Linked Books: un indice citazionale per la storia di Venezia

Colavizza, G.; Romanello, M.; Giuliano, A.; Mataloni, M.C.; Grandin, D.

Publication date
2019
Document Version
Final published version
Published in
DigItalia

DigItalia

Anno XIV, Numero 1 - **2019**

Rivista del digitale nei beni culturali

ICCU-ROMA

ICCU

Istituto centrale per il catalogo unico
delle biblioteche italiane e per le informazioni bibliografiche
https://www.iccu.sbn.it

DigItalia

Rivista del digitale nei beni culturali
ISSN 1972-6201
Anno XIV, Numero 1 - Giugno 2019

In copertina:
L'immagine è una libera elaborazione grafica della testa della statua di Apollo del I sec. d.c. (Civitavecchia, Museo Nazionale), copia da un originale greco avvicinabile all'Apollo di Leochares (IV sec. a.c.)

Produzione e Stampa
Istituto Poligrafico e Zecca dello Stato S.p.A.
Roma

Editore
ICCU
Istituto centrale per il catalogo unico
delle biblioteche italiane
e per le informazioni bibliografiche
Viale Castro Pretorio, 105
00185 Roma
T +39 06 49.210.425
F +39 06 49.59.302
email: digitalia@iccu.sbn.it
http://digitalia.sbn.it

In attesa di registrazione al Tribunale di Roma

ICCU

Comitato Scientifico

Osvaldo Avallone
Giovanni Bergamin
Dimitri Brunetti
Simonetta Buttò
Rossella Caffo
Rosaria Campioni
Maria Carla Cavagnis Sotgiu
Laura Ciancio
Flavia Cristiano
Gianfranco Crupi
Andrea De Pasquale
Maria Cristina Di Martino
Pierluigi Feliciati
Marina Giannetto
Maria Guercio
Mauro Guerrini
Klaus Kempf
Patrizia Martini
Maurizio Messina
Maria Cristina Misiti
Maria Teresa Natale
Marco Paoli
Don Valerio Pennasso
Alberto Petrucciani
Massimo Pistacchi
Marco Pizzo
Paola Puglisi
Roberto Raieli
Gino Roncaglia
Maria Letizia Sebastiani
Giovanni Solimine
Laura Tallandini
Anna Maria Tammaro
Costantino Thanos
Antonella Trombone
Paul Gabriele Weston

SOMMARIO

giugno 2019

SEGNALAZIONI

Linked Books: un indice citazionale per la storia di Venezia

Giovanni Colavizza - *University of Amsterdam (NL)*
Matteo Romanello - *Digital Humanities Laboratory, EPFL (CH)*
Andrea Giuliano - *Biblioteca Nazionale Centrale di Roma (BNCR)*
Maria Cristina Mataloni - *Istituto Centrale per il Catalogo Unico (ICCU)*
Daniela Grandin - *Università Ca' Foscari Venezia*

Presentiamo i risultati del progetto Linked Books, che ha portato alla creazione di un prototipo per un indice citazionale capace di collegare il catalogo bibliotecario nazionale italiano (Opac SBN) con il sistema informativo dell'Archivio di Stato di Venezia e con i portali d'autorità e "metamotori" di ricerca internazionali (VIAF.org, Europeana). Il prototipo include 3.850.581 citazioni estratte da un corpus di 2.475 volumi, di cui 1.905 monografie e 552 numeri di rivista, da cui 5.496 articoli, riguardanti la storia di Venezia. Il progetto Linked Books ha permesso di esplorare la fattibilità e l'opportunità della creazione di un indice citazionale per discipline umanistiche, e di affrontare e risolvere ostacoli tecnici quali la creazione di un corpus rappresentativo a partire da risorse e competenze bibliografiche, la digitalizzazione dei materiali nel rispetto dei diritti d'autore, l'estrazione automatica di citazioni e lo sviluppo di interfacce di ricerca per il pubblico.

Introduzione

I materiali bibliografici e le fonti per lo studio nelle discipline umanistiche sono allo stesso tempo abbondanti e dipendenti dal contesto in cui si trovano. Una collezione bibliotecaria o un fondo archivistico richiedono, di conseguenza, una cura specifica per la loro preservazione, studio e messa a disposizione degli studiosi interessati. In questo modo, e specialmente in paesi come l'Italia, istituti culturali come archivi e biblioteche si sono specializzati nel valorizzare specifiche collezioni. Anche servizi digitali inter-istituzionali, come il Catalogo del Servizio Bibliotecario Nazionale[1] ed il Sistema Archivistico Nazionale[2] non comunicano tra loro, pur portando organicità e permettendo l'accesso centralizzato a (meta)dati all'interno di uno stesso ambito. Tuttavia, nella letteratura scientifica collegamenti espliciti a materiali bibliografici e fonti primarie di varia natura sono presenti sotto forma di *citazioni*, reperibili in riferimenti a piè di pagina o in liste apposite. Un *in-*

[1] https://opac.sbn.it.
[2] http://san.beniculturali.it.

