# Shared Entity Management Infrastructure di OCLC

## Analisi retrospettiva e sintesi del progetto

**MICHAEL PHILLIPS**
OCLC
phillipm@oclc.org

**ANNE WASHINGTON**
OCLC
washinga@oclc.org

DOI: 10.3302/0392-8586-202206-044-1

## 1. OCLC Research e i linked data

### *1.1. Introduzione*

Per le biblioteche e gli altri enti culturali i linked data significano funzioni di ricerca ottimizzate, inaspettate opportunità di scoperta e una migliore interoperabilità informatica. OCLC, essendo uno dei principali fornitori di metadati bibliografici, è attivamente coinvolto nelle ricerche in materia di linked data nel tentativo di rispondere alle esigenze della prossima generazione di consumatori di dati. Attingendo alla propria pluriennale esperienza nel campo dei linked data, OCLC ha recentemente portato a termine il progetto shared entity management infrastructure per la realizzazione di un'infrastruttura di gestione condivisa delle entità, nella quale i dati bibliografici e gli authority data dei tradizionali record sono stati convertiti in entità interrelate tra loro tramite una serie di legami semantici. La conclusione del progetto segna una tappa significativa per lo sviluppo di un'infrastruttura di linked data solida, permanente ed estensibile.

### *1.2 Retrospettiva*

I passati progetti di ricerca di OCLC sui linked data hanno costituito le fondamenta su cui l'attuale progetto ha preso piede. Nel 2009 OCLC ha sviluppato il VIAF (Virtual International Authority File) e le tecnologie FAST (Faceted Application of Subject Terminology), atte a fornire record di autorità in versione linked data. Nel 2013 il progetto EntityJS ha evidenziato come diversi criteri di visualizzazione e navigazione dei linked data migliorassero le funzionalità di scoperta nella rete di relazioni. Nel 2014 il progetto pilota Person Entity Lookup ha portato alla realizzazione di una API (interfaccia di programmazione delle applicazioni) di accesso e interoperabile, ampliando la comprensione dell'applicazione dei linked data come servizio web. Tra il 2014 e il 2015 il progetto CONTENTdm® Metadata Refinery ha sviluppato una serie di strumenti per aiutare gli istituti culturali nelle fasi di creazione dei linked data, rafforzando i collegamenti tra i record di metadati prima della loro conversione in linked data. Nel 2017-2018 il Project Passage di OCLC ha segnato la creazione del primo, vasto reticolo di linked data per risorse bibliografiche. Per questo progetto OCLC ha utilizzato un pacchetto di strumenti forniti da un'istanza Wikibase, mentre per la valutazione dei risultati si è avvalso della consulenza degli esponenti della comunità bibliotecaria. Tra il 2019 e il 2020, con il CONTENTdm Linked Data Pilot, sono proseguite le indagini sull'utilizzo di Wikibase come ambiente di lavoro con i linked data. Contemporaneamente è stato implementato lo standard IIIF (International Image Interoperability Framework) per la descrizione delle risorse digitali legate al patrimonio culturale.
Da un progetto all'altro il grado di complessità della ri-

cerca è andato sempre crescendo, e questo ha migliorato la comprensione dei linked data e delle loro possibili applicazioni, ha messo in luce la necessità di identificatori univoci e persistenti come base per i sistemi di linked data e ha confermato la possibilità che sia proprio OCLC a realizzare questa struttura. Inoltre, i passati progetti hanno permesso la creazione e lo sviluppo di un pacchetto di strumenti pensati appositamente per la creazione e il miglioramento delle entità linked data, strumenti che possano, appunto, gestire la prammatica importazione e conversione dei record tradizionali in entità linked data così come la creazione *ex novo* delle entità da parte degli utenti tramite l'apposita interfaccia utente. Questo approccio su ambo i fronti consentirà di far confluire in un unico nuovo ambiente digitale sia i vecchi record convertiti sia le entità create *ex novo*, permettendo una fluida transizione tra passato e futuro.



### *1.3 Shared Entity Management Infrastructure: obiettivi*

Tramite un finanziamento della Mellon Foundation, OCLC ha dato avvio a un progetto per un'infrastruttura di gestione condivisa delle entità con l'obiettivo di sviluppare un sistema per la produzione su larga scala di linked data. Per realizzare questo scopo sono stati identificati tre obiettivi.

