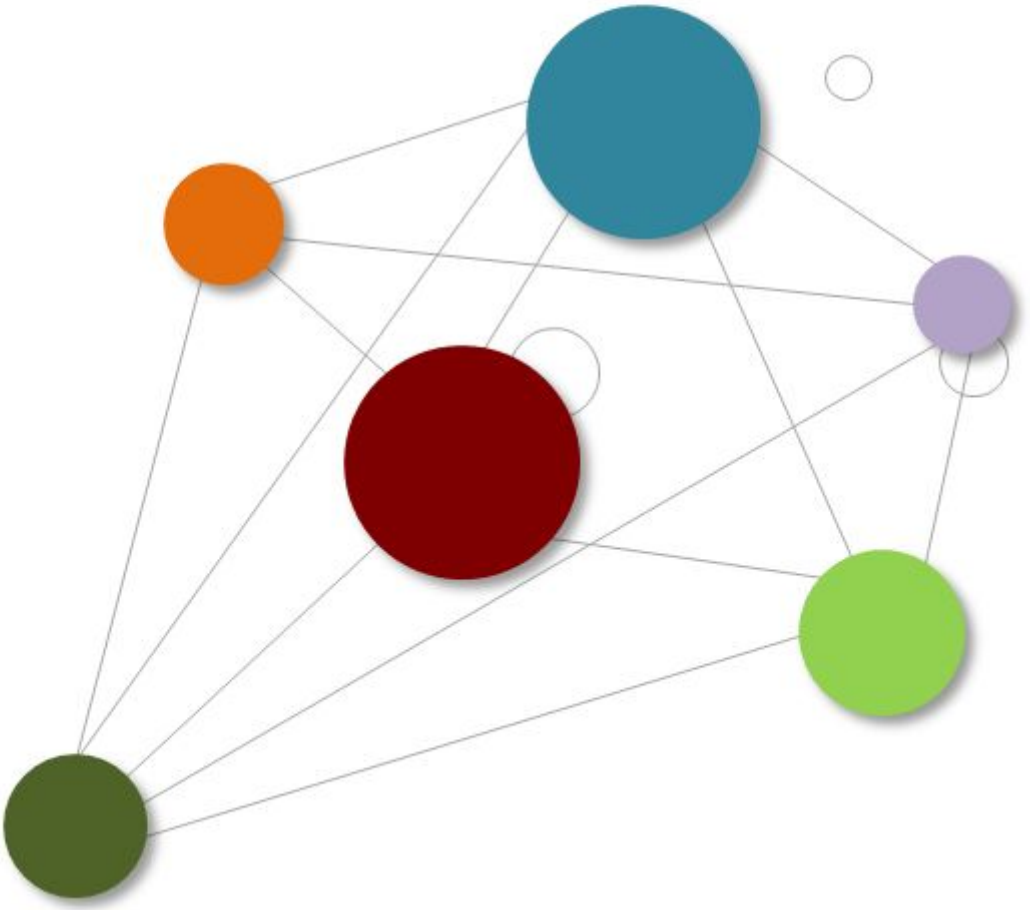
LOD Platform

Linked Data as a Service



@cult

Linked Data per la gestione della conoscenza condivisa nelle biblioteche, archivi e musei (LAM)

Indice

Per maggiori informazioni prego scrivere a info@atcult.it

1. Contesto

Biblioteche, archivi e musei (LAM) possiedono grandi quantità di dati e risorse che, spesso, possono restare nascoste all'interno dei cataloghi e degli archivi.
Sfruttare il potenziale di queste informazioni diffondendole a un pubblico più ampio rappresenta un'opportunità per arricchire ulteriormente il World Wide Web, promuovere una cultura di apertura alla conoscenza e creare una serie di vantaggi per tutti gli attori coinvolti nella catena produttiva dell'informazione.
In questo modo il patrimonio culturale mondiale può essere preservato a beneficio delle generazioni future e la ricerca può essere mantenuta nelle sue lingue originali, preservando varietà e vitalità culturale.
I linked data sono lo strumento che può rendere possibile tutto questo.

