

Il catalogo che cresce: gli autori della Beic Digital Library in Wikidata

Federica Viazzi

Fondazione BEIC

Abstract. Presentazione delle fasi di lavorazione di una delle attività del progetto GLAM/BEIC: l'esportazione in Wikidata dei record relativi agli autori della Biblioteca Digitale BEIC. Vengono descritti i passaggi tecnici necessari al riversamento dei dati nel database. Prima la preparazione dei metadati associati agli autori persona, la loro importazione in uno strumento wiki che permette di confrontare i record presenti in un catalogo con gli elementi esistenti in Wikidata, e di abbinarli nel caso si riferiscano alle stesse entità: Mix'n'match. Successivamente, l'inserimento tramite QuickStatements, un altro tool che permette di intervenire attraverso procedure semi-automatiche, di coppie proprietà-valore secondo la struttura propria a Wikidata. L'ultima parte descrive infine la creazione di nuovi elementi Wikidata, nel caso in cui non vi fossero corrispondenze tra i record del catalogo e il database, sempre tramite lo strumento QuickStatements.

Keywords. Metadati, Wikidata, Biblioteche, Digital humanities

Introduzione

La Fondazione BEIC partecipa dal 2014, a un progetto GLAM-wiki (Galleries, Libraries and Museums) in collaborazione con l'associazione Wikimedia Italia, il capitolo nazionale della Wikimedia Foundation. Nell'ottica di una condivisione aperta dei contenuti e di diffusione della conoscenza libera, ha l'obiettivo di disseminare le risorse della Biblioteca Digitale all'interno di Wikipedia e dei progetti fratelli.

Dall'inizio del GLAM a oggi, le attività hanno spaziato dalla contribuzione alle voci dell'enciclopedia libera, al caricamento di immagini digitalizzate in Wikimedia Commons, alla collaborazione con i progetti di Wikisource. Negli ultimi anni, tuttavia, è cresciuto l'interesse verso Wikidata, tra i progetti fratelli di Wikipedia quello più affine al mondo bibliotecario.

1. Wikidata

Wikidata è un database libero, collaborativo, multilingue e secondario che raccoglie dati strutturati per fornire supporto a Wikipedia, a Wikimedia Commons, agli altri progetti del movimento Wikimedia, e a chiunque nel mondo.

Si tratta del progetto che più di tutti si avvicina ai modelli di gestione dei dati bibliografici: esso infatti assume fra le basi concettuali della sua struttura principi che sono fondamentali nella teoria della catalogazione, come il controllo di autorità. L'affinità di Wikidata con le prassi di lavoro bibliotecarie è ulteriormente confermata dal fatto che le entità descritte in Wikidata hanno cominciato precocemente ad arricchirsi degli identificativi

persistenti provenienti da dataset bibliografici autorevoli, come quello della Bibliothèque nationale de France o il CERL Thesaurus.

La decisione di riversare i metadati di autorità della Biblioteca digitale BEIC in Wikidata risale al 2017. I metadati BEIC sottoposti a controllo di autorità sono molti: ovviamente gli autori, ma anche gli editori, i luoghi, i titoli di opere e di espressioni.

2. Le fasi di lavorazione

L'insieme di partenza ha incluso i soli autori di tipo persona che ricorressero come intestazioni principali dei record bibliografici, per un totale di circa 5000 nomi; successivamente, si è deciso di includere anche tutti gli autori persona con intestazione secondaria, arrivando a comprendere un set di quasi 15000 nomi.

2.1 Preparazione dei dati

Per arrivare al riversamento dei dati di autorità in Wikidata, la fase preliminare necessaria è stata la preparazione dei metadati di partenza. La Biblioteca digitale BEIC utilizza lo standard MARC21 per la creazione dei record bibliografici, pertanto tutte le informazioni relative agli autori di tipo persona sono state rintracciate all'interno dei tag di tipo 100 per le intestazioni principali, e di tipo 700 per le intestazioni secondarie. Ciascuno di questi campi è arricchito con informazioni strutturate che permettono l'identificazione univoca della persona: nome e cognome, date di nascita e morte, eventuali titoli associati, numerali nel caso di cariche politiche (ad es. papi e imperatori).

È importante segnalare che il protocollo catalografico BEIC prevede l'inserimento delle date nel record di autorità in ogni caso, e non soltanto per disambiguare gli omonimi. Questo dato stabile ha avuto grande utilità nel progetto di riversamento in Wikidata, poiché ha permesso un'identificazione più certa e al contempo ha assicurato la presenza di un dato di tipo enciclopedico in tutte le registrazioni.

I metadati associati agli autori persona sono stati esportati dal catalogo, rielaborati e riversati in un file unico, che è stato a sua volta importato in uno strumento wiki specifico: Mix'n'match. Creato da Magnus Manske, uno dei membri della comunità wiki internazionale, Mix'n'match è un tool che permette con relativa facilità di confrontare i record presenti in un catalogo con gli elementi esistenti in Wikidata e abbinarli nel caso si riferiscano alle stesse entità.