1. Creare una robusta "colonna vertebrale di entità" che presentasse la descrizione di *Opere* e *Persone*, completa di link alle loro rappresentazioni in vocabolari e authority file esterni.
2. Creare gli strumenti di indicizzazione e di ricerca necessari per coprire un'ampia gamma di casi d'uso.
3. Sviluppare strumenti che permettessero alla comunità bibliotecaria di arricchire la descrizione delle suddette entità.

Quest'ultimo punto si era già dimostrato di vitale importanza nelle precedenti esperienze di ricerca di OCLC. Senza un'interfaccia capace di incoraggiare gli operatori di biblioteca a creare e modificare le entità – e dunque a migliorare la qualità dei dati tramite la partecipazione attiva della comunità di riferimento – il progetto non avrebbe infatti avuto esito positivo.

## 2. Descrizione e design del progetto

### *2.1. Design del progetto*

Il team di progettazione ha coinvolto tutti i diversi reparti di OCLC, inclusi il settore di gestione di progetto, strategia del prodotto, di user experience e design, qualità dei dati, ricerca, programmi comunitari, comunicazione e marketing, architettura del sistema, ingegneria dei dati, ingegneria del software, garanzia della qualità e fornitura del software. Durante i due anni di finanziamento, la squadra così composta ha lavorato all'unisono per raggiungere i seguenti obiettivi con traguardi semestrali:

- Prima tappa (luglio 2020): selezione dei dati, iniziale caricamento e indicizzazione su Wikibase; messa a punto delle operazioni di ricerca e lettura.
- Seconda tappa (gennaio 2021): incremento della banca dati fino a 100 milioni di entità; messa a punto delle funzioni di creazione, lettura, modifica ed eliminazione delle entità tramite interfaccia utente e API.
- Terza tappa (luglio 2021): spostamento sull'infrastruttura di OCLC; definizione e implementazione dei parametri qualitativi per la creazione delle entità; messa a punto degli endpoint SPARQL e di ricerca tramite le API; estensione delle dimensioni dell'agglomerato di entità oltre le 100 milioni di unità; traduzione delle interfacce; creazione di un modello di sviluppo sostenibile.
- Quarta tappa (dicembre 2021): perfezionamento e correzione dei processi e dell'infrastruttura per il lancio dei prodotti e dei servizi nel rispetto delle tempistiche dovute.

Un'altra collaborazione di fondamentale importanza è stata quella con il gruppo di studio per la gestione delle entità, formato dai rappresentanti di ventisette biblioteche di sette paesi diversi. Il campione selezionato voleva essere rappresentativo delle varie tipologie di organizzazione (biblioteche pubbliche e accademiche,

musei e archivi), tenendo conto anche della dimensione e del luogo. Relazionandosi con questo gruppo di studio, OCLC ha potuto stabilire i criteri necessari per rispondere alle esigenze di un'utenza variegata, ottenendo feedback sui dati e sull'interfaccia utente, nonché sul flusso di lavoro a cui questo servizio avrebbe dovuto arrecare beneficio. Gli aderenti al gruppo di studio hanno partecipato a incontri mensili su molteplici argomenti, a workshop occasionali e sessioni di focus group, effettuando test sui dati, sulle interfacce utente e sulle API.

## *2.2. Resoconto delle tappe semestrali*

### 2.2.1 Primo semestre

La priorità durante questa fase è stata quella di assicurare che l'infrastruttura tecnica funzionasse correttamente. All'epoca, furono caricate all'incirca 15 milioni di entità all'interno di un ambiente Wikibase locale e furono attivate le operazioni di lettura e ricerca semplice. Il gruppo di studio ha verificato le funzioni di login e ha effettuato delle ricerche elementari usando i nomi e gli identificatori delle entità. I tester hanno trovato facile svolgere queste mansioni tramite l'interfaccia utente fornita da Wikibase, considerarono tuttavia incompleti i dati delle entità. Questa iniziale valutazione, relativa alla qualità dei dati, ha rivelato la necessità di aumentare sia la mole sia la ricchezza dei dati delle entità.