2. Web semantico e Linked Data

Il termine linked data (LD) si riferisce a un insieme di regole per la pubblicazione e la connessione di dati strutturati sul Web e si basa sul concetto di relazione qualificata.
L'idea centrale dei LD consiste nell'usare collegamenti ipertestuali non solo per identificare i documenti Web (come avviene nel World Wide Web tradizionale), ma anche entità arbitrarie del mondo reale con lo scopo finale di creare una vera e propria rete di dati univoci e certificati che ne costituiscano la base per la navigazione.
Viene così superato il problema dell'integrazione di cataloghi di oggetti diversi, grazie a un sistema che li interconnette indipendentemente dalla loro tipologia (libri, opere etc.) e dalla loro provenienza (biblioteche, musei, archivi etc.), creando un tessuto di informazioni estremamente più funzionale e completo per chi ricerca, e nello stesso tempo rendendo la conoscenza condivisa uno strumento collaborativo per le istituzioni culturali.
I LD pongono quindi le basi per ulteriori miglioramenti nelle dinamiche di collaborazione tra istituzioni e per il riutilizzo dei dati in contesti diversi, consentendo identificazione e possibilità di discovery più efficienti basate su Web.
Questo nuovo approccio è per biblioteche, archivi e musei un'occasione preziosa per fornire agli utenti strumenti di ricerca integrati, sempre più sofisticati e in linea con l'evoluzione delle nuove tecnologie web, mentre per i produttori di dati si configura come un passo decisivo per la creazione e la promozione di nuove forme di cooperazione.
I potenziali benefici sono significativi; flessibile e sostenibile, questo approccio cooperativo consente la creazione di un sistema ben strutturato per l'organizzazione dei dati, valorizzando la specificità delle varie tradizioni culturali e sfruttando le opportunità offerte dalle nuove tecnologie.
Le figure chiave nel settore potranno contribuire al fianco di organizzazioni più piccole, pronte a condividere la propria esperienza e l'unicità delle proprie risorse.
Biblioteche, archivi, musei, editori e fornitori saranno protagonisti nella generazione di nuovi dati e nella scoperta di nuove risorse, superando i confini dei loro domini specifici per creare opportunità di arricchimento dei dati che prima sarebbero state impensabili.

3. Che cos'è la LOD Platform

Con l'espressione LOD Platform ci si riferisce a un sistema tecnologico altamente innovativo, un ecosistema integrato per la gestione dei dati bibliografici, archivistici e museali, e la loro trasformazione in linked data, estensibile a seconda di scopi specifici.

Il nucleo della piattaforma LOD Platform è stato progettato nel progetto ALIADA, finanziato dall'UE, con l'idea di creare sistemi di gestione dei linked data scalabili e configurabili e interfacce di discovery in grado di adattarsi a ontologie di diversi domini di biblioteche, archivi e musei (LAM), in grado di automatizzare i processi di creazione e pubblicazione di linked open data, indipendentemente dal formato dei dati originali.
Lo scopo del framework è quella di aprire le possibilità offerte dai LD a biblioteche, archivi e musei fornendo maggiore interoperabilità, visibilità e reperibilità per tutti i tipi di risorse.
La partecipazione di istituti di differente natura e tipologia richiede ovviamente un'attenta analisi degli standard, formati e modelli utilizzati; la stessa organizzazione concettuale e tecnologica della piattaforma, fondata sull'ontologia bibliografica BIBFRAME 2.0, potrà essere arricchita da ulteriori ontologie, vocabolari e modellazioni a seconda delle esigenze specifiche.

Recependo standard, modelli e tecnologie riconosciuti come elementi chiave per la creazione di nuovi processi di gestione e fruizione delle conoscenze, LOD Platform consente:

- la creazione di una struttura dati basata sulle entità Agent, Work, Instance, Item, Place, come definite da BIBFRAME, ed estensibile per riconciliare altre entità;
- l'arricchimento dei dati tramite il collegamento a fonti di dati esterne;
- la riconciliazione e clusterizzazione di entità create dai dati originali;
- la conversione dei dati secondo il modello standard indicato dal W3C per i LOD, RDF – Resource Description Framework, mediante ontologie di settore;
- la consegna all'ente dei dati convertiti e arricchiti per il riutilizzo nei propri sistemi;
- la pubblicazione del dataset in LOD su storage RDF (triplestore);
- la costruzione di un portale di consultazione con interfaccia di navigazione basata su BIBFRAME o altre ontologie definite nel progetto specifico.

4. Panoramica di alto livello

Nella realizzazione di un progetto LOD Platform, i dati di biblioteche, archivi e musei vengono trasformati in LD tramite processi di identificazione delle entità, di riconciliazione e di arricchimento. Gli attributi sono utilizzati per identificare in modo univoco una persona, un'opera o altre entità, con le forme varianti riconciliate per formare un cluster di dati riferito alla stessa entità.
I dati sono successivamente riconciliati e arricchiti con ulteriori fonti esterne, così da creare una rete di informazioni e risorse.
Il risultato è un database di relazioni, aperto, e una knowledge base di cluster in RDF (CKB).
Il database utilizza i paradigmi del web semantico ma consente alle istituzioni che usano la LOD Platform di continuare a gestire i propri dati in modo indipendente e offre:

- arricchimento dei dati con URIs, sia per i record originali sia per l'output delle entità in linked data; per una panoramica delle fonti utilizzate per l'arricchimento si veda SVDE_URI_table_for_external_sources;
- conversione dei dati in RDF usando il vocabolario BIBFRAME e altre ontologie;
- creazione di una piattaforma di discovery con interfaccia utente;
- creazione di un database di relazioni e cluster accessibile in RDF attraverso un triplestore;
- implementazione di strumenti per interagire direttamente con i dati, consentendo la validazione, l'aggiornamento, il controllo a lungo termine e la manutenzione dei cluster e degli URI che identificano le entità;
- procedure di aggiornamento automatiche/batch;
- disseminazione dei dati in modalità automatica/batch;