dice citazionale per le discipline umanistiche che consideri questi collegamenti può (e deve) mettere in relazione ambiti documentari per ora separati, e fornire agli utenti un punto di accesso e uno strumento di navigazione finora insperato.
Il progetto *Linked Books* è nato con l'obiettivo di creare un prototipo per un indice citazionale capace di collegare il catalogo bibliotecario nazionale italiano (Opac SBN) con i sistemi informativi archivistici e con i portali d'autorità oppure con i "metamotori" di ricerca internazionali (VIAF.org, Europeana)[3]. Il progetto si è focalizzato su un caso di studio specifico: la storiografia su Venezia e il collegamento tra l'Opac SBN e il sistema informativo dell'Archivio di Stato di Venezia (SiASVe)[4]. L'interesse scientifico del progetto risiede nella volontà di studiare la fattibilità di un indice citazionale per le discipline umanistiche. Il caso di studio proposto trova invece la sua motivazione negli scopi specifici del progetto *Venice Time Machine*[5].
Il progetto *Venice Time Machine*, iniziato nel 2012 da una collaborazione tra l'*École polytechnique fédérale de Lausanne* (EPFL), l'Archivio di Stato di Venezia e l'Università Ca' Foscari di Venezia, ha come scopo la creazione di un sistema informativo digitale riguardante ogni aspetto di Venezia e della sua storia, dall'urbanistica alla geografia, dall'arte ai materiali bibliografici ed archivistici. Questi ultimi – le collezioni dell'Archivio di Stato – hanno costituito fin dal principio il cantiere di lavoro principale. La necessità, dapprima pratica poi vieppiù ragionata, di reperire dati sull'uso delle collezioni dell'Archivio che la comunità di ricerca ha fatto e continua a fare, è stata all'origine del progetto *Linked Books*: un indice citazionale per la storia di Venezia. Il progetto, iniziato nel 2014, è stato ufficializzato nel 2015 e sviluppato fino al 2018. Al progetto hanno collaborato l'EPFL, l'Archivio di Stato di Venezia, l'Università Ca' Foscari tramite il Sistema Bibliotecario di Ateneo e la Biblioteca di Area Umanistica (BAUM), la Biblioteca Nazionale Marciana, l'Istituto Veneto di Scienze, Lettere ed Arti, la Biblioteca Europea di Informazione e Cultura (BEIC) e l'Istituto Centrale per il Catalogo Unico (ICCU)[6].
Nel corso di questi anni, il progetto *Linked Books* ha realizzato la digitalizzazione di un corpus di circa 2000 monografie e oltre 500 numeri di rivista, da cui sono stati estratti circa 4 milioni di riferimenti a fonti primarie e secondarie. Due strumenti di ricerca digitali sono stati inoltre sviluppati e pubblicati in linea: una biblioteca digitale (*Venice Scholar – Library*)[7] e l'indice citazionale (*Venice Scholar*)[8]. Questi due strumenti comunicano tra di loro e con tutti i cataloghi o sistemi informativi menzionati precedentemente. Il progetto *Linked Books* ha dimostrato come

[3] https://viaf.org, https://www.europeana.eu.
[4] http://www.archiviodistatovenezia.it/siasve.
[5] https://vtm.epfl.ch.
[6] Il progetto è stato sostenuto dal Fondo di ricerca nazionale svizzero con due finanziamenti, progetti 205121_159961 e P1ELP2_168489, da un *research grant* di Europeana e dal supporto continuo di EPFL e della Fondazione Lombard Odier di Ginevra.
[7] http://venicescholar.dhlab.epfl.ch/library.
[8] https://venicescholar.dhlab.epfl.ch.

un indice citazionale possa agire come aggregatore ed integratore di sistemi informativi provenienti da ambiti culturali differenti, come archivi e biblioteche. Ha inoltre offerto lo spunto per un approccio alla creazione di indici citazionali collaborativi e distribuiti, sul modello dei cataloghi bibliotecari nazionali.
La sezione che segue offre una breve digressione sulla storia degli indici citazionali e la loro copertura delle discipline umanistiche. In seguito, presentiamo il nostro approccio alla selezione dei materiali bibliografici da cui estrarre citazioni, il contesto legale per quanto riguarda il diritto d'autore e l'acquisizione dei dati e metadati. Dettagliamo, quindi, la metodologia di estrazione delle citazioni con procedure automatizzate, nonché lo sviluppo delle interfacce di ricerca per gli utilizzatori. Il contributo si conclude con una discussione e degli spunti di lavoro per il futuro.

Indici citazionali e discipline umanistiche

Gli indici citazionali nascono dalla volontà di superare i limiti della catalogazione per soggetto, in presenza di una crescente massa di materiale pubblicato, e offrire uno strumento non mediato per il reperimento della letteratura d'interesse. Il primo indice citazionale, *Science Citation Index (SCI)*, ebbe per antecedenti empirici gli strumenti di indicizzazione della giurisprudenza (secolo diciannovesimo) e gli esperimenti in bibliografia statistica risalenti ai decenni subito precedenti. Al primo volume dello SCI, pubblicato nel 1963, ha fatto seguito un'espansione degli studi quantitativi sulla scienza (*scientometrics*). Il legame simbiotico tra lo SCI e lo sviluppo del campo "scientometrico" è stato paragonato a quello fra il sincrotrone e la fisica delle particelle moderna[9]. Nonostante il lancio già nel 1978 di un indice specializzato per le discipline umanistiche, l'*Arts & Humanities Citation Index (A&HCI)*, il suo impatto per gli studiosi e per gli stessi "scientometristi" è stato quantomeno marginale[10].
Diversi indici citazionali commerciali sono attualmente a disposizione della comunità scientifica, e l'offerta è in continua espansione. Gli indici a pagamento includono *Web of Science* (l'attuale evoluzione dello SCI) e *Scopus*, mentre *Google Scholar, Dimensions* e *Microsoft Academic*, per limitarsi a quelli con maggiore copertura, sono a libero accesso. Esistono oltretutto indici specializzati, come *Meta* per la letteratura biomedica o *Semantic Scholar* per quella in informatica. La copertura dei dati di questi indici è oggetto di costanti analisi da parte della comunità di ricerca in bibliometria, tuttavia la maggior parte degli studi concordano nel sottolineare che i dati sono in costante miglioramento, sia in quantità che qualità, e specialmente per le pubblicazioni più recenti[11]. Cenerentola restano le discipline

[9] Nicola De Bellis, *Bibliometrics and citation analysis: from the Science Citation Index to cybermetrics*, Lanham (Md): Scarecrow Press, 2009, p. 38.
[10] Jordy Ardanuy, *Sixty years of citation analysis studies in the humanities (1951- 2010)*, «Journal of the American Society for Information Science and Technology», 64 (2013), n. 8, p. 1751-1755.
[11] Anne-Will Harzing – Satu Alakangas, *Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison*, «Scientometrics», 106 (2016), n. 2, p. 787-804.

umanistiche e le arti (*Arts and Humanities*) e, in parte, le scienze sociali, nonostante segnali di miglioramento anche in questo settore[12].
Come hanno sottolineato precedenti studi, la copertura delle discipline umanistiche negli indici citazionali non è sempre del tutto esaustiva per almeno due gruppi di fattori[13]:
– Ragioni intrinseche che dipendono da alcune caratteristiche della letteratura umanistica. Per esempio, l'uso di varie tipologie di pubblicazioni, specialmente monografie; la varietà di lingue; la diffusa pratica di pubblicare prevalentemente in cartaceo; l'ampia varietà di fonti utilizzate e del modo di citare[14].
– Ragioni estrinseche, legate invece agli ecosistemi informativi in cui l'estrazione delle citazioni avviene. Facciamo riferimento alla possibilità non soddisfacente di accedere in maniera aperta e centralizzata ai metadati necessari per identificare una fonte citata, a partire da identificativi permanenti (*PID: Permanent Identifiers*)[15].