### 2.2.2 Secondo semestre

Uno dei principali obiettivi per il primo anno di ricerca è stato quello di confermare l'estensibilità dell'infrastruttura Wikibase; per questo fu deciso di dare priorità alla crescita quantitativa delle entità piuttosto che concentrarsi sul miglioramento qualitativo dei dati. Prefiggendosi l'obiettivo di raggiungere i 100 milioni di entità, la squadra di progettazione ha raggiunto 94 milioni di unità.
Il team si è trovato davanti a un ostacolo quando ha tentato di caricare una grande quantità di dati su Wikibase. Diversamente dal primo semestre, i processi di caricamento duravano, infatti, settimane invece di giorni. Identificare e correggere gli errori di caricamento e i rallentamenti diventò, quindi, la priorità assoluta; per evitare ritardi sulla consegna del progetto gli errori di caricamento dovevano essere risolti immediatamente. Concentrandosi sulla velocità e la scalabilità del caricamento dati, ci furono meno occasioni per prestare attenzione alla ricchezza e alla qualità dei dati delle entità. Allo scadere dei 12 mesi, l'*équipe* fu in grado di caricare all'incirca 94 milioni di entità, pur nella consapevolezza che la qualità dei dati sarebbe comunque risultata insufficiente.
Una specifica lacuna nei dati ebbe un significativo impatto sulla valutazione effettuata dal gruppo di studio in questa fase. Ai membri fu chiesto di concentrarsi sui dati delle entità e sulla visualizzazione della pagina delle informazioni. I risultati delle ricerche in Wikibase includono, infatti, un'etichetta per le entità, un identificatore univoco e la relativa descrizione. Ad esempio, un risultato di ricerca per "Toni Morrison" mostrerebbe un'etichetta "Toni Morrison", l'ID "Q1234", e una sommaria descrizione, "romanziere americano, saggista e accademico (1931-2019)". Essendo la priorità quella di assicurare dei tempi di caricamento ottimizzati, la squadra non ebbe il tempo di creare le descrizioni per molte delle entità. Questo comportò che nei risultati di una ricerca, gli utenti riscontrassero una lista di etichette per le entità, senza ulteriori descrizioni. Tali entità non potevano essere disambiguate se non tramite un processo manuale di selezione e poi comparando i dettagli su altre pagine. I tester manifestarono frustrazione verso questo limite, specialmente quando si trovarono a lavorare con entità che condividevano parole o titoli comunemente usati o persone con nomi relativamente comuni.
Ai collaudatori venne, inoltre, richiesto di valutare le informazioni che apparivano nelle pagine di descrizione delle entità: erano sufficienti per disambiguare un'entità da un'altra? Le informazioni erano accurate? I dati che si aspettavano di trovare erano già tutti presenti? Tutto sommato la valutazione fu negativa, le informazioni presenti nelle descrizioni delle entità non rispondevano a dei criteri descrittivi standard e ciò comportò l'inclusione di elementi aggiuntivi per la descrizione. OCLC fece tesoro dei loro *feedback* realizzando una serie di "minimi criteri per la valida descrizione delle entità", e una scala di parametri valutativi per assicurare la qualità dei dati (vedi paragrafo 3.4).

### 2.2.3 Terzo semestre

A seguito della revisione del gruppo di studio, la squadra mutò il proprio approccio verso il caricamento dati durante la terza fase, puntando a migliorarne la qualità. Solo una porzione delle entità – circa 36 milioni – furono selezionate, giacché i loro dati erano sufficientemente ricchi per fornire una base per generare informazioni descrittive.
Allo stesso tempo, il team fu traslato dall'ambiente Wikibase verso un'altra infrastruttura tecnica. Wikiba-