- progressiva implementazione di workflow aggiuntivi come API per ILS, retro-conversione per sistemi di acquisizione e amministrazione locali, reportistica

Nelle varie fasi di avanzamento del progetto, le istituzioni partecipanti sono chiamate a collaborare attivamente, in modo che le decisioni relative alle fasi successive siano basate su fondamenta solide per un'implementazione su vasta scala nella comunità delle biblioteche, degli archivi e dei musei.
I dati utilizzati non vengono semplicemente convertiti in RDF, ma sono anche arricchiti con identificatori e interconnessi tra loro, così da consentirne l'utilizzo nell'ambiente LD.
L'obiettivo è quello di far sì che una grande mole di dati, che spesso rimangono nascosti o inespressi in silos ("contenitori") chiusi, riveli finalmente la sua ricchezza all'interno delle collezioni esistenti.

5. Quali sono i vantaggi

La LOD Platform, sviluppata secondo il principio di funzionalità, fornisce vari ambienti e interfacce per la creazione e l'arricchimento dei dati e offre flussi di lavoro in grado di rispondere alle differenti esigenze dei bibliotecari/archivisti/operatori museali, professionisti, studiosi, ricercatori e studenti partecipanti. I vantaggi sono molteplici:

- integrazione dei processi di un ambiente collaborativo con sistemi e strumenti locali;
- integrazione nel web semantico mantenendo la proprietà e il controllo dei propri dati, beneficiando dell'amministrazione semplificata dell'ambiente creato e di un ampio pool di dati;
- sviluppo continuativo e supporto per le funzioni di metadatazione secondo gli standard del web semantico;
- standard e infrastrutture per dati "a prova di futuro", assicurando cioè che siano compatibili con la struttura dei LD e del web semantico;
- arricchimento dei dati con ulteriori informazioni e relazioni precedentemente non espresse dal formato in uso, aumentando le possibilità di discovery per tutti i tipi di risorse;
- creare un ambiente che sia utile sia per gli utenti sia per gli operatori degli istituti (bibliotecari; archivisti; operatori museali);
- consentire ai bibliotecari una più ampia e diretta interazione con le entità linked data, con la possibilità di editarle attraverso il Cluster Knowledge Base Editor (maggiori dettagli più avanti);
- interfacce di ricerca avanzate per migliorare l'esperienza d'uso e fornire risultati di ricerca più ampi agli utenti;
- ricercabilità delle risorse ottimizzata, con la possibilità di scoprire dati che altrimenti sarebbero rimasti nascosti nei silos, consentendo perciò agli utenti finali di accedere a una grande quantità di informazioni che possono essere sia importate che esportate dalle istituzioni partecipanti.

Questo approccio sfrutta appieno il potenziale dei linked data, collegando le informazioni della biblioteca in un ambiente di ricerca dinamico a vantaggio di studenti, ricercatori e di tutti gli utenti della biblioteca che possono sperimentare nuovi modi di accedere alla conoscenza.