Se gli ostacoli intrinseci derivano perlopiù dalla cultura scientifica umanistica, e vanno considerati come dati di fatto da risolvere con accorgimenti appositi, gli ostacoli estrinseci ricadono invece nell'area di responsabilità delle istituzioni che si occupano di curare questi servizi.
La conseguenza è evidente: se da un lato un indice citazionale per le discipline umanistiche può offrire accesso sistematico a collezioni documentarie e bibliotecarie in maniera integrata, dall'altro la sua messa in opera offre uno stimolo alle istituzioni culturali del settore per la messa a disposizione digitale, ed in forma programmaticamente interrogabile, dei propri metadati. Punto di partenza per il raggiungimento di questo obiettivo è la collaborazione su progetti specifici ed esplorativi, come *Linked Books*.

Creazione del corpus: selezione e digitalizzazione

Il punto di partenza del progetto è stata la domanda: in cosa consiste la storiografia su Venezia? La necessità di reperire pubblicazioni sulla storia di Venezia ricade in teoria nell'ambito della ricerca per soggetto, ma richiede in pratica un approccio

[12] Björn Hammarfelt, *Beyond coverage: Toward a bibliometrics for the humanities*, in: *Research assessment in the humanities*, a cura di M. Ochsner, S. E. Hug, H.-D. Daniel, Cham: Springer International Publishing, 2016, p. 115-131.
[13] Giovanni Colavizza – Matteo Romanello, *Citation mining of humanities journals: The progress to date and the challenges ahead*, «Journal of European Periodical Studies», in corso di pubblicazione.
[14] Björn Hellqvist, *Referencing in the humanities and its implications for citation analysis*, «Journal of the American Society for Information Science and Technology», 61 (2009), n. 2, p. 310-318; Anton Nederhof, *Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A review*, «Scientometrics», 66 (2006), n. 1, p. 81-100; Chris Alen Sula – Matthew Miller, *Citations, contexts, and humanistic discourse: Toward automatic extraction and classification*, «Literary and Linguistic Computing», 29 (2014), n. 3, p. 452-464.
[15] Cfr., ad esempio, il progetto europeo *Freya*: <https://www.project-freya.eu>.

pragmatico e differenziato. In questo contesto, la BAUM ha supportato adeguatamente le finalità del progetto, considerato che al suo interno raccoglie una collezione significativa di materiale bibliografico (fondi speciali, pubblicazioni, volumi di pregio) sulla storia di Venezia. La biblioteca ha consentito la conservazione e l'accessibilità di questi materiali, facilitate dall'adozione di una collocazione del materiale bibliografico prevalentemente a scaffale aperto, secondo un sistema classificatorio per materia, basato sulla Classificazione Decimale Dewey. I criteri per la selezione del corpus per il progetto sono stati: 1) la selezione di materiali principalmente dedicati a Venezia, ad esclusione della letteratura, peraltro molto abbondante, che tocca Venezia in maniera meno centrale; 2) la selezione di materiali tipicamente sottorappresentati negli indici commerciali, quali ad esempio monografie e lingue vernacolari (italiano, francese, ecc.); 3) la selezione di un volume di opere congruo ma non sovrabbondante rispetto alle risorse del progetto, anche in vista dell'espansione del corpus in fasi successive; ed infine, 4) la selezione di materiali perlopiù recenti (dal 1970 in poi).

Rispetto ai diritti d'autore, la soluzione trovata con l'accordo delle varie parti è stata quella di affidare ad EPFL la digitalizzazione delle opere su mandato delle istituzioni proprietarie delle singole copie utilizzate. L'istituzione proprietaria ha diritto a creare una copia digitale, per scopo di preservazione e di messa a disposizione dell'utenza all'interno dei propri locali fisici. Dalla copia digitale, EPFL ha creato una copia temporanea finalizzata al trattamento automatico dettagliato nel seguito, il cui obiettivo è l'estrazione di riferimenti bibliografici. Tali copie temporanee, distrutte al termine del trattamento, hanno permesso l'estrazione dei riferimenti bibliografici escludendo ogni altro contenuto sotto diritto d'autore.

Un primo gruppo di pubblicazioni individuato è stato prevalentemente di tipo monografico, in lingue nazionali che non fossero antecedenti al 1970. L'approccio euristico assunto nella progressiva costruzione del corpus si è affidato, nella fase iniziale, alle conoscenze dirette dei *subject librarians* in merito alla disponibilità di particolari fondi librari di studiosi della storia di Venezia le cui biblioteche personali erano state donate nel corso degli anni[16]. La conoscenza di questi fondi restituiva la certezza che le pubblicazioni estratte dal catalogo fossero coerenti con il tema del progetto. In questo modo si è potuta avviare la digitalizzazione per verificarne contestualmente gli esiti.

Successive estrazioni dal catalogo di Ateneo si sono focalizzate su una ricerca mirata per argomento, con lo scopo di recuperare risorse ordinate per classe, utilizzando la notazione numerica 945.31 (Storia di Venezia) in base al sistema di Classificazione Dewey. Il sistema classificatorio viene utilizzato in biblioteca anche per la collocazione del materiale bibliografico, un aspetto che ha consentito di

[16] In particolare, si fa riferimento ai fondi di Frederic C. Lane e di Gaetano Cozzi per un totale di 2.300 volumi. Le schede bio-bibliografiche sono accessibili ai seguenti indirizzi: <https://www.unive.it/pag/fileadmin/user_upload/SBA/documenti/BAUM/012_Fondo_Frederic_Lane.pdf> e <https://www.unive.it/pag/fileadmin/user_upload/SBA/documenti/BAUM/007_Fondo_Gaetano_Cozzi.pdf>.

operare un'ulteriore selezione direttamente a scaffale. Un approccio pragmatico che ha permesso di escludere dal corpus bibliografico, ad esempio, le pubblicazioni che presentavano nel testo un corredo di illustrazioni e immagini piuttosto significativo: libri coerenti con l'argomento ricercato, ma con una assente o ridotta lista di riferimenti.

