



Trasformazione dei
metadati in linked data
per migliorare la visibilità
delle collezioni digitali:
Un progetto pilota CONTENTdm

SINTESI ESECUTIVA

Durante lo svolgimento del progetto CONTENTdm Linked Data Pilot, OCLC ha collaborato con le istituzioni che gestiscono le loro collezioni digitali tramite il servizio CONTENTdm di OCLC per studiare i metodi e la fattibilità della trasformazione dei metadati in linked data per migliorare la visibilità e la gestione del materiale culturale digitalizzato e della sua descrizione. Questo report, *Transforming Metadata into Linked Data to Improve Digital Collection Discoverability (Trasformare i metadati in linked data per migliorare la visibilità delle collezioni digitali)*, descrive l'andamento del progetto e le sue principali aree di indagine, oltre a riassumere i principali risultati e le conclusioni generate da questo studio collaborativo.

Il progetto è stato ideato per aiutare il team OCLC e i partecipanti all'indagine pilota a capire meglio le seguenti domande:

- Quanto divergono le pratiche dei dati descrittivi tra le istituzioni che utilizzano CONTENTdm e quali strumenti sono necessari per fare questa valutazione?
- È possibile sviluppare un modello di dati condiviso ed estensibile per supportare le diverse esigenze e richieste di una serie di tipologie di materiali e di istituzioni?
- Qual è la giusta combinazione di attenzione umana e automazione per riconciliare in modo efficace le voci dei metadati con le entità dei linked data?
- Quali tipi di strumenti possono aiutare gli esperti del settore ad estendere la descrizione del materiale culturale?
- Dopo aver trasformato i metadati provenienti da diverse istituzioni e collezioni, esistono inediti strumenti di scoperta che possono aiutare i ricercatori a trovare nuove connessioni, o che prima erano invisibili, attraverso un sistema di scoperta centralizzato?
- Quali sono gli interessi istituzionali e personali nel cambio di paradigma del passaggio ai linked data?

Nel corso delle attività, il team del progetto pilota e i partner hanno osservato un miglioramento della gestione e della scoperta dei metadati in azione...

Al progetto hanno partecipato cinque organizzazioni che rappresentano in modo trasversale diverse tipologie di istituzioni: la Huntington Library, Art Museum, and Botanical Gardens; la Cleveland Public Library; la Minnesota Digital Library; le Temple University Libraries e le University of Miami Libraries. Il progetto pilota si è concentrato sullo sviluppo di flussi di lavoro efficienti per la trasformazione dei metadati, sulla valutazione delle interfacce esistenti per sfruttare i linked data, e sulla sperimentazione di applicazioni costruite nell'ambiente Wikibase per la gestione dei linked data appena creati.

Nel corso delle attività, il team del progetto pilota e i partner hanno osservato un miglioramento della gestione e della scoperta dei metadati e hanno riflettuto sui potenziali benefici: metadati di qualità superiore e più ricchi possono essere gestiti dal personale con maggiore efficienza; inoltre, i linked data possono essere utilizzati per aggiungere informazioni contestuali e per creare una rete di connessioni che rifletta meglio la conoscenza del mondo reale. Questo contesto e queste connessioni possono aiutare i ricercatori a comprendere in modo più completo i materiali della collezione, favorendo un maggiore coinvolgimento e utilizzo da parte dei membri della comunità.

I metadati più ricchi e di qualità più elevata possono essere gestiti dal personale con maggiore efficienza.

I linked data possono essere utilizzati per aggiungere informazioni contestuali e creare una rete di connessioni che rifletta meglio la conoscenza del mondo reale.

Sebbene i risultati del progetto pilota si basino su un insieme limitato di istituzioni e collezioni, essi suggeriscono con vigore che esiste un potenziale significativo per una migliore scoperta e una gestione più efficiente dei dati quando i materiali digitalizzati sono descritti utilizzando un modello di dati condiviso, dove le voci sono associate a entità e relazioni di linked data, e quando le entità e le relazioni sono riunite in un'unica aggregazione.

La domanda principale che ha guidato il progetto sui linked data è stata: per un cambio di paradigma di questa portata, come si può fare in modo che i cambiamenti fondamentali siano più scalabili, accessibili e sostenibili? Il progetto ha dimostrato che la portata e l'entità dello sforzo necessario per analizzare, trasformare e riconciliare completamente tutti i metadati descrittivi attuali in linked data coerentemente modellati vanno oltre la portata di una singola agenzia centralizzata. Tale iniziativa richiederà l'impegno di risorse sostanziali e condivise da parte di una comunità decentralizzata di professionisti che dovranno essere dotati di strumenti e flussi di lavoro facilmente accessibili per effettuare la transizione. Le prove raccolte durante il progetto, e dettagliate in questo report sulla modellazione dei dati, la riconciliazione dei metadati e l'analisi dei dati forniscono nuove conoscenze su come questi strumenti e flussi di lavoro potrebbero essere progettati e utilizzati.