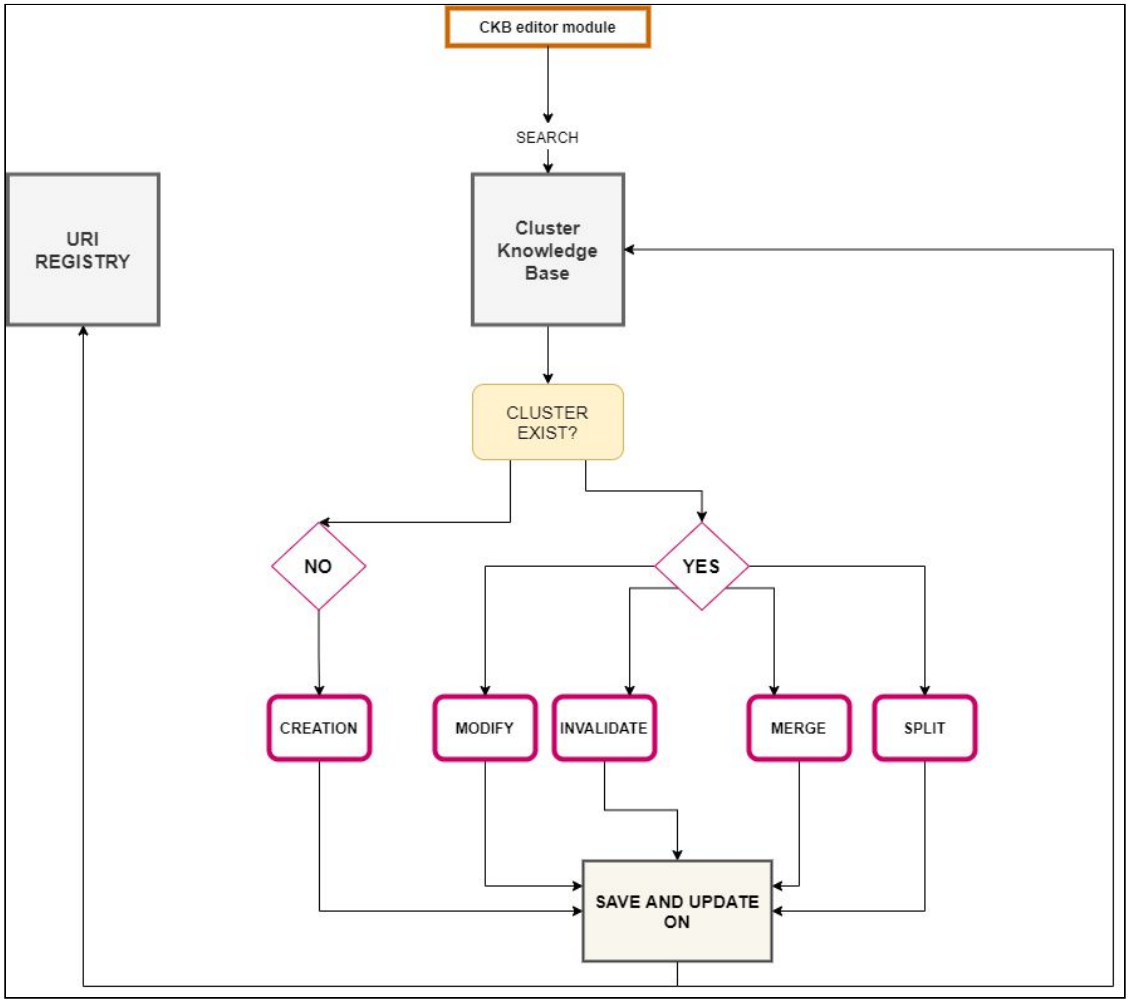
6. Valore aggiunto

È particolarmente importante evidenziare che è in corso un potenziamento della LOD Platform con un modulo dedicato alla modifica e all'aggiornamento delle entità nella Cluster Knowledge Base (CKB). Questo editor della CKB è stato chiamato J.Cricket nel contesto del sistema Share-VDE in cui è stato progettato, ed è concepito come un ambiente collaborativo con diversi livelli di accesso e interazione con i dati, consentendo diverse azioni manuali e automatiche su i cluster di entità salvati nel database, inclusa la creazione, la modifica, l'unione di cluster di Work, Agent etc.

J.Cricket consiste di due layers principali:

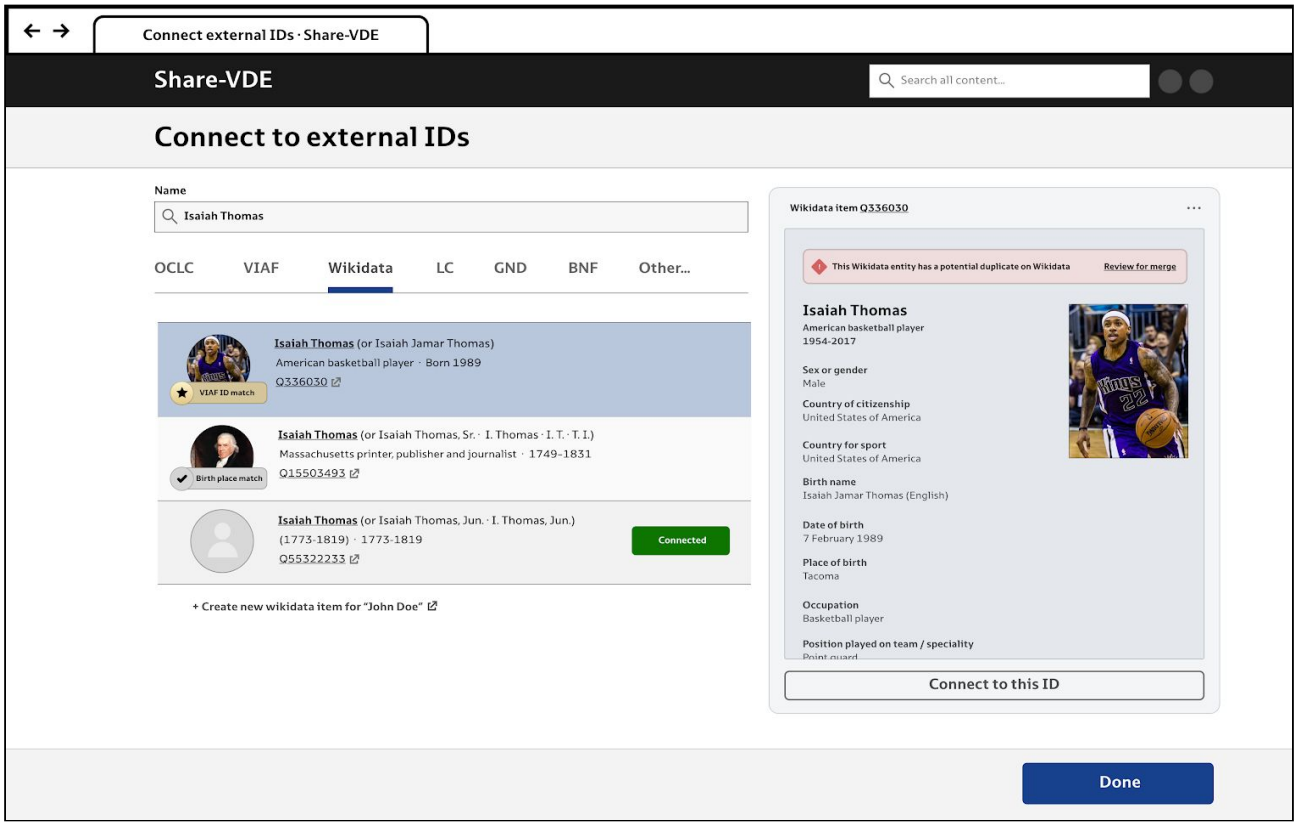
1. controlli automatici e aggiornamento dei dati effettuati dalla LOD Platform;
2. controlli manuali e modifica dei dati effettuati dall'utente tramite interfaccia web.