Successive estrazioni si sono affidate ad una ricerca per soggetto utilizzando più descrittori; su tutti, "Venezia" e "Venezia-Storia". La ricerca per soggetto ha restituito un complesso di risorse piuttosto consistente, anche se più limitato rispetto alle opere disponibili nelle collezioni. Non tutte le notizie bibliografiche, infatti, sono corredate da accessi semantici e potevano quindi essere viziate dall'uso, in passato, di vocabolari non uniformi o controllati. D'altra parte, si deve considerare che il catalogo è frutto di un processo di progressiva stratificazione che nel tempo ha sovrapposto norme e persone e può quindi soffrire di una mancanza di uniformità nella qualità delle descrizioni bibliografiche. Nell'espansione del corpus bibliografico sono state inoltre comprese pubblicazioni periodiche focalizzate sulle tematiche della storia veneziana che hanno richiesto la collaborazione con altre istituzioni veneziane, in particolare l'Archivio di Stato di Venezia e la Biblioteca Nazionale Marciana, per la ricostruzione della serie completa delle annate[17].

Le liste di titoli elaborate dai *subject librarians* nel corso del tempo hanno prodotto complessivamente più di 22.000 risultati che sono stati oggetto di una valutazione e di una deduplicazione, prima di essere sottoposti alla digitalizzazione per l'estrazione delle citazioni. La digitalizzazione delle opere è stata messa in atto utilizzando uno scanner fornito da EPFL e personale specializzato, avendo cura di associare ad ogni volume digitalizzato i metadati disponibili dal polo SBN di Venezia. La taglia finale del corpus, dopo una successiva campagna di espansione, è di 2.475 volumi, di cui 1.905 libri, perlopiù monografie, e 552 numeri di rivista, da cui 5.496 articoli sono stati estratti[18].

Creazione del corpus: citazioni e uso dei dati del catalogo SBN

La creazione di un corpus di citazioni bibliografiche è un processo che consiste di due fasi distinte: in primo luogo l'estrazione (*citation extraction*) e, in secondo luogo, la loro disambiguazione *(citation matching)*. L'estrazione richiede di classificare ciascuna parola del testo di partenza come facente parte di una citazione oppure no e se sì, di determinare il ruolo che essa ricopre al suo interno (ad es. in-

[17] Si tratta in particolare di pubblicazioni periodiche: Archivio Veneto, Studi Veneziani, gli Atti dell'Istituto Veneto di Scienze, Lettere ed Arti e Ateneo Veneto.

[18] Per un approfondimento riguardo alla selezione del corpus e sua coerenza rispetto alle citazioni da esso estratte in seguito, rimandiamo il lettore a Giovanni Colavizza – Matteo Romanello – Frédéric Kaplan, *The references of references: A method to enrich humanities library catalogs with citation data*, «International Journal on Digital Libraries», 19 (2017), n. 2-3, p. 151-161.

dicazione dell'autore, titolo, anno di pubblicazione, ecc.). La disambiguazione, invece, consiste nello stabilire, per ogni citazione così estratta, un legame tra la citazione stessa ed un record bibliografico in modo da poter ricondurre due o più citazioni della stessa pubblicazione al medesimo oggetto bibliografico citato[19].

Risorse esistenti di metadati bibliografici, come il Catalogo del Servizio Bibliotecario Nazionale (SBN), rivestono un ruolo fondamentale per la disambiguazione di citazioni in quanto contengono già delle descrizioni dettagliate delle pubblicazioni che troviamo poi citate nel nostro corpus sulla storia di Venezia. Tali descrizioni sono estremamente preziose per qualsiasi algoritmo che cerchi di stabilire una corrispondenza tra citazione e pubblicazione citata.

È in questo frangente che il progetto *Linked Books* ha visto la collaborazione tra EPFL e ICCU al fine di migliorare la fruibilità dei *record* bibliografici contenuti in SBN. Tale collaborazione ha permesso di risolvere alcuni ostacoli esistenti nell'utilizzo di SBN, dovuti principalmente alle modalità di accesso ai dati – soprattutto vista la loro mole imponente – e al formato dei dati stessi, permettendo di accedere facilmente ai metadati contenuti in ciascun *record*.

La prima opzione considerata è stata l'utilizzo delle API (*Application Programming Interface*) implementate per l'app "OPAC SBN"[20], realizzata per l'ICCU da Inera Srl[21]. Fra le altre funzionalità messe a disposizione da queste API c'è l'estrazione di un record in formato JSON dato il suo BID, l'identificatore univoco che viene assegnato ad ogni notizia bibliografica in SBN. JSON è un formato di dati testuale aperto, molto diffuso in ambito web per lo scambio di dati tra applicativi. Il formato JSON è ben noto alla maggioranza degli sviluppatori, a differenza di UNIMARC che, pur estremamente ricco, è comunque poco diffuso al di fuori dell'ambito biblioteconomico. L'*output* in JSON sarebbe stato facilmente utilizzabile per gli scopi del progetto, ma questo approccio presenta evidenti criticità.

La prima riguarda le prestazioni: una prova di *download* di 10.000 BID di tipologia varia richiede circa 45 minuti. Pertanto, per scaricare gli oltre 16 milioni di record attualmente in SBN, sarebbero stati necessari almeno cinquanta giorni, senza interruzioni. Oltre ad essere incompatibile con le tempistiche del progetto, questo approccio avrebbe avuto un serio impatto sul sistema OPAC, rischiando di degradare le sue prestazioni. Una seconda criticità è la mancanza, nell'*output* prodotto dalle API, di alcune informazioni importanti come soggetti, classificazioni Dewey, altri titoli e relazioni gerarchiche.

Per risolvere queste criticità, è stato realizzato un programma che, partendo dai

[19] Usiamo il termine "citazione" per indicare sia la porzione di testo contenente un riferimento bibliografico, che il legame tra una pubblicazione (citante) ed una fonte (citata). Il contesto chiarisce l'accezione usata.

[20] <https://play.google.com/store/apps/details?id=it.inera.opacmobile&hl=it> e <https://itunes.apple.com/it/app/opac-sbn/id786155255>.

[21] https://www.inera.it.

record UNIMARC estratti periodicamente dall'Indice SBN e destinati all'aggiornamento del Catalogo online, producesse record in un formato JSON ispirato a quello prodotto dalle API per l'applicazione mobile, con l'aggiunta delle informazioni mancanti ed eventuali altre migliorie. Le prime prove, effettuate su un campione di 100.000 record, hanno dato risultati molto confortanti anche dal punto di vista delle prestazioni: eseguiti su un normale *personal computer*, hanno richiesto meno di dieci minuti. Questo approccio, basato su un'elaborazione offline, evita di sovraccaricare il Catalogo online da un carico di lavoro eccessivo[22].

