le TIC per promuovere un'amministrazione digitale intelligente, sostenibile e innovativa,* Commissione Europea, 2010, Bruxelles, 15/12/2010. Com (2010) 743 definitivo http://eur-lex.europa.eu/legal-content/it/txt/?uri=celx:52010dc0743

[5] OpenAIRE. https://www.openaire.eu

[6] John Houghton, *Open Access-What are the economic benefits? A comparison of the United Kingdom, Netherlands and Denmark*, June 22, 2009. https://ssrn.com/abstract=1492578 or http://dx.doi.org/10.2139/ssrn.1492578

[7] *Verso un accesso migliore alle informazioni scientifiche: aumentare i benefici dell'investimento pubblico nella ricerca*, Commissione Europea, Bruxelles, 17.7.2012 COM (2012) 401 final. https://www.researchitaly.it/uploads/7309/com_401.pdf?-v=a901bf7

[8] Open data. https://ec.europa.eu/digital-single-market/en/open-data

[9] Portale Open Data dell'Unione Europea. https://data.europa.eu/euodp/it/about

[10] Horizon 2020. https://ec.europa.eu/programmes/horizon2020/

[11] *Sharing Clinical Trial Data. Maximizing Benefits, Minimizing Risk*, Institute of Medicine of the National Academie, 2015. https://www.nap.edu/read/18998/chapter/1

[12] Naana Afua Jumah et al., *Has open data arrived at the British Medical Journal (BMJ)? An observational study*, BMJ Open 2016, 6(10). http://bmjopen.bmj.com/content/6/10/e011774

[13] *European Medicines Agency agrees policy on publication of clinical trial data with more user-friendly amendments*, Press Release 12/06/2014. http://www.ema.europa.eu

[14] European Medicines Agency Clinical Data. https://clinicaldata.ema.europa.eu/web/cdp/home

[15] Frank P. Rockhold, *Data Access and Sharing: Are we being transparent about clinical research? Let's do what's right for patients*, Ann Oncol 2017 Apr 5. doi: 10.1093/annonc/mdx123 [Epub ahead of print]

[16] Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca. http://www.anvur.org

[17] Force M.M., Robinson N.J.J. Encouraging data citation and discovery with the Data Citation Index. Comput Aided Mol Des 2014;28:1043

[18] Simone Aliprandi, *Fare open access. La libera diffusione del sapere scientifico nell'era digitale*, Milano, Ledizioni, 2017.

[19] Ferruccio Diozzi - Silvia Molinari - Francesca Gualtieri - Ivana Truccolo, *Cinque tesi sui social network*, "Biblioteche oggi", 32 (2014), n.4, p. 5-9.

[20] Reviewer Credits https://reviewercredits.com

[21] *Utopia scientifica: apertura dell'inchiesta psicologica di comunicazione scientifica*, 2012. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1492578

Open data e open source per una biblioteca digitale aperta

Il tema del convegno delle Stelline di quest'anno era "La biblioteca aperta", e ci è sembrato particolarmente indicato per raccontare a tutti la nostra versione di biblioteca aperta. O meglio: *come stiamo costruendo* la sezione di Risorse Open nei progetti MLOL, OpenMLOL, cioè una biblioteca digitale che sia aperta, libera, partecipativa.

Al momento MLOL (nelle sue varie istanze: i portali bibliotecari e anche i portali MLOL Scuola) possiede infatti due principali collezioni:

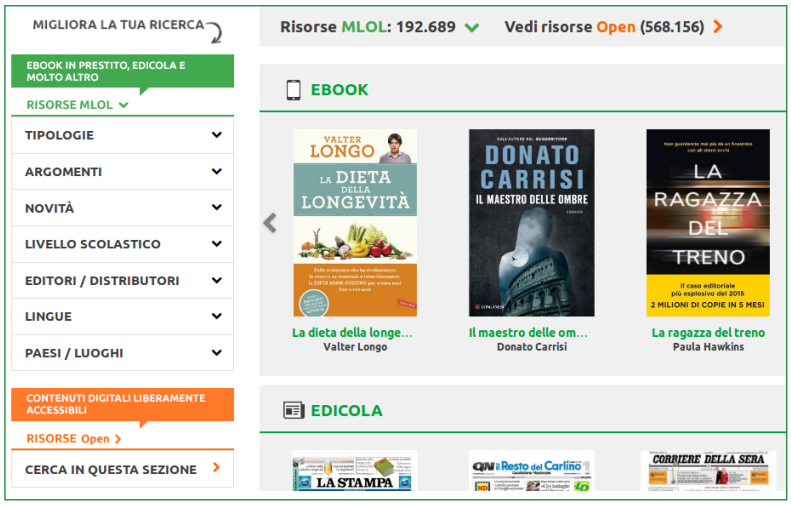
- una collezione composta di risorse ancora *sotto copyright* (es. Edicola e ebooks ancora in catalogo);
- una collezione di risorse *open*, gratuite e con licenze aperte Creative Commons. Questa collezione è anche accessibile in un portale autonomo, chiamato openMLOL.

Biblioteca come piattaforma

Il termine "biblioteca digitale" è sempre stato vago e ambiguo, anche fra gli addetti ai lavori. Vengono definite biblioteche digitali progetti diversissimi: Google Books, Europeana, archivi istituzionali, Internet Archive, siti bibliotecari non più aggiornati con poche decine di scansioni in JPG, la Digital Public Library of America, progetti di trascrizione gestiti da volontari su Internet.

In un articolo[1] del 2012 su Library Journal, David Weinberger proponeva un cambio di prospettiva con l'idea della "library as a platform", cioè una *biblioteca come piattaforma*.

L'idea cioè di una biblioteca digitale focalizzata sui dati, sul loro uso ma soprattutto *riuso*. Una biblioteca che lavora i dati e li restituisce sempre in ma-



HomePage MLOL, la suddivisione fra risorse commerciali ed Open

niera *open*, attraverso API aperte, e che apra ad altri l'innovazione e la creazione di servizi che noi stessi non abbiamo pensato o non sappiamo realizzare.
Un modello può certamente essere la Digital Public Library of America: portale di aggregazione di collezioni locali di decine di istituzioni negli Stati Uniti, con una complessa e ricca possibilità di interfaccia per dati e contenuti, per cui sono a disposizione app sulle collezioni fatte da sviluppatori indipendenti.
Un altro modello, ancora più estremo, può essere Internet Archive: una gigantesca libreria con milioni di risorse digitali, che vanno dalle digitalizzazioni del libro antico ai videogame degli anni Ottanta e Novanta, leggibile da umani ma soprattutto da *macchine*.
Internet Archive infatti punta moltissimo al riutilizzo dei propri dati: mantiene una struttura delle proprie URL altamente logica e modulare, facendo sì che le proprie API siano facili da capire e da utilizzare anche per programmatori non professionisti.
La biblioteca digitale del futuro, come quella del presente, deve assolutamente tenere in conto il *riuso* e la *riaggregazione* dei propri dati, secondo i principi degli open data e dell'open source.
Non è un caso che istituzioni culturali come la New York Public Library, o musei importanti come il MET e il MOMA, abbiano da tempo abbracciato la filosofia open source e rilascino liberamente online le proprie collezioni libere da diritto d'autore, e anche i metadati attorno ad esse.
Dati e metadati sono disponibili su GitHub, cioè la più grande collezione di codice libero al mondo, e il luogo giusto dove offrire agli sviluppatori di tutto il mondo le proprie collezioni di immagini, digitalizzazioni e metadati.
Vediamo, in questo senso, alcune innovazioni:
- l'utilizzo, per le proprie collezioni di dati bibliografici, di licenze estremamente aperte come la CC0, cioè di fatto una completa *liberalizzazione* del dato, secondo la filosofia *open source* e *open data*. Il dataset della NYPL diventa così un bene comune digitale, e può potenzialmente andare ad arricchire progetti come Wikidata;
- l'utilizzo di API aperte e di "standard web" (es. API REST, JSON), in modo da rendere più accessibili questi dati al di fuori del mondo bibliotecario;
- l'utilizzo di piattaforme come GitHub, estremamente popolari e standard *de facto* della comunità di sviluppatori open source.