Tutte le modifiche alle entità, sia automatiche che manuali, sono riportate nell'URI registry, una fonte (disponibile anche in RDF) che tiene traccia degli aggiornamenti di ciascuna entità, soprattutto quando ciò ha un impatto sulla persistenza degli URI delle entità.
Il piano di sviluppo di J.Cricket e si estende nell'anno in corso e nel prossimo. Il sistema verrà rilasciato in fasi progressive e, indicativamente, il completamento delle funzioni di autenticazione, ricerca avanzata e merge di entità della CKB attualmente in fase di sviluppo è pianificato per la nella seconda metà del 2021. Altre funzionalità saranno implementate in un piano di sviluppo più ampio durante l'anno 2022.
Di seguito è riportato il flusso di alto livello dell'editor della CKB.



Qui di seguito il link a uno sketch di una delle funzioni manuali dell'editor CKB. Questo sketch rappresenta l'interfaccia in cui l'utente autorizzato può modificare i dati dell'entità Work: modifica dell'entità.

Un ulteriore valore aggiunto è la capacità della LOD Platform di interagire direttamente con fonti di dati esterne come ISNI e Wikidata. L'interazione con Wikidata è attualmente in fase di analisi e verrà attivata dallo stesso editor CKB, consentendo la ricerca dall'editor a Wikidata e l'arricchimento delle entità della LOD Platform con informazioni provenienti da Wikidata e viceversa. In questo modo l'editor consentirà la creazione di nuovi identificatori sia nelle sorgenti dati esterne (dove possibile o applicabile) sia nella Cluster Knowledge Base.
Lo sketch seguente illustra l'interfaccia dell'editor in cui vengono visualizzati i risultati di una query fatta su Wikidata: l'editor è pronto per arricchire l'entità con le informazioni di Wikidata che verranno salvate nella Cluster Knowledge Base.



7. Come funziona la LOD Platform

I componenti e gli strumenti sviluppati hanno lo scopo di creare un ambiente utile per la gestione della conoscenza, con interfacce di ricerca avanzate per migliorare l'esperienza d'uso e fornire risultati più ampi a biblioteche, archivi, musei e ai loro utenti:

- **Authify**, un modulo RESTFul che fornisce servizi di ricerca e full-text di dataset esterni (scaricati, memorizzati ed indicizzati nel sistema), relativi soprattutto ad Authority file (VIAF, Library of Congress Name Authority file, etc.) ma estendibile anche ad altre tipologie di dataset. È composto da due parti principali: un'infrastruttura SOLR per l'indicizzazione dei dataset e relativi servizi di ricerca, ed un livello logico che orchestra tali servizi per trovare una corrispondenza all'interno dei cluster delle entità.

- **Cluster Knowledge Base**, su database PostgreSQL, è il risultato del processo di elaborazione e arricchimento dei dati con fonti esterne per ogni entità; tipicamente: cluster di nomi (forme, autorizzate e varianti, dei nomi degli Agent) e cluster di titoli (punti di accesso autorizzato e forme varianti per i titoli delle Opere).
- **Lodify**, modulo RESTFul che automatizza l'intero processo di conversione e pubblicazione dei dati in RDF secondo l'ontologia BIBFRAME 2.0 in modo lineare e scalabile. È flessibile e adattabile a molteplici situazioni: permette, quindi, di gestire le classi e le proprietà non solo di BIBFRAME ma anche di altre ontologie, a seconda delle esigenze.
- **Triplestore**: la LOD Platform può attualmente essere integrata con due triplestore, uno open source (Blazegraph), più adatto per progetti di scala ridotta (fino a circa 2 milioni di record bibliografici), e uno commerciale (Stardog) adeguato per set di dati di grandi dimensioni. Per consentire agli utenti di accedere al triplestore Stardog, è disponibile lo strumento dedicato Stardog Studio (si veda https://stardog.studio e Stardog documentation). Inoltre, Stardog fornisce un set di API standard per interrogare i dati disponibili nel triplestore, si veda https://www.stardog.com/docs/#_api_overview.
- **Portale** di presentazione dei dati, per una modalità user-friendly di reperimento e navigazione dei dati.