Dopo una prova effettuata sull'intero scarico UNIMARC di oltre 16 milioni di record, durato solo qualche ora, è cominciata la fase di perfezionamento dell'output per renderlo più adatto alle esigenze del sistema di estrazione delle citazioni. Rispetto all'output delle API mobile sono state aggiunte, ad esempio, le classificazioni Dewey, i soggetti e altri titoli. Per quanto riguarda le localizzazioni, invece, si è preferito esportare solo il codice ISIL[23], omettendo le denominazioni per esteso delle biblioteche e rimandando agli *open data* disponibili sul sito dell'Anagrafe delle biblioteche italiane per la decodifica dei codici ISIL.

Lo scarico dei dati SBN in formato JSON viene poi indicizzato in un'istanza locale di *Elastic Search*, un sistema *open source* di indicizzazione e ricerca di dati testuali, liberamente scalabile ed altamente performante. Questa soluzione permette di cercare nei metadati dei record SBN, permettendo così di usarli per disambiguare i riferimenti bibliografici estratti. Tuttavia, il limite principale di questa soluzione, che d'altro canto è inevitabile vista l'attuale infrastruttura tecnica di SBN, è la perdita di sincronizzazione con i dati in SBN[24]. Tali dati, infatti, sono tutt'altro che statici: nuove acquisizioni producono nuovi record, ed altri scompaiono in seguito alla rimozione di duplicati. Una soluzione ottimale dovrebbe permettere di usare SBN e altri sistemi analoghi come una corrente continua di dati (dinamica e costantemente in divenire), piuttosto che come uno scarico statico e destinato a diventare obsoleto.

L'estrazione di citazioni, nel progetto *Linked Books*, è avvenuta in due fasi successive. Nella prima fase citazioni a fonti primarie e secondarie sono state estratte dal corpus di pubblicazioni tramite un modello statistico che fa uso di citazioni manualmente estratte ed annotate. Tale modello consente di determinare, con un certo livello di accuratezza, l'inizio e la fine di ciascun riferimento, così come di classificare le informazioni che compongono una citazione (tipicamente nome degli autori, titolo, anno di pubblicazione, editore, ecc.). Questo modello ha poi per-

[22] Il programma è stato realizzato in Java 8, utilizzando le librerie MARC4J, JSONSimple e GSON, tutte distribuite sotto licenze aperte (LGPL 2.1 e Apache 2).

[23] https://anagrafe.iccu.sbn.it/it/informazioni/cosa-e-lisil.

[24] A questo riguardo, si segnala l'interessante proposta di utilizzare il software aperto Wikibase per la pubblicazione linked open data di dati UNIMARC, avanzata da Giovanni Bergamin – Cristian Bacchi, *New ways of creating and sharing bibliographic information: an experiment of using the Wikibase Data Model for UNIMARC data*, «JLIS.it» 9 (2018), p. 35-74.

messo di estrarre 3.850.581 citazioni dal corpus considerato. La seconda fase, invece, ha permesso di collegare una citazione ad una autorità bibliografica o archivistica, e più precisamente SiASVe per i riferimenti a fonti di archivio ed i dati di ICCU e del *Virtual International Authority File* (VIAF) per le fonti secondarie[25]. Le citazioni così estratte possono essere esplorate dagli utenti utilizzando due applicazioni di ricerca, discusse nella sezione che segue.

Design e sviluppo di interfacce di ricerca ed esplorazione: Venice Scholar e Venice Scholar – Library

Due applicazioni per la gestione e ricerca dei dati sono state sviluppate nel corso del progetto *Linked Books*: la *Venice Scholar Library (VSL)*[26] e il *Venice Scholar Index (VSI)*[27]. Si tratta, nel primo caso, di un applicativo per la biblioteca digitale (*digital library software)* che permette agli utenti di gestire ed esplorare i materiali digitalizzati. Nel secondo caso invece, del vero e proprio indice citazionale che permette agli utenti di esplorare le pubblicazioni sulla storia di Venezia seguendo i collegamenti tra di esse stabiliti dalle citazioni bibliografiche e di accedere ai dati disponibili per ciascuna pubblicazione (ad esempio, numero di citazioni date e ricevute).

Il design e sviluppo di entrambe le interfacce si è svolto seguendo un processo iterativo. In una prima fase di concezione e design, il team del progetto ha lavorato a stretto contatto con un designer per far sì che le interfacce fossero il più possibile intuitive ed ergonomiche. Terminato il design si è passati all'implementazione delle interfacce sotto forma di prototipi, che sono stati poi sottoposti ad una fase di *beta testing* al fine di ottenere dei commenti che potessero migliorarne l'usabilità.

Entrambe le interfacce sono state sottoposte ad un periodo di *beta testing* da parte di due gruppi di utenti. Un primo gruppo di cinque bibliotecari della biblioteca BAUM dell'Università Ca' Foscari di Venezia, ed un secondo gruppo composto da storici, archivisti e bibliotecari provenienti dalle altre istituzioni parte del progetto. Questa fase di prova è risultata di grande utilità per migliorare l'usabilità delle interfacce in quanto ha permesso di individuare elementi che erano fonte di confusione per l'utente (ad esempio, denominazioni dei pulsanti) o funzionalità che si sono rivelate insoddisfacenti, e pertanto sono state migliorate (come la precisione della ricerca testuale).

Passiamo ora ad esaminare le funzionalità principali di ciascuna interfaccia. La VSL offre una serie di funzionalità che sono tipiche di qualsiasi *digital library*, affiancate

[25] Per maggiori dettagli sui processi automatici di estrazione e disambiguazione e sulla loro accuratezza si veda Giovanni Colavizza – Matteo Romanello – Frédéric Kaplan, *The references of references: A method to enrich humanities library catalogs with citation data*, «International Journal on Digital Libraries», 19 (2017), n. 2-3, p. 156-158.

[26] https://veniceschol ar.dhlab.epfl.ch/library.

[27] https://veniceschol ar.dhlab.epfl.ch.