L'interoperabilità di una biblioteca digitale può quindi essere definita sia a livello *legale* che a livello *informatico-tecnologico*.
Non basta certamente infatti digitalizzare il proprio patrimonio bibliografico per far sì che esso sia davvero accessibile.
L'apertura legale ed informatica è necessaria perché i contenuti possano essere condivisi, trasmessi, usati e riusati anche in altri contesti. Significa mettere a disposizione di altri aggregatori (es. Europeana a livello istituzionale, ma anche Internet Archive in maniera più informale e "comunitaria") le proprie collezioni, in modo che la grande visibilità di quegli aggregatori possa rendere i nostri contenuti davvero aperti e accessibili.
L'apertura di un contenuto permette infine non solo la possibilità di riutilizzo da parte di altri, ma è anche, paradossalmente, il modo migliore per *sapere* dove sta andando il proprio patrimonio: tutti i progetti che formano la "galassia" open hanno infatti creato una comunità molto attiva e attenta nel restituire la *provenance* della risorsa utilizzata. Entrare a far parte di una comunità che è viva e in costante sviluppo può dare idee nuove e diverse su come utilizzare i propri contenuti, a volte semplicemente relegati al proprio sito, spesso non abbastanza conosciuto e visitato.

La filiera dell'open

Una piccola biblioteca di provincia possiede del materiale digitalizzato da qualche anno, ma al momento non possiede competenze interne né budget per costruire un portale web che renda queste digitalizzazioni disponibili su web.
Possiede i PDF completi di una dozzina di libri, oppure le immagini in JPG, divise per cartelle e visibili solo da terminali presenti fisicamente in biblioteca.
Decide dunque di caricare le scansioni su Internet Archive: con un semplice inserimento dei metadati descrittivi, in qualche ora di lavoro si ritrova con un libro accessibile a tutti direttamente su uno dei siti più visitati al mondo.[2] Internet Archive, inoltre, provvede a:

- fornire un visualizzatore ebook, completo di varie opzioni e ricerca termini;
- fare l'OCR sul testo;
- derivare il file originale in diversi formati, per garantire la *preservazione digitale;*
- fornire un'interfaccia API, perché altri possa usare e riusare i metadati.

In poco tempo, dunque, la nostra collezione può diventare globale e inserirsi all'interno dell'"ecosistema open".
Da Internet Archive, infatti, è possibile caricare automaticamente[3] la digitalizzazione su Wikisource, biblioteca digitale wiki e progetto-fratello di Wikipedia. Su Wikisource sarà poi possibile:

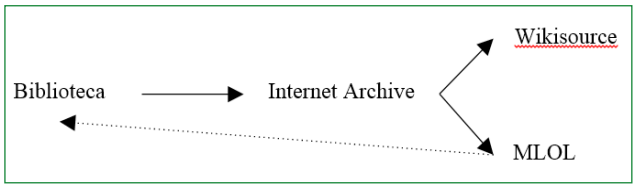
- correggere tutti gli errori dell'OCR;
- inserire link, rendendo il libro ipertestuale e connesso con altri autori e libri;
- scaricare il libro riletto e corretto in EPUB, MOBI, PDF.