8. Il ciclo di lavorazione dei dati in un progetto che utilizzi la LOD Platform

Negli schemi di seguito si illustra un processo 'tipo' di trattamento dei dati, dalla ricezione dei record originali alla pubblicazione sul portale. Il flusso è solo indicativo, ma esprime in modo esaustivo gli step inclusi in un processo di elaborazione dei dati.
Partendo dalla sinistra del **grafico 1**, i dati sono ricevuti dagli enti partecipanti (biblioteche, archivi, musei) in differenti formati (MARC, xml etc.). I dati possono essere bibliografici e di autorità.
I dati ricevuti sono elaborati secondo processi di Text analysis e String matching (rappresentati nel box "Similarity's score"), per identificare le entità incluse nei testi 'piatti' (i record), e preparare la creazione dei cluster.
Questa funzione di identificazione delle entità è potenziata ed estesa attraverso analoghi processi di Text analysis e String matching lanciati su fonti esterne (VIAF, ISNI, NAF, GND, Nuovo soggettario etc.), attraverso il framework Authify: questi processi generano l'arricchimento del dato con altre forme varianti provenienti dalle fonti esterne e con gli URI attraverso i quali la medesima entità è identificata su queste fonti (riconciliazione): il cluster di origine si arricchisce e consentirà, nel processo di conversione in Linked data, di attivare la funzione di interlinking, essenziale per la condivisione e il riuso dei dati nel web.

Il risultato di questi processi è triplice:
- Identificazione delle entità
- Arricchimento dei dati
- Creazione dei cluster attraverso processi di riconciliazione

I dati così ottenuti sono pronti per essere nuovamente processati, attraverso differenti canali:
- arricchimento manuale e quality check (nel caso in cui la biblioteca richiedesse uno specifico servizio ad agenzie esterne - quali Casalini Libri - o gestisse internamente i dati arricchiti ricevuti);

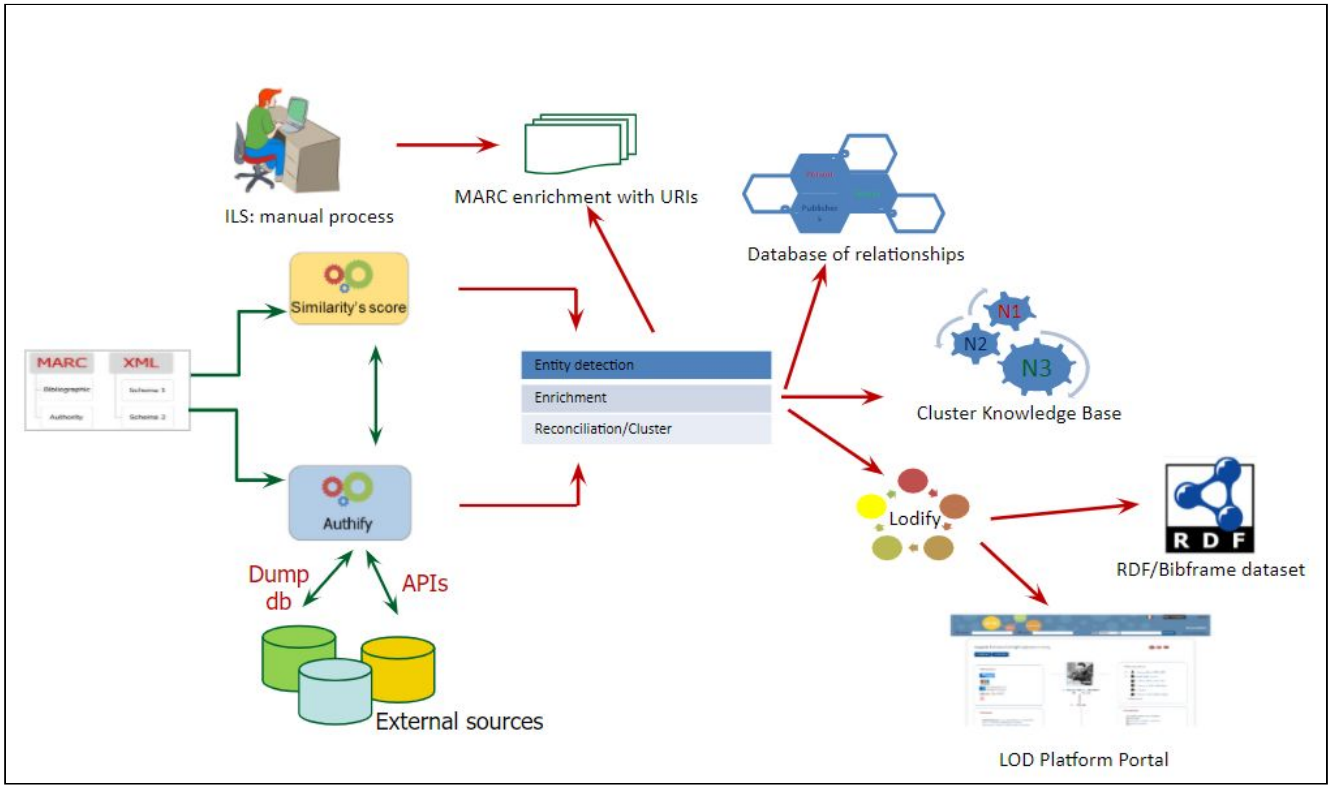
- estrazione delle relazioni ‘nascoste’ per la generazione e alimentazione di un database delle relazioni (che verrà riutilizzato in successivi step di arricchimento del dato e nelle fasi di pubblicazione, per estendere i collegamenti tra dati);
- generazione della Cluster Knowledge Base di progetto, disponibile in RDF (quindi come Linked Open Data) e accessibile come end point per query SPARQL e API;
- elaborazione/conversione in RDF, seguendo il modello BIBFRAME e/o altre ontologie di dominio suggerite dal progetto.