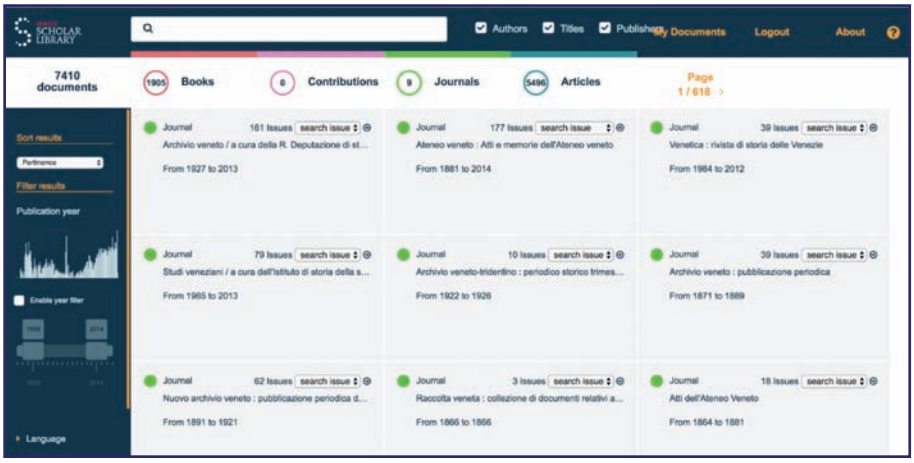

Figura 1. *Venice Scholar LIbrary: visualizzazione dei risultati di ricerca sui metadati.*

Figura 2. *Venice Scholar Library: modalità di lettura del materiale digizalizzato con immagine digitale ed OCR visualizzati una accanto all'altro.*

da alcune funzionalità avanzate che permettono agli utenti amministratori di modificare manualmente le citazioni estratte automaticamente dalla VSL. Nella prima tipologia di funzionalità rientrano la visualizzazione e ricerca sui metadati del digitalizzato (Fig. 1), una modalità di lettura dove immagine digitale e testo prodotto dal riconoscimento automatico dei caratteri (OCR: *Optical Character Recognition*) sono visualizzati una accanto all'altro (Fig. 2) e la possibilità di scorrere rapidamente attraverso le miniature delle pagine digitalizzate. Per quanto riguarda le funzionalità avanzate, invece, la VSL offre agli utenti una modalità di lettura del testo dove le citazioni bibliografiche estratte automaticamente possono essere visualizzate ed eventualmente modificate (nel caso di utenti amministratori) (Fig. 3).

Sebbene nel corso del progetto *Linked Books* non ci sia stata alcuna attività di correzione manuale delle citazioni estratte, avere uno strumento che consente di

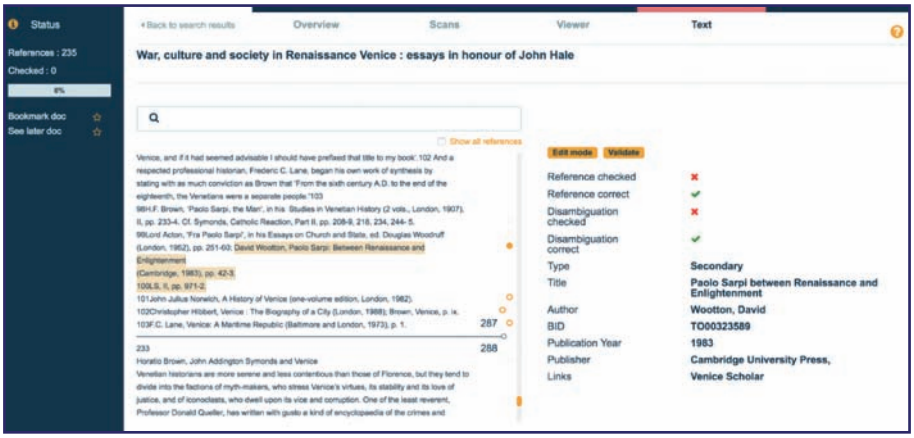

Figura 3. *Venice Scholar Library: modalità di lettura per utenti amministratori con funzionalità di correzione dei riferimenti bibliografici.*

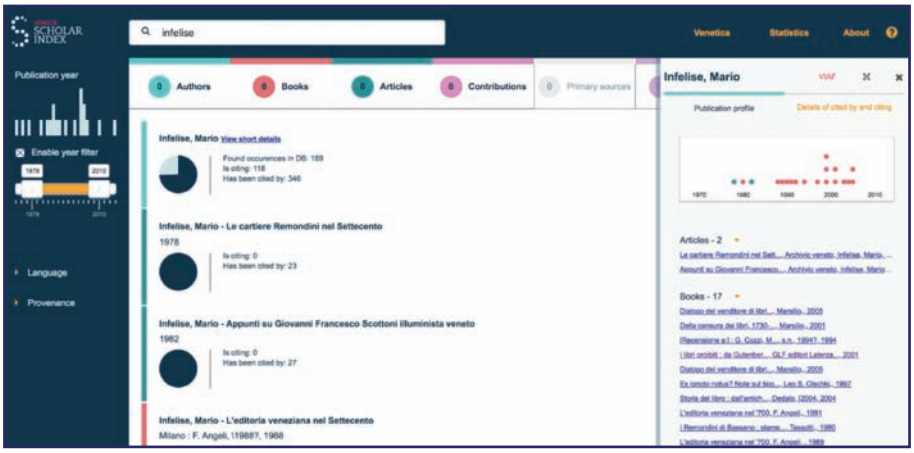

Figura 4. *Venice Scholar Index: pagina dei risultati di ricerca.*

farlo apre la strada ad un possibile coinvolgimento attivo di bibliotecari e archivisti non solo nell'attività di digitalizzazione ma anche in quella di miglioramento e correzione dei dati citazionali ricavati dalle pubblicazioni digitalizzate[28]. Tale correzione manuale è necessaria ed auspicabile al fine di garantire la maggiore accuratezza possibile delle analisi quantitative che un indice citazionale rende possibili.
L'indice di citazioni vero e proprio può essere esplorato attraverso una seconda interfaccia, il *Venice Scholar Index*. A differenza della VSL, l'accesso al VSI è intera-

[28] Il coinvolgimento dei bibliotecari nel miglioramento dei dati citazionali è al centro di un altro progetto di ricerca, il *Linked Open Citation Database*, sul quale si veda Anne Lauscher – Kai Eckert – Lukas Galke – Ansgar Scherp – Syed Tahseen Raza Rizvi – Sheraz Ahmed – Andreas Dengel – Philipp Zumstein – Annette Klein, *Linked Open Citation Database*, in: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, New York, ACM Press, 2018.