se aveva permesso alla squadra di familiarizzare con la struttura e la gestione dei linked data, ma fin dal principio non fu considerato come una soluzione a lungo termine per l'infrastruttura pensata da OCLC. I problemi riscontrati nel caricamento di ingenti quantità di dati, così come la manutenzione tecnica richiesta da Wikibase, dissuasero dalla possibilità di considerarlo come l'ambiente definitivo. Il team sviluppò un'infrastruttura locale, con un sistema di indicizzazione e archiviazione più modulare, realizzato per assicurare un'infrastruttura scalare sul lungo periodo. Si occupò anche di sviluppare la prima versione di una nuova interfaccia utente per la visualizzazione delle entità e un nuovo framework API.
La nuova interfaccia utente e le entità caricate furono nuovamente valutate allo scadere dei 18 mesi. Sebbene il grosso dell'attenzione per questa fase fosse rivolta alla valutazione qualitativa dei dati, OCLC era anche interessata a dei primi feedback sulla nuova interfaccia utente e le API. Il gruppo di studio si trovò soddisfatto per quanto riguardava la qualità dei dati sulle persone, sia per il grado di disambiguazione raggiunto sia per la ricchezza delle descrizioni. Le informazioni sulle Opere, invece, furono considerate subottimali, rendendo il lavoro di disambiguazione più impegnativo. Inoltre, apparve chiaro che sarebbe stato utile ottenere un maggior numero di collegamenti tra le entità. Per quanto concerne le funzioni di ricerca e di lettura delle entità, il gruppo di studio rispose positivamente alla nuova interfaccia utente e alle API.

#### 2.2.4 Quarto semestre

Durante gli ultimi sei mesi di progettazione furono portati avanti la valutazione e il caricamento dei dati, venne ultimata la realizzazione di strumenti per creare e modificare le entità attraverso la user interface e le API, furono aggiunte diverse lingue per l'interfaccia e i parametri valutativi per la qualità dei dati (vedi paragrafo 3.4).
La conclusione del progetto era programmata per dicembre e alla luce dei miglioramenti OCLC stabilì durante questa fase due periodi di test per il gruppo di studio, piuttosto che limitarsi a un'unica e più lunga fase di revisione alla fine del progetto. In ciascuno dei due periodi sarebbero state incorporate le ultime modifiche e aggiunte nuove proposte.
L'enfasi per entrambe le finestre di revisione verteva sulla gestione delle entità nell'interfaccia utente, comprendendo la possibilità di modificare l'UI, la visualizzazione dei punteggi di qualità dell'entità e l'aggiunta di ulteriori lingue per l'interfaccia. I dati valutati rimasero invariati dalla precedente fase, mentre gli informatici continuarono a raffinare i criteri di selezione dei dati e a migliorare i processi di caricamento.
L'interfaccia utente in quest'ultima fase presentava un differente layout delle informazioni, ovvero raggruppava le proprietà delle entità in categorie, quali "date significative" e "persone collegate". Complessivamente il gruppo di studio trovò l'interfaccia facile da usare e agile per completare le mansioni richieste durante il test, tra le quali la ricerca di un'entità e la navigazione nella descrizione dei dettagli. Etichette esplicite e la navigazione di pagina aiutarono i tester a trovare velocemente gli elementi della descrizione che cercavano. Inoltre, essi apprezzarono particolarmente la lista dei link con collegamenti esterni, per esempio il collegamento al VIAF delle entità, e ritennero che il riferimento a questi altri ID e fonti di autorità potesse svolgere un ruolo chiave nel loro flusso di lavoro.
L'ultima finestra di analisi e revisione tenutasi a dicembre introdusse le funzionalità di creazione e modificazione delle entità. I feedback positivi a riguardo sottolinearono l'importanza di un'interfaccia utente capace di creare e gestire le entità come complemento alle API. Una volta completata la traduzione dell'interfaccia, i tester potevano facilmente spostarsi da una lingua all'altra scegliendo tra inglese, francese, spagnolo, tedesco e olandese. I suggerimenti risultanti da queste due finestre di valutazione furono presi in considerazione per l'interfaccia finale.
Il 31 dicembre 2021 ha segnato la formale conclusione del progetto finanziato. Oltre questa data i costi per concludere i report e l'attività di lancio sono stati coperti da OCLC.