Al termine di questi processi, i dati sono pronti per essere indicizzati sul portale ed essere pubblicati su vari siti, in RDF.
Il sistema mette inoltre a disposizione procedure di aggiornamento periodico dei dati (secondo periodicità definita dal Cliente).

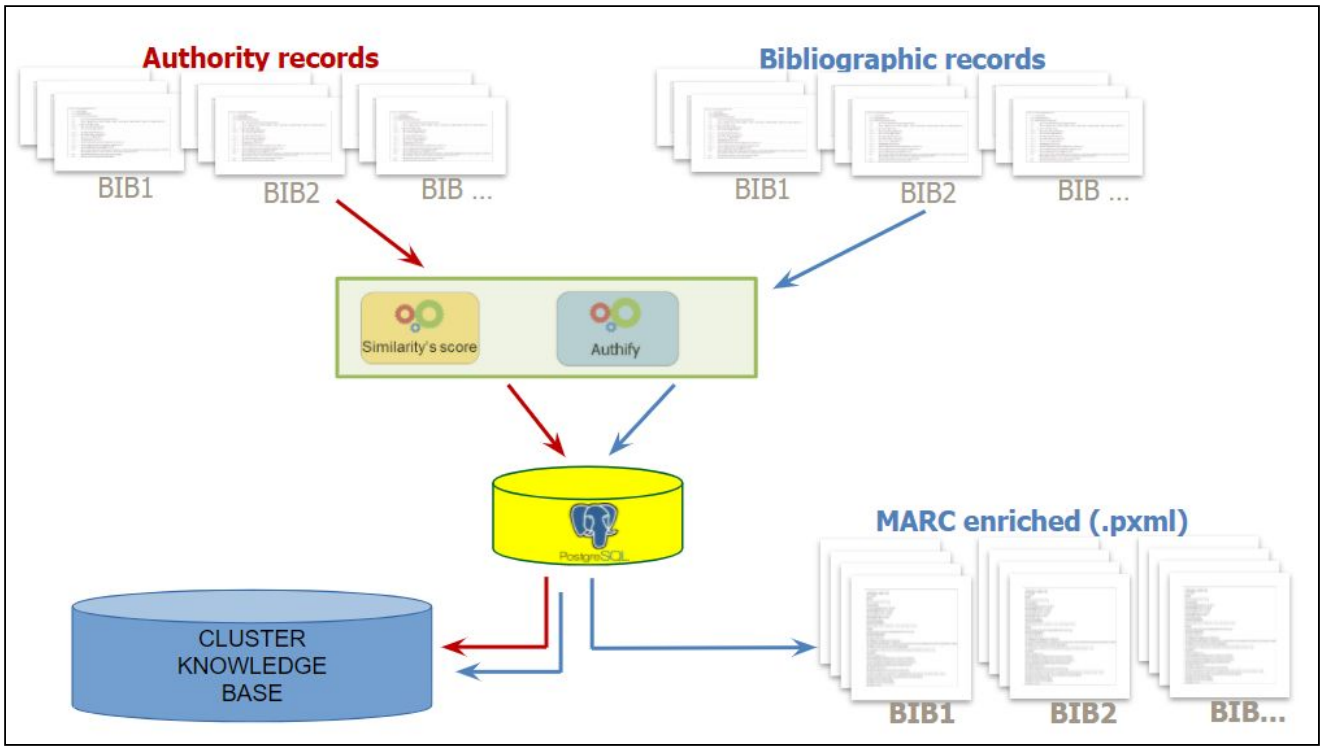
I grafici 2 e 3 mostrano più in dettaglio alcuni passaggi del workflow di alto livello rappresentato nel grafico 1.