Figura 5. *Venice Scholar Index: profilo citazionale dell'autrice Patricia Fortini Brown. Le cifre sul numero di citazioni fatte o ricevute sono cumulative rispetto a tutte le pubblicazioni dell'autrice che sono state indicizzate.*

mente pubblico in quanto i dati che contiene e su cui si basa (soprattutto metadati bibliografici) non sono soggetti a diritto d'autore.

La prima pagina che accoglie l'utente presenta la consueta maschera di ricerca attraverso la quale è possibile effettuare una ricerca su a) nomi di autori, b) titoli di pubblicazioni citate (articoli o monografie), c) nomi di fonti d'archivio (fonti primarie) e d) sul testo stesso delle citazioni estratte. I risultati della ricerca vengono poi mostrati suddivisi per categoria e con la possibilità di filtrarli ulteriormente sulla base di criteri come anno di pubblicazione, lingua e biblioteca o archivio di provenienza del materiale digitalizzato (Fig. 4).

Ogni autore citato all'interno del corpus possiede un profilo delle pubblicazioni ed uno delle citazioni. Le pubblicazioni includono titoli che sono stati digitalizzati nel corso del progetto (e sono pertanto disponibili per la lettura attraverso la VSL), così come titoli non digitalizzati ma citati da altre pubblicazioni. Il profilo delle citazioni, disponibile sia per autori che singole pubblicazioni, presenta da un lato una visualizzazione della frequenza delle citazioni (sia in entrata che in uscita, cioè fatte e ricevute) e dall'altro offre la possibilità di esaminare la lista di pubblicazioni citate/citanti (Fig. 5).

Un aspetto che distingue il VSI da altri indici citazionali esistenti, come *Google Scholar*, è il fatto di indicizzare le citazioni a fonti primarie, e nello specifico i documenti archivistici. Il profilo citazionale di ogni autore contiene una sezione dedicata alle fonti primarie dove è possibile esplorare quali fondi, serie o sottoserie d'archivio sono stati citati da un certo autore nella totalità delle sue pubblicazioni (Fig. 6).

Un altro punto di accesso alle fonti primarie nell'interfaccia del VSI è fornito dal profilo citazionale disponibile per ogni fonte primaria citata all'interno del corpus.

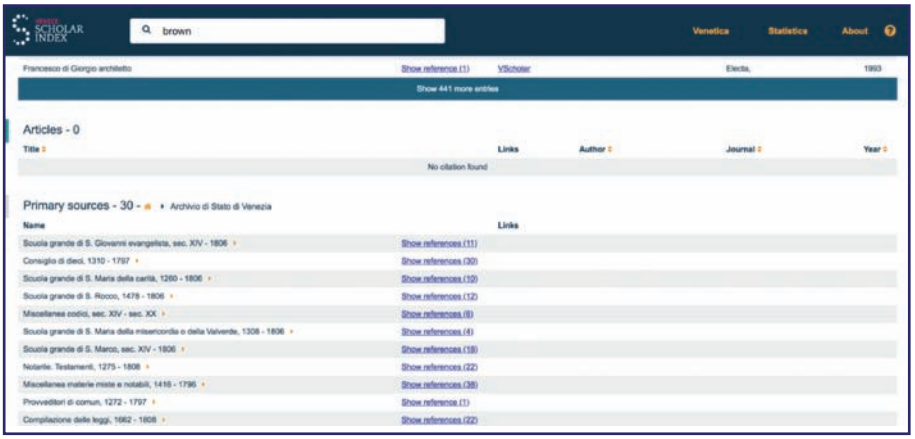

Figura 6. *Venice Scholar Index: fonti primarie citate dall'autrice Patricia Fortini Brown.*

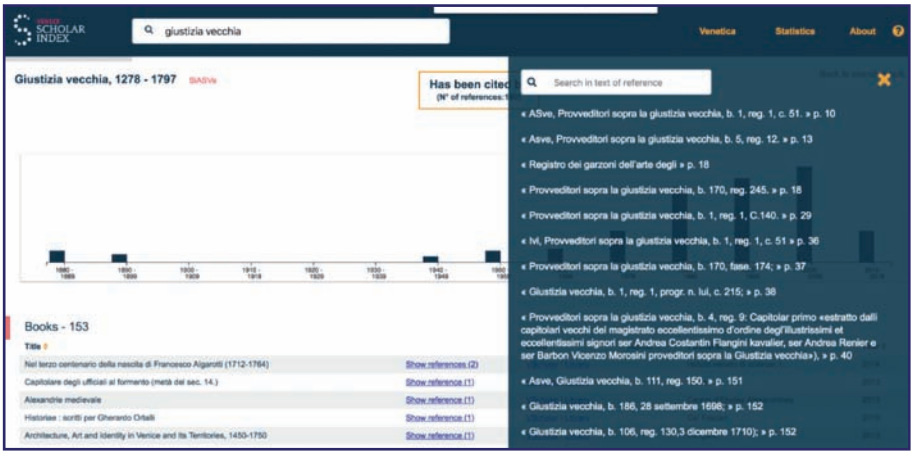

Figura 7. *Venice Scholar Index: profilo citazionale del fondo "Giustizia Vecchia" dell'Archivio di Stato di Venezia.*

Per ciascuna fonte il VSI fornisce a) una *timeline* che mostra la frequenza delle citazioni ricevute per decennio (Fig. 7) e b) una lista dettagliata delle pubblicazioni citanti, suddivise per tipologia (articolo di rivista, monografie, ecc.).
A concludere questa carrellata, una funzionalità sperimentale che è stato possibile sviluppare ed aggiungere al VSI grazie ad un finanziamento messo a disposizione da *Europeana Research* nel quadro del suo *Research Grants Programme 2018*[29]. Per ogni risorsa disponibile nel VSI (autore moderno, fonte secondaria, fonte primaria), nella pagina ad essa dedicata, un pulsante con il logo di Europeana consente di accedere ad un menù laterale con oggetti rilevanti a tale risorsa e provenienti da Europeana (Fig. 8). Tale lista di risorse è creata dinamicamente aggre-

[29] https://pro.europeana.eu/post/developing-the-venice-scholar-matteo-romanello-europeana-research-grant-winner.