Allo stesso modo, quando una risorsa è presente in Internet Archive è più semplice per altri aggregatori poter accedere ai metadati: a MLOL viene usato quotidianamente per trovare ebook in pubblico dominio e ad accesso aperto.
Facciamo un esempio concreto: *La cucina futurista*, di Filippo Tommaso Marinetti, digitalizzato dall'Università di Torino. Il libro è stato caricato dapprima su Internet Archive poi su Wikisource, nell'ambito di una collaborazione fra Gruppo di lavoro AIB Piemonte, Università di Torino e comunità di Wikisource.
I metadati del libro sono stati poi inseriti dentro MLOL, in modo tale che tutti i portali Medialibrary adesso possiedono il libro in EPUB (corretto, formattato, riletto).
Ovviamente, per tutte le biblioteche che hanno compiuto l'integrazione fra OPAC e MLOL, il libro è inoltre accessibile agli utenti tramite ricerca sul catalogo.
Dunque, in poche settimane e con un bassissimo numero di ore di lavoro, si è passati da una digitalizzazione sul repository ad un ebook disponibile gratuitamente per i lettori della biblioteca.
In questo senso, progetti come Internet Archive e Wikisource (e, in maniera minore, MLOL con le sue Risorse Open) fanno parte di una *filiera dell'open*, che origina dalla biblioteca per potenzialmente ritornarci.

Controllo di autorità degli autori

Una "biblioteca digitale come piattaforma" ha però altre opportunità.
Ad esempio, come MLOL, abbiamo iniziato ad indagare le possibilità di lavoro con i *Linked Open Data*.

La scheda di un ebook Open



VIAF

Il primo passo è stato pulire il nostro database di autori tramite un controllo di autorità, come il VIAF (Virtual International Authority File), un database che raccoglie i dati di milioni di autori dai cataloghi bibliografici di tutto il mondo.
Attraverso un processo semi-automatico di *riconciliazione*, abbiamo dunque trovato circa 84.000 identificatori VIAF su un totale di circa 148.000 autori.
È inoltre importante notare come questi numeri siano in crescita, dato che la nuova procedura di caricamento prevede da ora in poi una riconciliazione degli autori inseriti con le nuove risorse che vengono settimanalmente caricate su OpenMLOL.
La riconciliazione degli autori è una classica operazione bibliotecaria che permette di distinguere autori omonimi ma diversi: famoso e scellerato il caso dei "Alexandre Dumas" padre e figlio, scrittori che condividono lo stesso identico nome.
In questo modo, è possibile mostrare per ogni autore le sue opere *e le sue soltanto*.
L'identificazione permette anche di risolvere il problema opposto: lo stesso autore che possiede nomi diversi. È il caso di autori classici e presenti nelle biblioteche di tutto il mondo: Omero, Aristotele, Platone, Cicerone, Dante. Oppure di autori russi, cinesi, giapponesi che sono stati traslitterati in modi diversi, anche nella stessa lingua.
Grazie alla riconciliazione dei dati, vengono dunque eliminati (dove possibile) i problemi di *sinonimia* e *omonimia*.

La scheda autore riconciliata attraverso VIAF

Wikidata

Un ulteriore passaggio è stato collegare gli autori con Wikidata: un collegamento possibile proprio perché il database Wikidata ha già al suo interno la corrispondenza con il VIAF. Questo ci ha permesso di collegarci, tramite API, direttamente con la pagina Wikipedia e creare degli *snippet* riassuntivi degli autori.
Un modo per rendere le nostre schede più chiare e ricche, e garantire ai lettori un'esperienza di ricerca migliore.
Infine, abbiamo anche integrato le risorse open con Wikidata inserendo un "identificatore openMLOL" *all'interno* di Wikidata, instaurando un *collegamento diretto* con quello che è a tutti gli effetti il progetto di database semantico più importante finora. Un inizio per altri futuri, e maggiori, esperimenti.
Tramite questo identificatore, infatti, abbiamo la possibilità di poter svolgere query complesse (attraverso l'endpoint SPARQL di Wikidata) sugli autori MLOL ad esso collegati. Possiamo, ad esempio, conoscere quanti autori maschi e femmine sono presenti nella nostra biblioteca oppure indagare sulla loro nazionalità.
È importante ricordare come la query possa esser fatta in *real time*: nel momento in cui la collezione si amplia, o i dati di Wikidata diventino essi stessi più ricchi e granulari, il risultato viene aggiornato.

