Aggregatori di dati e linked data: grande opportunità per il patrimonio culturale

Fabio Di Giammarco
Biblioteca di Storia moderna e contemporanea

La valorizzazione del processo globale di digitalizzazione in corso è soprattutto affidata ai grandi aggregatori di dati del patrimonio culturale. Si tratta di organizzazioni – sia pubbliche che private – di livello europeo, nazionale, regionale e locale che raccolgono metadati da tutta una serie di fornitori per poi renderli operabili mediante piattaforme ad hoc. Tra questi, diversi sono ormai i modelli di caratura internazionale: ad esempio, riguardo l'arte dell'antica Grecia e di Roma, *Claros*[1], aggregatore frutto della collaborazione delle università di Oxford, Colonia e Parigi che consente ricerche simultanee all'interno di importanti collezioni in istituti di ricerca e musei.

In Italia, il più importante aggregatore nazionale – progetto nato nell'ambito dell'impegno dell'Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche (ICCU) per la digitalizzazione del patrimonio culturale[2] e che si avvale della consulenza scientifica della Scuola Normale di Pisa – è CulturaItalia[3]. Si tratta di un sistema nella doppia veste portale/aggregatore che, grazie a soluzioni informatiche innovative, è in grado di raccogliere, organizzare e mettere a disposizione milioni di informazioni riguardanti il patrimonio culturale del paese. I dati sulle risorse non sono prodotte direttamente da CulturaItalia, ma forniti dai soggetti che le posseggono e le gestiscono. Tutte le parti dell'ecosistema culturale – istituzioni pubbliche, imprese private ecc. – possono trasferire nel database di CulturaItalia i metadati, ovvero le informazioni descrittive delle loro risorse digitali. L'utente poi, interagendo con il portale/aggregatore, accede alla base dati costituita dai fornitori convenzionati. In questo modo può scoprire risorse di vario genere provenienti da musei, archivi fotografici, archivi audiovisivi, biblioteche, ma anche mostre, monumenti ecc. Risorse in grado di soddisfare non solo semplici curiosità culturali ma anche intenti di ricerca scientifica.

Il progetto CulturaItalia è basato sulla collaborazione. I vari fornitori mettono a disposizione i metadati che descrivono le risorse delle proprie basi dati. In pratica, quando un nuovo *content provider* si propone, si procede al *mapping* dei dati cui segue l'installazione del *repository* OAI-PMH[4], quindi si estraggono i dati. A questo punto, effettuato l'*harve-*

[1] http://www.clarosnet.org/XDB/ASP/clarosHome/.
[2] http://leo.cineca.it/index.php/jlis/article/view/8726/7886.
[3] http://www.culturaitalia.it/.
[4] OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting* o *Protocollo per il raccoglimento dei metadati dell'Open Archive Initiative*) è un protocollo sviluppato dall'Open Archives Initiative come infrastruttura di comunicazione per l'Open access. È utilizzato per raccogliere (o collezionare) i metadati dei documenti in un archivio affinché i servizi possano essere costruiti utilizzando metadati da più

sting dei metadati si passa alla loro pubblicazione su CulturaItalia. Al momento, l'indice dei metadati è alimentato da circa trenta tra istituzioni e organizzazioni della cultura italiana e contiene oltre 2 milioni di dati.

Sia per la sua competenza nel dare supporto tecnico e formativo agli istituti culturali durante l'aggregazione dei contenuti, che per la sua capacità di mettere a sistema il capitale informativo di tanti soggetti diversi, l'aggregatore CulturaItalia si propone come punto di riferimento non solo a livello nazionale ma anche europeo. Tra l'altro, rivendicando un'impostazione fortemente "open": e non solo per le sue caratteristiche di sistema "aperto", ma anche e soprattutto per l'utilizzo nella regolazione dei diritti della Licenza CCO 1.0 Public Domain Dedication corrispondente a una dichiarazione di rinuncia al copyright con applicazione universale.

In generale, la configurazione dei dataset del patrimonio culturale europeo si avvale di un sistema multilivello dove aggregatori più piccoli – nazionali, regionali, locali ecc. - fungono nello stesso tempo da raccoglitori ma anche da fornitori per aggregatori più grandi. CulturaItalia, che è il principale aggregatore nazionale, è anche fornitore del principale aggregatore europeo: *Europeana*[5], vale a dire del grande progetto della UE al quale partecipano tutti gli Stati membri. Si tratta di un portale multilingue (l'interfaccia è disponibile in 29 lingue) dietro a cui opera un sistema di aggregazione e indicizzazione di dati afferenti a risorse digitali provenienti da oltre 1500 istituzioni – per la maggior parte musei, biblioteche, archivi, centri culturali ecc. - sparse sul territorio europeo.