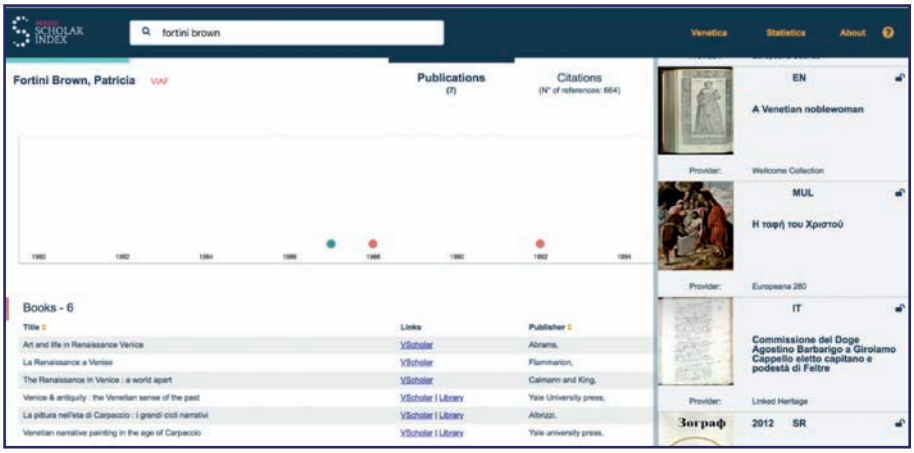

Figura 8. *Venice scholar index: Europeana sidebar.*

gando e filtrando dati provenienti dalla *Europeana Search API* [30]. Tra i milioni di oggetti digitali contenuti nel meta-aggregatore, un sistema basato sulle parole chiave estratte dai titoli delle pubblicazioni seleziona gli oggetti potenzialmente più rilevanti per l'utente. Tale funzionalità ci ha permesso, da un lato, di mostrare come strumenti diversi possano venire integrati insieme in modo produttivo e, dall'altro, di creare una funzionalità capace di amplificare l'aspetto di serendipità[31] inerente ad ogni ricerca effettuata attraverso un motore di ricerca.

Conclusione

Il progetto *Linked Books* costituisce un'esplorazione della fattibilità e delle potenzialità degli indici citazionali nelle discipline umanistiche. Pensato con l'obiettivo di collegare cataloghi bibliotecari, sistemi informativi archivistici e rispettive piattaforme digitali per il tramite della storiografia, il progetto ha permesso di porre ed affrontare una serie di questioni chiave rispetto al futuro dell'indicizzazione di risorse bibliografiche e documentarie in un futuro sempre più digitale. Prima di tutto, la definizione di un corpus rappresentativo, seguita da digitalizzazione e analisi in un contesto legale appropriato. Cruciale, in questa fase, non solo la presenza di collezioni bibliografiche adeguate, ma anche di competenze pluridisciplinari (esp. *subject librarians*). A seguire, l'uso di metodi automatizzati per estrarre citazioni dalla letteratura, che richiede la disponibilità di un rapido accesso ai cataloghi o sistemi informativi di riferimento per quanto riguarda identificativi univoci e metadati (e.g. SBN). Per concludere, lo sviluppo di prototipi di applicativi: una biblioteca digitale (*Venice Scholar – Library*) e un indice citazionale (*Venice Scholar*).

[30] https://pro.europeana.eu/resources/apis/search.

[31] Wikipedia s.v. serendipità: «Il termine serendipità indica la fortuna di fare felici scoperte per puro caso e, anche, il trovare una cosa non cercata e imprevista mentre se ne stava cercando un'altra".

Merita di essere sottolineato quello che è forse ad oggi l'ostacolo principale per progetti come *Linked Books*: la possibilità (o sua assenza) di accedere programmaticamente (per esempio tramite API aperte) a metadati archivistici e bibliotecari strutturati in maniera uniforme e secondo stabiliti standard internazionali. Idealmente, sotto forma di *linked data*. Costruire dei servizi che si basino su questi metadati e sulla loro aggregazione sarà sempre difficoltoso, se non irrealizzabile, senza un maggiore sforzo istituzionale rivolto non solo alla cura di metadati e alla loro pubblicazione tramite interfacce dedicate, ma anche alla loro accessibilità per un uso programmatico/automatico. Anche per queste ragioni, l'ICCU sta avviando un progetto denominato Sistema di Ricerca Integrato (SRI) che porterà alla realizzazione di un accesso *linked open data* ai dati del catalogo SBN.
Il progetto *Linked Books* ha infine permesso di concepire un progetto di più ampio respiro, lo *Scholar Index*: un indice citazionale aperto, collaborativo e distribuito[32]. Seguendo la tradizione dei cataloghi bibliotecari, questo progetto propone di affidare alla responsabilità di ciascuna istituzione culturale facente parte del progetto, una parte relativamente piccola della letteratura da indicizzare. Compito dell'istituzione è quindi quello di digitalizzare le pubblicazioni di sua competenza e, con l'ausilio dell'automazione permessa dagli applicativi sviluppati, curare l'estrazione delle citazioni. Questi dati citazionali possono poi essere aggregati, ad esempio in infrastruttura esistente come *Open Citations*[33], e messi a disposizione dell'utenza tramite interfacce o come *linked data*. Riteniamo che il futuro dell'indicizzazione digitale di materiali documentari sarà marcato dalla connettività, e specificamente dal superamento di confini informativi istituzionali. I legami citazionali offrono in questo senso un'opportunità ancora del tutto inesplorata.

> *We present the outcomes of the Linked Books project, resulting in a prototype citation index interlinking the Italian national library catalog (Opac SBN) with the information system of the State Archive of Venice and international authority records or "metaengines" such as VIAF.org and Europeana. Our prototype includes 3.850.581 citations extracted from a corpus of 2.475 volumes, of which 1.905 monographs, and 552 journal volumes, or 5.496 articles therein. The corpus is focused on the history of Venice. The Linked Books project allowed us to explore the feasibility and desirability of a citation index for the humanities, and to face and solve technical challenges including: the selection of a thematically representative corpus from bibliographic resources and expertise, the digitization of these materials within the bounds of copyright, the automatic extraction of citations and the development of public search interfaces.*

[32] <https://scholarindex.eu>, il progetto è in discussione come potenzialmente facente parte del progetto europeo Time Machine (<http://timemachine.eu>).

[33] <http://opencitations.net>. Silvio Peroni – Alexander Dutton – Tanya Gray – David Shotton, *Setting Our Bibliographic References Free: Towards Open Citation Data*, «Journal of Documentation», 71 (2015), n. 2, p. 253-77.

L'ultima consultazione dei siti web è avvenuta nel mese di giugno 2019