2.2 I due insiemi

Una volta completata questa fase, Mix'n'match ha presentato due insiemi di risultati: uno che comprende tutti gli elementi abbinati, cioè tutti i casi in cui un nome presente nel catalogo di partenza corrisponde a un elemento già esistente in Wikidata; un altro che include tutti gli unmatched, ovvero i nomi presenti nel catalogo di partenza che non hanno alcun abbinamento con entità di Wikidata.

2.3 L'arricchimento con l'entità Biblioteca digitale BEIC degli elementi esistenti

Nel primo caso, gli elementi Wikidata coinvolti avrebbero dovuto solamente essere ar-

ricchiti da una nuova coppia proprietà-valore che evidenziasse la corrispondenza con le singole entità del Catalogo BEIC. Per fare ciò, è stato necessario, innanzitutto, creare un elemento Wikidata che descrivesse la Biblioteca digitale BEIC.

Data la mole di dati, sarebbe stato impensabile procedere modificando uno per volta i singoli elementi da arricchire, pertanto si è deciso di utilizzare un altro strumento che permette di intervenire in Wikidata in maniera semiautomatica: QuickStatements. Tramite questo strumento, partendo da un elenco di elementi Wikidata, è possibile inserire lo stesso tipo di informazione in tutti con un'unica operazione. Nel nostro caso, l'informazione da inserire era la forma controllata dell'autore, con la modalità propria di Wikidata, cioè tramite una coppia proprietà-valore. I dati sono quindi stati impostati secondo una struttura che accoppia la proprietà "Descritto nella fonte" al valore "Biblioteca digitale BEIC", a sua volta arricchito da due riferimenti: la forma del nome e l'URL che lancia una ricerca corrispondente nel Discovery tool della biblioteca digitale.

2.4 La creazione di nuovi elementi Wikidata

Il secondo possibile risultato del lavoro in Mix'n'Match include tutti i casi in cui non è stato possibile creare una corrispondenza fra un autore BEIC e un elemento Wikidata. Questa eventualità è risultata essere comunque minoritaria e ha determinato una fase di lavoro differente.

Gli elementi da creare non avrebbero dovuto contenere solo la coppia proprietà-valore relativa alla presenza nella Biblioteca digitale BEIC, ma anche le informazioni usate per descrivere le persone: nome, cognome, date di nascita e morte. Queste informazioni sono state desunte dai metadati bibliografici presenti nel catalogo e successivamente normalizzate in una struttura adatta all'immissione in Wikidata.

Oltre a ciò, si è scelto di fare uno sforzo ulteriore: per ciascun elemento sono state specificate le proprietà 'sesso' e 'occupazione', per le quali è stato necessario incrociare informazioni derivanti da fonti diverse.

3. Conclusioni

Il risultato finale di questo lavoro vede l'entità "Biblioteca digitale BEIC" associata tramite la proprietà "Descritto nella fonte" a 14666 elementi Wikidata. Il prossimo passo sarà quello di provare a chiudere il cerchio della condivisione: la prospettiva è infatti di integrare le informazioni sempre più corpose e precise presenti nelle entità di Wikidata all'interno dei record di autorità, a partire dagli identificativi persistenti e dalle forme varianti del nome.

Un'ulteriore prospettiva è data dalla volontà di migliorare la struttura dei record di autorità del Catalogo BEIC, creando per ciascuno di essi un identificativo persistente, che dovrà essere il riferimento stabile al di fuori del catalogo e, ovviamente, negli elementi Wikidata. L'adozione di un identificativo semplificherà significativamente le operazioni su Wikidata, in particolare nelle fasi di aggiornamento.

Una volta concluso il progetto sugli autori di tipo persona, l'obiettivo è quello di estendere la prassi fin qui descritta anche agli altri metadati sottoposti al controllo di autorità.

In particolare, si lavorerà sulle forme normalizzate degli autori di tipo ente, degli editori antichi e moderni – includendo anche quelli presenti nel catalogo dell'Archivio della Produzione Editoriale Lombarda, sempre allestito dalla BEIC – e dei titoli di opere ed espressioni.

Concludendo, l'incontro e talvolta lo scontro tra catalogo e Wikidata hanno generato risultati senza dubbio molto positivi: l'esposizione dei dati di autorità della Biblioteca digitale BEIC al di fuori del contesto del catalogo ha permesso il loro arricchimento e contemporaneamente ha dato visibilità alle risorse descritte.

Riferimenti bibliografici

GLAM su Meta-Wikimedia <https://meta.wikimedia.org/wiki/GLAM>

Introduzione a Wikidata <https://www.wikidata.org/wiki/Wikidata:Introduction/i>

Autrice



Federica Viazzi - federica.viazzi@gmail.com

Dopo la laurea in Conservazione dei Beni Culturali presso l'Università di Pavia si occupa per molti anni di libro antico approfondendone la conoscenza alla British Library di Londra. Parallelamente coltiva interessi per i metadati e le biblioteche digitali. Attualmente occupa il ruolo di responsabile per i metadati strutturali presso la biblioteca digitale della Fondazione BEIC di Milano e contribuisce allo sviluppo del progetto GLAM in collaborazione con Wikimedia Italia.