Europeana nel suo ruolo di grande piattaforma di aggregazione del patrimonio culturale, utilizza il modello dati EDM[6]: uno standard intermedio studiato appositamente per rendere più semplice il processo di trasformazione dei metadati in entrata e nello stesso tempo di garantire la persistenza dei record anche in caso di future evoluzioni degli standard di riferimento. Massima apertura, invece, per quanto riguarda la gestione dei diritti.

Anche per *Europeana* – come per CulturaItalia – la scelta è andata sulla licenza CCO 1.0 Public Domain che rende possibile utilizzare i metadati pubblicati per qualunque scopo, compresi quelli commerciali. Unica condizione richiesta ai fornitori è la sottoscrizione dell'accordo DEA[7] che comunque concede libertà alle istituzioni che rilasciano i propri dati di decidere quanti e quali (limitando, ad esempio, i permessi solo ad alcuni set di metadati) rendere interoperabili attraverso *Europeana*.

Insomma, *Europeana* con il suo enorme data set di quasi 26 milioni di metadati che descrivono e rimandano a libri, audiovisivi, immagini, manoscritti, documenti sonori, applicazioni in 3D ecc., rappresenta un'eccezionale piattaforma attraverso la quale costruire strategie operative per la valorizzazione (anche economica) e la fruizione del patrimonio culturale comune europeo. Non a caso, uno dei cardini del progetto è il riuso aperto dei dati. Una politica con ricadute importanti per tutti i soggetti pubblici e privati aggregati al progetto: maggiore visibilità e traffico verso i rispettivi siti, sviluppo di servizi innovativi e stimolo per l'industria creativa, maggiori opportunità di generare entrate mediante distribuzione di risorse digitali ecc. La chiave per attuare queste strategie espansive e innovative è nella tecnologia dei *linked open data*.

Con l'avvio nel 2012 del progetto LOD (Linked Open Data), *Europeana* si è fatta pro-

archivi. Una implementazione dell'OAI-PMH deve supportare metadati rappresentati in Dublin Core, ma può supportare altre rappresentazioni (Wikipedia).

[5] http://www.europeana.eu/.

[6] Europeana Data Model.

[7] Europeana Data Exchange Agreement.

motrice dell'iniziativa di incentivare lo sviluppo, in ambito culturale, di applicazioni innovative per la produzione di contenuti culturali e la creazione di nuovi servizi sul web. Infatti, è dall'1 luglio 2012 che *Europeana* presenta i metadati anche sotto forma di LOD con tanto di licenza di pubblico dominio e autorizzazione al riuso in base a quanto previsto dall'accordo DEA. I LOD sono dati "grezzi" pubblicati in formato RDF[8], soggetti poi a operazioni di potenziamento per essere usati come base per fornire servizi a valore aggiunto per cittadini e imprese. Tra queste operazioni di potenziamento, quella, fondamentale, dell'arricchimento semantico è effettuata con diverse modalità: per i luoghi (GeoNames), per i concetti (Thesaurus GEMET), per le persone (Dbpedia) e per i periodi di tempo un vocabolario ad hoc.

La tecnologia dei linked data è la rivoluzione che può rendere interoperabili e quindi produttivi i grandi dataset del patrimonio culturale. I collegamenti che si vengono a creare tra le entità descritte nei dataset rendono possibili realizzazioni di applicazioni in grado poi – utilizzando le "relazioni" del reticolo linked data – di "saltare" da un dataset all'altro. I vantaggi sono molteplici ed esponenziali: vanno dalla riduzione della duplicazione delle informazioni, alla possibilità di condivisione per un uso efficiente delle risorse pubbliche, fino alla capacità di fornire dati di alta qualità e soprattutto contestualmente utili per un loro riutilizzo diretto in settori fondamentali, quali l'impresa privata, la formazione, la ricerca scientifica e la cultura.