## 3. WorldCat® Entities: il modello di dati

### *3.1 Introduzione*

Oltre all'aver esplorato e sviluppato l'architettura di sistema, un eguale sforzo è stato compiuto nel campo della modellazione dei dati per costituire l'aggregazione delle entità. L'obiettivo era quello di poter generare in massa entità di tipo *Persona* e *Opera* estraendo i dati necessari dai record bibliografici di WorldCat, di Wikidata e dal VIAF (Virtual International Authority File). I dati ricavati da queste tre fonti sono stati poi combinati per creare delle entità con identificatori univoci persistenti e con relazioni semantiche verso altre entità. Il modello di dati ha un ruolo cruciale quando si tenta di raccogliere ingenti quantità di dati provenienti da fonti diverse, dato che fornisce il modello concettuale tramite cui le entità vengono create e interrelate tra loro. I

motori di ricerca, grazie a questa modellazione dei dati, possono inoltre rispondere correttamente a delle query complesse che altrimenti richiederebbero un'inferenza umana. Questo permette di ampliare la nostra definizione di "utente" includendo i computer, i quali diventano così consumatori di dati alla stregua degli esseri umani.

### 3.1.1 Classi e proprietà

Gli elementi fondanti di un modello di dati sono le classi e le proprietà. Il termine *classe* fa riferimento alla tipologia di entità che viene descritta. L'ambito di questo progetto era limitato alle entità *Persona* e *Opera*, essendo questi gli elementi basilari che costituiscono i dati bibliografici. Ciò nonostante, un modello dati espanso potrebbe includere altre classi di entità, come Luoghi, Eventi, e Concetti. Le *proprietà* fanno riferimento alle relazioni instaurate tra le entità. Esprimono un'azione nella forma di una tripla. Le triple sono costruite seguendo un formato soggetto-predicato-oggetto al cui interno un'entità (il soggetto) è unita tramite una proprietà (il predicato) a una seconda entità (l'oggetto). Le proprietà e le classi sono pensate per contestualizzare un'entità immettendola in una rete di altre entità correlate piuttosto che definendola da sola o in modo decontestualizzato. Questo facilita le funzioni di ricerca, la ricchezza dei risultati e coadiuva i processi di disambiguazione. Il modello è stato realizzato per fornire queste prime classi e proprietà al fine di raggiungere gli obiettivi preposti e fornire una solida base per eventuali sviluppi futuri.

### 3.1.2 Vincoli

All'interno del modello i vincoli definiscono come le proprietà possono essere usate per legare le entità. Esse rafforzano la interconnettività logica del modello e permettono l'identificazione e la risoluzione di dati anomali. I vincoli delle proprietà includono i concetti di *dominio* e *intervallo*, laddove esse determinano quale soggetto (il dominio) possa essere usato in congiunzione con quale oggetto (l'intervallo). Per esempio, sebbene una persona abbia una data di nascita e un'opera una data di ideazione, non sarebbe altrettanto logico se una persona avesse una data di ideazione e un'opera una data di nascita. Pur essendo questo evidente per un individuo non lo è altrettanto per una macchina. I vincoli all'interno del modello di dati esplicitano queste distinzioni necessarie al fine di facilitare la creazione e l'analisi delle relazioni tra entità.

## *3.2 Prime idee sulle entità Opere*

Dopo aver definito la struttura fondamentale del modello, OCLC ha considerato la definizione dell'entità *Opera* relativa al modello LRM (Library Reference Model). Il modello LRM stabilisce una gerarchia di tipo Opera-Espressione-Manifestazione-Item (WEMI), ove in LRM un'opera si realizza mediante un'espressione che si materializza tramite una manifestazione ed è esemplificata da un item. Il modello, quindi, muove dalla più astratta concettualizzazione di una certa risorsa fino al singolo oggetto fisico che la rappresenta. Questo modello, mentre fornisce un chiarissimo quadro concettuale, presenta una difficoltà di applicazione nel creare entità Opere (nella visione di LRM) per i dati bibliografici in WorldCat. Infatti, i record MARC al massimo descrivono le risorse a livello di manifestazione. Per questo i dati relativi alle entità Opere avrebbero dovuto essere estratti dai dati delle manifestazioni, ma ciò avrebbe comportato una grossa sfida nell'unire e deduplicare le risultanti entità, dato che un'Opera prevede un notevole numero di manifestazioni.
Una proposta iniziale per derivare entità Opera fu, invece, quella di crearle usando dei *cluster* di dati FRBR (Functional Requirements for Bibliographic Records) già presenti in WorldCat. Questi cluster FRBR di WorldCat sono pensati per raccogliere tutte le versioni di una data risorsa, incluse le traduzioni, le edizioni dei diversi editori e le ristampe. Questo era un iniziale tentativo da parte di OCLC per avvicinarsi all'intento aggregativo dell'entità Opera di LRM. Questi cluster vengono creati attraverso processi automatici che comparano gli elementi di metadatazione fondamentali nei record bibliografici di WorldCat per identificare e raggruppare i record fondati sulle stesse Opere. Tuttavia, OCLC raccoglie i dati da centinaia di diverse biblioteche e fonti istituzionali, e l'inevitabile inconsistenza nei dati (dovuti a diversi processi di catalogazione, così come a banali errori umani) presentava un effetto controproducente, in quanto i processi automatici di creazione dei cluster non garantivano un'accuratezza del 100%. Di conseguenza è nata la preoccupazione che l'inconsistenza di questo processo si riflettesse sulla qualità delle entità.
Un'altra valutazione fatta da OCLC è stata quella relativa alla visualizzazione e all'accesso da parte dell'utente finale. Una singola opera può avere centinaia di entità derivate: un ampio numero di traduzioni, adattamenti e riduzioni. Considerato il fatto che le ricerche lato utente tendono a concentrarsi sui dettagli delle espressioni del modello LRM, vedi un determinato formato materiale o una lingua in cui è scritta l'opera, una ricerca fondata sul livello Opera di LRM avrebbe potuto presentare troppi risultati irrilevanti per l'utente. Per questa