INTRODUZIONE

Il progetto **CONTENTdm**¹ Linked Data Pilot (a cui ci si riferisce, in questo report, anche come “progetto Linked Data”) è l’ultimo (dal 2020) di una serie di indagini² che OCLC ha organizzato e condotto per diversi anni nell’interesse di sviluppare una comprensione condivisa di come biblioteche, archivi e musei possano effettuare la transizione verso i linked data. OCLC collabora con queste istituzioni per aumentare la capacità dei ricercatori di scoprire, valutare e utilizzare il materiale culturale digitalizzato, principalmente attraverso il supporto del servizio CONTENTdm per la creazione, la conservazione e la presentazione delle collezioni digitali uniche di una biblioteca.

Questo progetto Linked Data si è focalizzato sull’immaginare e valutare i sistemi e i flussi di lavoro scalabili e accessibili che saranno necessari per produrre ricche rappresentazioni di linked data di entità e relazioni, che contribuiranno a rendere visibili connessioni che prima erano invisibili. Il progetto si è basato sulla proposta di valore dei linked data, secondo la quale le migliori pratiche per la pubblicazione di dati strutturati sul web - l’uso di URI (Uniform Resource Identifiers) come nomi da dare alle cose, l’uso di URI HTTP per permettere alle persone di cercare quei nomi, la fornitura di informazioni utili sfruttando gli standard quando le persone cercano un URI, e l’inclusione di collegamenti ad altri URI per permettere alle persone di scoprire altre cose - porteranno a una rete globale interconnessa di dati che potrà servire sia agli sviluppatori che ai ricercatori³.

Al progetto hanno partecipato, in qualità di partner, cinque organizzazioni che rappresentano in modo trasversale diverse tipologie di istituzioni: la Huntington Library, Art Museum, and Botanical Gardens; la Cleveland Public Library; la Minnesota Digital Library; le Temple University Libraries e le University of Miami Libraries. I partecipanti al progetto pilota hanno collaborato con OCLC su una serie di studi mirati, tra cui lo sviluppo di flussi di lavoro efficienti per trasformare i metadati di origine in linked data, la valutazione delle personalizzazioni dell’interfaccia di CONTENTdm per sfruttare i linked data per la scoperta, la sindacalizzazione e la sperimentazione di nuove applicazioni costruite nell’ambiente Wikibase per il recupero dei dati, l’annotazione di immagini, la modifica e l’analisi dei metadati, e la scoperta.

Questo rapporto descrive l’andamento del progetto CONTENTdm Linked Data Pilot e le sue principali aree di indagine, condivide inoltre le esperienze delle cinque istituzioni partner che hanno partecipato al progetto, e riassume i risultati e le conclusioni chiave generate da questo studio collaborativo.

L’attenzione del progetto Linked Data alla sostenibilità e alla scalabilità ha posto molti quesiti da risolvere, tra cui:

Quanto divergono le pratiche dei dati descrittivi tra le istituzioni che utilizzano CONTENTdm e quali strumenti sono necessari per fare questa valutazione?

Il grande volume di metadati descrittivi del materiale culturale conservato in CONTENTdm ha offerto un eccellente banco di prova per valutare una transizione su larga scala verso i linked data. Inoltre, i risultati del progetto Metadata Refinery dell'OCLC, completato nel 2016, e del prototipo di linked data Project **Passage**⁴, completato nel 2018, hanno fornito importanti indicazioni su come implementare un sistema che agevoli la mappatura, la riconciliazione, l'archiviazione e il recupero di dati strutturati per materiali digitali unici. Questo progetto pilota si basa su tali intuizioni e successi. Le sezioni che seguono, che descrivono l'ambiente **Wikibase**⁵, le fasi di raccolta e trasformazione dei metadati e l'applicazione prototipo "Field Analyzer" evidenziano le sfide poste dall'applicazione di questo lavoro su larga scala.

È possibile sviluppare un modello di dati condiviso ed estensibile per supportare le diverse esigenze e richieste di una serie di tipologie di materiali e di istituzioni?

L'ampia varietà di modelli di dati e di pratiche descrittive attualmente utilizzate in CONTENTdm potrebbe essere significativamente più facile da gestire per il personale se fosse disponibile un modello di dati condiviso, e se tale modello condiviso potesse anche supportare una ricca scoperta per i ricercatori in un unico sistema di scoperta aggregato. Questo progetto si è proposto di sviluppare un modello di dati condiviso, basandosi sugli standard esistenti, ma consentendo estensioni in base alle prove emerse nei metadati di origine per classi e relazioni aggiuntive. La seguente sezione sullo sviluppo del modello di dati fornisce una panoramica e degli esempi dei risultati di questo lavoro.

Qual è la giusta combinazione di attenzione umana e automazione per riconciliare in modo efficace le voci dei metadati con le entità dei linked data?