Machine-learning

Un ulteriore filone di ricerca e sperimentazione è il *machine learning*: nello specifico, la possibilità di avere una soggettazione automatica o semi-automatica delle risorse. La questione della catalogazione di risorse è uno dei temi

centrali del lavoro bibliotecario, che può essere coadiuvato dallo sviluppo costante di sistemi di relazione e semantizzazione delle informazioni. Nel caso di MLOL, aggregando un'enorme quantità di contenuti aperti dalla rete abbiamo necessità di un supporto automatico alla soggettazione: è infatti impossibile e anti-economico catalogare "a mano" oltre mezzo milione di risorse.
Tramite l'accesso a Wikidata, abbiamo la possibilità di avere informazioni sugli autori che non avremmo altrimenti: la professione, la data di nascita e morte, il sesso, il genere degli scritti dell'autore, saggi o romanzi più rilevanti ecc.
Da tutti questi dati è possibile quindi trarre informazioni sul tipo di opere a cui siamo di fronte. Una mole accessibile facilmente tramite query SPARQL e API.
Questi esperimenti trattano essenzialmente la rielaborazione dei metadati, che possono essere rimodulati in modo da estrapolare informazioni terze. In questo modo il lavoro bibliotecario si ibrida e i metadati stessi assumono una concezione differente: non più semplicemente metadati descrittivi o strutturali, ma diventano uno strumento di operabilità del documento. In quest'ottica è estremamente importante il discorso sull'omogeneità degli identificativi, che permettono appunto il collegamento tra dati differenti.
In questo modo è possibile sperimentare elementi di machine-learning per inserire anche nel campo e nel lavoro del bibliotecario alcuni elementi di intelligenza artificiale, normalmente di attinenza informatica, per proiettare in un contesto più complesso il lavoro di catalogazione e offerta delle informazioni.

ANDREA ZANNI
MLOL
andrea.zanni@medialibrary.it

NOTE

[1] http://lj.libraryjournal.com/2012/09/future-of-libraries/by-david-weinberger/
[2] Al 18 aprile 2017, numero 264 secondo il ranking mondiale Alexa.
[3] Attraverso il sito https://tools.wmflabs.org/ia-upload/commons/init.

Cavalcare la tigre dei social network

L'iniziativa del MAB Lombardia, realizzata nell'ambito del Convegno delle Stelline con la collaborazione della Regione Lombardia, ha registrato un notevole successo, considerando il numero delle presenze varie tra gli iscritti di professionisti dei beni culturali. L'obiettivo del workshop MAB (Musei, Archivi, Biblioteche) era particolarmente ambizioso, come lasciava intendere il titolo stesso: *Cavalcare la tigre dei social network: musei, archivi e biblioteche tra Open Access e Big Data*.
L'uso dei social network e l'Open Access come paradigmi generalizzati e la creazione di Big Data offrono alle istituzioni culturali della memoria opportunità tuttora inesplorate.
Alla base del tema scelto i risultati molto stimolanti di un progetto di ricerca dell'Osservatorio innovazione digitale nei beni e attività culturali del Politecnico di Milano presentato recentemente.[1] La collaborazione offerta dal gruppo di lavoro coordinato da Eleonora Lorenzini è stata basilare per il workshop del MAB. Dalla ricerca, articolata in diverse fasi e assai complessa, si sono enucleati ai fini del workshop alcuni aspetti che apparivano rilevanti come suggerimenti operativi. I risultati della ricerca costituiscono infatti una preziosa e inaspettata fonte di indicazioni e ipotesi di lavoro, riferite principalmente all'ambito museale, ma da percorrere e trasporre in buona misura anche in quello bibliotecario e archivistico. Ciò che tali risultati ci fanno capire è che occorre adottare un approccio metodologico diverso dal tradizionale, che individui nella misurazione delle performance degli istituti della memoria un modello necessario da applicare alla gestione digitale dei beni e dei servizi nella quotidianità, volto alla conoscenza dei medesimi e per creare valore attraverso l'innovazione digitale, oltre che contribuire alla loro gestione