ragione, OCLC ha scelto nel suo progetto di considerare l'entità Opera equivalente al livello di Espressione in LRM, in quanto questo avrebbe fornito sia un'affidabile creazione delle entità che risultati utili per le ricerche.
Il progetto sull'entità Opere è stato poi derivato da un'analisi dei cluster FRBR per identificare in ciascuno di essi una "espressione rappresentativa" sulla quale basare il progetto delle entità Opere. Diversi fattori sono stati considerati per scegliere l'espressione rappresentativa, inclusa la completezza, il livello di catalogazione, e la lingua usata durante la fase di catalogazione. Una volta scelto il record, il processo di estrazione dati ha identificato i campi chiave che sono stati estratti e aggiunti per creare la descrizione dell'entità Opera.

## *3.3 MVED e parametri qualitativi*

Rispetto a un piano di progetto realistico, il modello di entità implementato non intende fornire un insieme esaustivo di tutti i possibili attributi e proprietà potenzialmente relativi a una determinata entità; è stato, invece, stabilito un set di parametri, denominato MVED (Minimum Viable Entity Description), per ciascuna tipologia di entità. IL MVED definisce la soglia minima di attributi e di relazioni necessari per rendere chiara ciascuna entità e garantirne la disambiguazione.
Sulla base del MVED, OCLC ha definito dei parametri qualitativi per stabilire l'idoneità allo scopo di una determinata entità di progetto. Questa metrica valutativa della qualità è stata definita con l'intenzione di fornire agli utenti un completo, trasparente e comprensibile profilo dei dati di un'entità in un preciso momento. Gli utenti sono in questo modo assistiti nel prendere decisioni su come migliorare e correggere tali dati. Questi parametri sono aggiornati in tempo reale man mano che definizioni e dichiarazioni vengono aggiunte, corrette o rimosse, permettendo all'utente un feedback immediato sull'impatto del proprio lavoro. Il controllo qualità è calcolato sulla base e sul peso di tre parametri: completezza (25%), attendibilità (25%), e grado di disambiguazione (50%). La completezza è determinata dall'aderenza dell'entità ai principi del MVED. L'attendibilità è determinata dalla presenza o assenza di collegamenti a fonti esterne controllate, che conferma la provenienza e le dichiarazioni di un'entità. La disambiguazione è determinata dal numero e dalle tipologie delle relazioni; le entità sono, quindi, differenziate da altre entità in virtù delle loro relazioni semantiche. La combinazione tra MVED e la metrica sulla qualità fornisce un affidabile set di strumenti con cui analizzare e controllare le entità.

## *3.4 WorldCat Entities: caratteristiche del modello*

### 3.4.1 Caratteristiche comuni

Le caratteristiche comuni a tutte le entità, indifferentemente dalla tipologia di appartenenza, sono le seguenti: un identificativo univoco e persistente, un'etichetta e un campo tipologia. Inoltre, ogni definizione di un'entità include un riferimento relativo alla provenienza di quella dichiarazione, e degli identificatori esterni. Questi identificatori appaiono sotto forma di URI (Uniform Resource Identifiers) e collegano la nostra entità a una diversa versione della medesima entità presente in un altro sistema (per esempio un authority file).