Rispetto al mondo delle biblioteche, la possibilità, creata dai reticoli LOD, di connettere i grandi aggregatori del patrimonio culturale può significare un cambiamento epocale. Oggi ci sono milioni di dati bibliografici conservati negli OPAC delle biblioteche di tutto il mondo non raggiungibili via Web perché creati e registrati con un formato ormai totalmente superato, il MARC. Il rapporto del W3C[9] denominato *Library Linked Data Incubator*[10] già da tempo ha sostenuto la fondamentale importanza dei collegamenti tra dati bibliografici e informazioni prodotte all'esterno. Non dovrebbe più esistere distinzione – sia per la forma che per il luogo di registrazione – tra dati prodotti dalle biblioteche e dati di natura diversa. Insomma, il concetto da cui scaturisce tutta l'analisi del *Library Linked Data Incubator* è che nel Web esistono solo dati condivisibili, modulari, riutilizzabili.

Uno dei grandi vantaggi nell'applicare la tecnologia dei linked data ai cataloghi bibliotecari, è quella di "contestualizzare" l'informazione creando collegamenti tra saperi diversi durante le procedure di ricerca. Utilizzando, in altre parole, il "web dei dati" come "contesto esteso" per l'esplorazione dei contenuti. Quando i dati del catalogo sono strutturati come linked data si vengono a creare record formati da triple RDF tutte indentificate da URI[11]. In questo modo, ogni record è potenzialmente connesso al reticolo illimitato dei dati del Web semantico.

Un altro esempio dell'estensione della ricerca grazie ai linked data è la possibilità di ricostruire intorno a un documento – ad esempio un libro – il contesto storico, culturale e letterario nel quale si inserisce: il che significa gestire nuovi ambienti esplorabili mediante collegamenti trasversali di dati bibliografici, oggetti digitali, contenuti multimediali e altri materiali prodotti dalle comunità del web.

[8] Resource Description Framework.

[9] Il W3C è un'organizzazione non governativa internazionale che ha come scopo quello di sviluppare tutte le potenzialità del World Wide Web. Al fine di riuscire nel proprio intento, la principale attività svolta dal W3C consiste nello stabilire standard tecnici per il World Wide Web inerenti sia i linguaggi di markup che i protocolli di comunicazione.

[10] http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/.

[11] Uniform Resource Identifier.

In pratica, tramite la rete di collegamenti tra i dataset, dal patrimonio culturale diverrà automatico passare da un soggetto a un oggetto museale (quadro, scultura ecc.) che lo rappresenta, oppure attivare connessioni tra dati bibliografici e cose reali descritte nel web. Ad esempio, un utente, sfruttando il collegamento tra dataset delle biblioteche e database geografici, sarà in grado di visualizzare mappe geografiche di autori che pubblicano in un determinato paese o che hanno avuto maggiore influenza in una certa area geografica.
Dal punto di vista dell'utente, un primo effetto del cambiamento è che un qualsivoglia oggetto di ricerca sarà raggiungibile – indifferentemente – tanto da un motore di ricerca che da un OPAC di un servizio bibliotecario. La fondamentale differenza tra le due opzioni di ricerca, sarà casomai che i linked data creati dalle biblioteche dovranno avere standard di qualità, di coerenza e autorevolezza decisamenti superiori; mentre, da un punto di vista tecnico, la ricerca dell'utente mediante "web dei dati", funzionerà, in pratica, recuperando tutte le relazioni tra certi dati identificati a loro volta da specifiche URI: ad esempio, le relazioni tra il "dato libro" con il "dato titolo", "dato autore", "dato luogo di stampa" ecc.

Infine, per quel che riguarda la visualizzazione dei nuovi dati bibliografici – ovvero le interfacce OPAC – i cambiamenti introdotti dalle nuove tecnologie del Web sembrano procedere con un passo ancora più spedito. Le visualizzazioni possono già avvalersi di pagine dinamiche per ogni entità di interesse del catalogo, come un'opera, un soggetto, un autore ecc. È quello che, in pratica, sta facendo la Bibliothéque nationale de France con il suo progetto: data.bnf.fr[12]. A partire da un determinato soggetto viene, infatti, creata dinamicamente una pagina in grado di stabilire collegamenti a documenti, film, immagini, risorse multimediali e collegamenti esterni. Visualizzazioni avanzate create a partire dalle entità di interesse del catalogo consentono poi il recupero dell'informazione bibliografica in un contesto arricchito da altre risorse autorevoli provenienti dal web, realizzando così le condizioni per il passaggio – rispetto alla ricerca utente e alle interfacce OPAC – dal tipico recupero ispirato sull'*information retrieval* a un recupero basato, invece, sul modello "entità-relazioni", concettualmente molto efficace nell'ambito della nuova ricerca attraverso il Web semantico.

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

L'ultima consultazione dei siti Web è avvenuta nel mese di giugno 2014.