Il progetto ha dedicato molto tempo e sforzi alla verifica dei flussi di lavoro di riconciliazione e alla prototipazione di nuovi strumenti per rendere questo lavoro più efficiente, pur mantenendo un'elevata qualità. La prototipazione di un nuovo endpoint di riconciliazione dei metadati ci ha aiutato a capire il potenziale di miglioramento delle prestazioni di quello che può essere un processo automatizzato che richiede molto tempo. Lo sviluppo dell'applicazione web "Retriever" per la ricerca di entità correlate in altri sistemi, e la loro trasformazione in nuove entità Wikibase, ha affrontato un ostacolo relativo al flusso di lavoro dei catalogatori. Di seguito sono descritti entrambi i prototipi.

Quali tipi di strumenti possono aiutare a estendere la descrizione del materiale culturale agli esperti del settore?

Il team del progetto ha sviluppato - e i partecipanti hanno testato - un'applicazione prototipo "Image Annotator" che potrebbe essere utilizzato dal personale della biblioteca, o da esperti in materia esterni alla biblioteca, per associare i titoli dei soggetti alle entità raffigurate nelle immagini, immaginando come i dati trasformati, insieme a nuovi strumenti, potrebbero aprire le porte a descrizioni più numerose e ricche da parte di una comunità impegnata. La descrizione dell'applicazione "Image Annotator" di seguito riportata include una sintesi dei risultati dei test di usabilità.

Dopo aver trasformato i metadati provenienti da diverse istituzioni e collezioni, esistono inediti strumenti di scoperta che possono aiutare i ricercatori a trovare nuove connessioni, o che prima erano invisibili, attraverso un sistema di scoperta centralizzato?

L'applicazione prototipo "Explorer", sviluppata durante il progetto e di seguito descritta, ha dimostrato la capacità di ricerca tra i dati di una serie di archivi, con ricerche e sfaccettature alimentate da entità derivate da file di autorità e da vocabolari creati dai bibliotecari. La collezione virtuale "Transportation Hub" inclusa nell'Explorer ha fornito al team del progetto, e ai partecipanti, un modo per provare in azione la scoperta dei linked data, lavorando con descrizioni di elementi tematicamente correlati e forniti da una sezione trasversale di istituzioni e collezioni, e trasformati in entità e relazioni separate.

Quali sono gli interessi istituzionali e personali nel cambio di paradigma del passaggio ai linked data?

La stretta collaborazione tra OCLC e i partner del progetto pilota è stata uno degli aspetti più gratificanti del progetto. Dato che la maggior parte del progetto è stata condotta mentre le persone, e le organizzazioni per cui lavorano, stavano vivendo significative interruzioni della loro vita e del loro lavoro con l'arrivo e lo sviluppo della pandemia COVID-19 nel 2020, all'inizio non era chiaro quale priorità e attenzione relativa potesse ricevere il progetto pilota. Ma l'attenzione e l'impegno dei partecipanti - e il sostegno di OCLC - non sono mai venuti meno, e abbiamo tratto reciprocamente grandi benefici da questa iniziativa. Per ulteriori approfondimenti e prospettive sull'impatto di questo progetto e sulle implicazioni per il nostro futuro comune, si vedano le seguenti sezioni sulla comunicazione della coorte e sulle riflessioni dei partner.

I risultati del progetto - illustrati in dettaglio in questa relazione - sulla modellazione dei dati, la riconciliazione dei metadati e l'analisi dei dati forniscono nuove conoscenze su come questi strumenti e flussi di lavoro potrebbero essere progettati e utilizzati, e presumiamo che informeranno le future ricerche sui linked data e gli sviluppi delle comunità di biblioteche, archivi e musei.

Il progetto CONTENTdm Linked Data Pilot rappresenta una nuova tappa nella crescente serie di attività di ricerca e sviluppo sui linked data che OCLC ha intrapreso negli ultimi dieci anni. I risultati del progetto - illustrati in dettaglio in questa relazione - sulla modellazione dei dati, la riconciliazione dei metadati e l'analisi dei dati forniscono nuove conoscenze su come questi strumenti e flussi di lavoro potrebbero essere progettati e utilizzati, e presumiamo che informeranno le future ricerche sui linked data e gli sviluppi delle comunità di biblioteche, archivi e musei.

NOTE

1. Panoramica del servizio di gestione dei contenuti digitali CONTENTdm di OCLC: <https://www.oclc.org/en/contentdm.html>
2. Panoramica della storia dei progetti di ricerca Linked Data di OCLC: <https://www.oclc.org/research/areas/data-science/linkedata/linked-data-outputs.html>
3. W3C. "Linked Data". <https://www.w3.org/wiki/LinkedData>
4. Panoramica del progetto pilota sui linked data dell'OCLC: <https://www.oclc.org/research/areas/data-science/linkedata/linked-data-prototype.html>