### 3.4.2 Entità Persona

Oltre alle caratteristiche comuni a tutte le entità, le proprietà distintive per le entità della classe *Persona* sono:

- le proprietà che contestualizzano l'entità di una persona in uno specifico luogo e tempo, necessari per la disambiguazione. Esse includono le date significative di nascita e morte e i relativi luoghi di nascita e morte;



- le proprietà che definiscono le relazioni tra l'entità persona e le altre entità. Tali proprietà sono un valido strumento di ricerca giacché permettono agli utenti di navigare senza sforzo attraverso il reticolo delle entità, ed includono le persone correlate all'entità in questione (fratelli/sorelle, genitori, figli e "influenzato da") così come le "opere degne di nota";
- le proprietà che indicano le varianti del nome e i nomi alternativi dell'entità (nome di nascita, alias, pseudonimo ecc.). Difatti, per fornire dei risultati di ricerca esaustivi è necessario ricondurre ad un'unica entità tutte le forme nome con cui un individuo si è fatto conoscere durante la sua vita.

#### 3.4.3 Entità Opera

Oltre alle caratteristiche comuni a tutte le entità, le entità della classe *Opera* presentano le seguenti proprietà distintive:

- le proprietà che definiscono le responsabilità autoriali degli agenti (compositore, architetto, sceneggiatore ecc.). Il sistema permette di scegliere tra una vasta gamma di responsabilità di tipo autore per identificare tutti coloro che hanno contribuito in qualche modo alla realizzazione dell'opera in questione. La specificità di queste relazioni semantiche è particolarmente importante se un'opera presenta più di una responsabilità autoriale, ad esempio quando l'autore e l'illustratore sono due persone distinte;
- le proprietà che contestualizzano il legame con altre opere (adattamenti, traduzioni, "basato su"). Queste proprietà immettono l'entità in questione in un più largo ecosistema di opere d'intelletto. Tale legame permette una più facile navigazione nella rete di relazioni che intercorrono tra opere derivate. Vi sono poi ulteriori proprietà tese a sottolineare le differenze tra opere in relazione tra loro (come "data di creazione" e "lingua");
- le proprietà che definiscono l'argomento o il tema trattato nell'opera. Queste aiutano gli utenti durante la ricerca per soggetto;
- una proprietà che definisce il sottotipo dell'entità (istanza di) con la quale si indica la specifica forma di un'opera, per esempio testo, suono, immagine in movimento ecc.

### *3.5 Riepilogo del modello di dati*

Lo sviluppo di un modello di dati pratico, preciso e scalabile è stato essenziale per realizzare questo progetto. Il framework fornito dal modello di dati ha reso possibile la trasformazione delle risorse basate su testo, quali i record MARC e gli authority file, in entità completamente basate sui linked data.

## 4. Previsioni future

Una volta terminato il progetto shared entity management infrastructure, OCLC ha lanciato in rete la sua versione pubblica, WorldCat Entities. Al suo interno sono presenti più di 150 milioni di entità delle classi Persona e Opera. Queste entità, garantendo degli URI permanenti e riusabili da chiunque, serviranno come fondamenta per lavorare con i linked data. Gli strumenti per la gestione delle entità, permettendo agli utenti di creare e modificare le entità, saranno utilizzabili tramite OCLC Meridian e le API. Queste ultime saranno rilasciate a seguito di una fase di sviluppo e consentiranno alla comunità di partecipare attivamente al mantenimento dei dati relativi alle entità.
Nuove aree di interesse per sviluppi futuri prese in considerazione da OCLC:

- sviluppo di meccanismi di sincronizzazione tra i record in formato MARC di WorldCat e le entità linked data in WorldCat. Sebbene il settore bibliotecario stia migrando in massa verso l'applicazione dei linked data, è chiaro che il formato MARC rimarrà uno dei principali formati in uso per molti anni a venire. OCLC continuerà a impegnarsi per mantenere aggiornata l'infrastruttura MARC e nel frattempo continuerà a sviluppare le tecnologie sui linked data. Questi due paradigmi informatici si valorizzeranno a vicenda attraverso la sincronizzazione e l'arricchimento reciproco dei dati delle risorse equivalenti, indifferentemente dal fatto che essi siano in formato MARC o in versione linked data.
- Integrazione dei collegamenti verso vocabolari controllati. Il progetto è stato basato su un modello di dati essenziale; tuttavia, centinaia di vocabolari controllati di natura specialistica potrebbero essere collegati all'infrastruttura delle WorldCat *entities*. Ciò garantirebbe alle funzioni di ricerca una maggiore qualità. Questo è un valido esempio di come "la colonna vertebrale di entità" definita dal progetto possa essere estesa per includere vocabolari specializzati in modo da arricchire il più ampio ecosistema informativo. Ciò detto, è importante che questo eventuale sviluppo sia portato avanti in modo ponderato: sarebbe difatti importante evitare di concentrarsi troppo sui lemmi della lingua inglese o manifestare una iper-attenzione verso i vocabolari "occidentali", poiché questo comprometterebbe la portata globale che OCLC aveva inizialmente pensato per le WorldCat Entities. La possibilità di non incappare in questo errore di-

penderà dalla continua collaborazione tra OCLC e le biblioteche afferenti al progetto.

- Modellazione dei dati relativi al posseduto per garantire agli utenti l'accesso diretto alle risorse. Il primo passo logico in questa direzione sarebbe la creazione di entità che rappresentino le stesse biblioteche che hanno riversato il proprio posseduto in WorldCat. Per realizzare questa idea OCLC si avvarrà dell'enorme bacino di dati relativo agli authority record delle istituzioni afferenti a OCLC, così come del WorldCat Registry. Una volta create le entità delle organizzazioni, potranno essere introdotti i dati granulari relativi al posseduto, così da permettere la consegna delle risorse agli utenti.

Il progetto shared entity management infrastructure, assieme alla sua applicazione per gli utenti (WorldCat Entities), rappresenta una significativa tappa nello sviluppo di un robusto, persistente e scalabile bacino di linked data. Con questa infrastruttura pronta a fornire nuovi modi per descrivere e collegare le entità, OCLC non vede l'ora di soddisfare le moderne esigenze degli operatori di biblioteca e dei loro utenti.
Per navigare le WorldCat Entities, ti preghiamo di visitare il sito entities.oclc.org.

## Ringraziamenti

OCLC desidera ringraziare la Mellon Foundation per la sua sovvenzione al progetto e tutti i membri del gruppo di studio per l'aiuto fornito.

## RIFERIMENTI BIBLIOGRAFICI

Bahnemann, Greta, Michael Carroll, Paul Clough, Mario Einaudi, Chatham Ewing, Jeff Mixter, Jason Roy, Holly Tomren, Bruce Washburn, Elliot Williams, *Transforming Metadata into Linked Data to Improve Digital Collection Discoverability: A CONTENTdm Pilot Project*, Dublin, OCLC Research, 2021, https://doi.org/10.25333/fzcv-0851.

Godby, Jean, Karen Smith-Yoshimura, Bruce Washburn, Kalan Davis, Karen Detling, Christine Fernsebner Eslao, Steven Folsom, Xiaoli Li, Marc McGee, Karen Miller, Honor Moody, Holly Tomren, Craig Thomas, *Creating Library Linked Data with Wikibase: Lessons Learned from Project* Passage. Dublin, OCLC Research, 2019, https://doi.org/10.25333/faq3-ax08.

OCLC, *WorldCat Entity Data guidelines and standards*, 24 maggio 2022, https://help.oclc.org/Metadata_Services/WorldCat_Entities/WorldCat_Entity_Data_guidelines_and_standards.

ABSTRACT
*OCLC, with grant assistance from the Mellon Foundation, built a shared entity management infrastructure to support linked data in a production capacity. This infrastructure provides persistent, centralized, and jointly curated entity data, to support libraries and cultural heritage institutions as they seek to implement new descriptive workflows. This paper will outline the entity infrastructure project, its outcomes, and the future outlook. It will include a detailed look at data modeling, as well as OCLC's close work with libraries throughout the grant.*