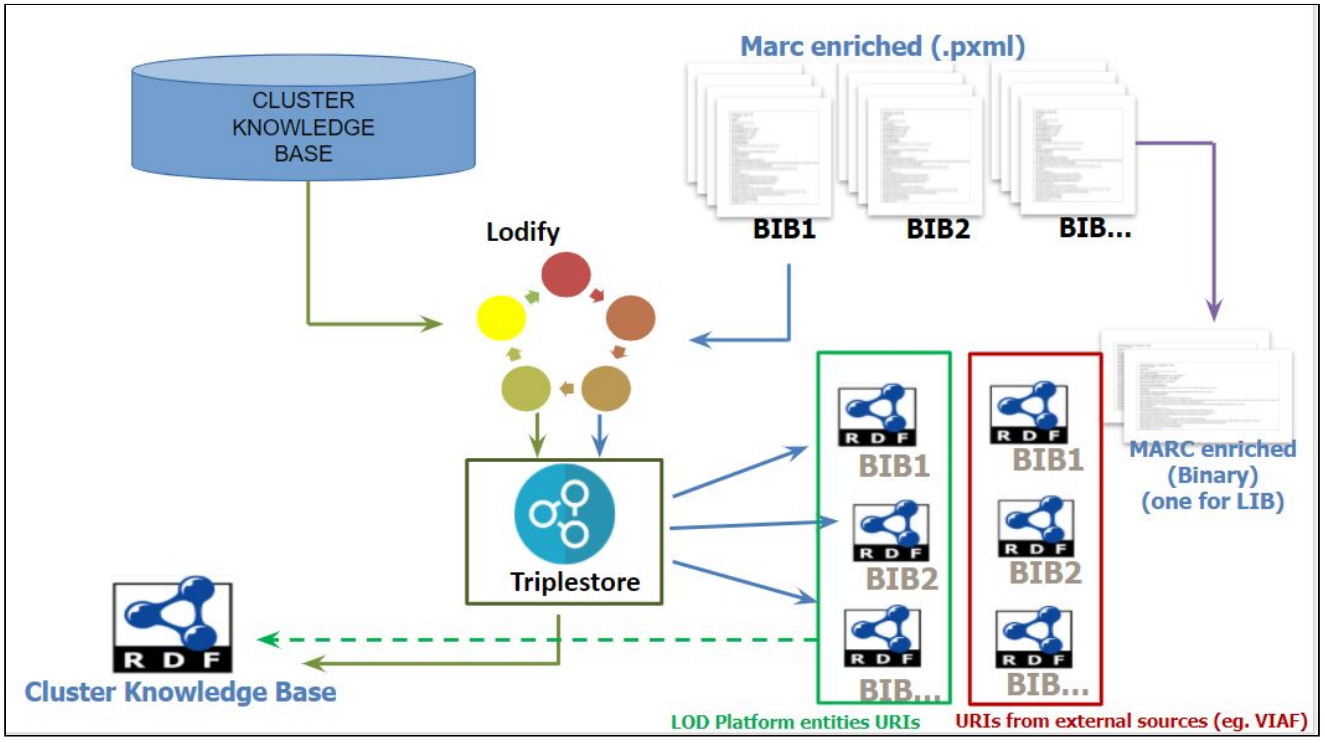
8.1. Il workflow della LOD Platform

Workflow di alto livello - grafico 1



Focus sui processi - grafici 2 e 3





8.2. Aggiornamento dei dati

La LOD Platform è in grado di gestire gli aggiornamenti dei dati. La gestione delle entità create con dati provenienti da fonti interne ed esterne è effettuata tramite approcci diversi che dipendono dalla fonte di dati (disponibilità di OAI-PMH o altri protocolli, aggiornamento periodico del dump etc.). Questi diversi approcci devono essere analizzati nel corso del progetto e ritagliati sulla singola istituzione, in modo che le entità possano essere aggiornate il più possibile mediante processi automatizzati. Inoltre, è in fase di sviluppo un editor di entità (J.Cricket) per consentire la gestione manuale delle entità, da parte di utenti autorizzati, attraverso un'interfaccia utente (si veda il capitolo dedicato).

Segue la descrizione di alto livello dei processi di aggiornamento automatico dei dati che sono già stati implementati per l'iniziativa Share-VDE e che possono essere adattati ad altri progetti.

Delta update di Share-VDE

Per aggiornamenti "delta" si intendono le modifiche ai record della biblioteca che vengono periodicamente inviati a Share-VDE. L'automazione dell'import in Share-VDE dei record delta ha lo scopo di aggiornare regolarmente i dati disponibili attraverso l'interfaccia di discovery e gli altri endpoint del flusso in cui i dati sono disponibili. Ciò significa aggiornare i dati delle entità clusterizzate e delle relative risorse ricercabili sull'interfaccia di discovery e nel triplestore, secondo la frequenza richiesta dalla biblioteca.

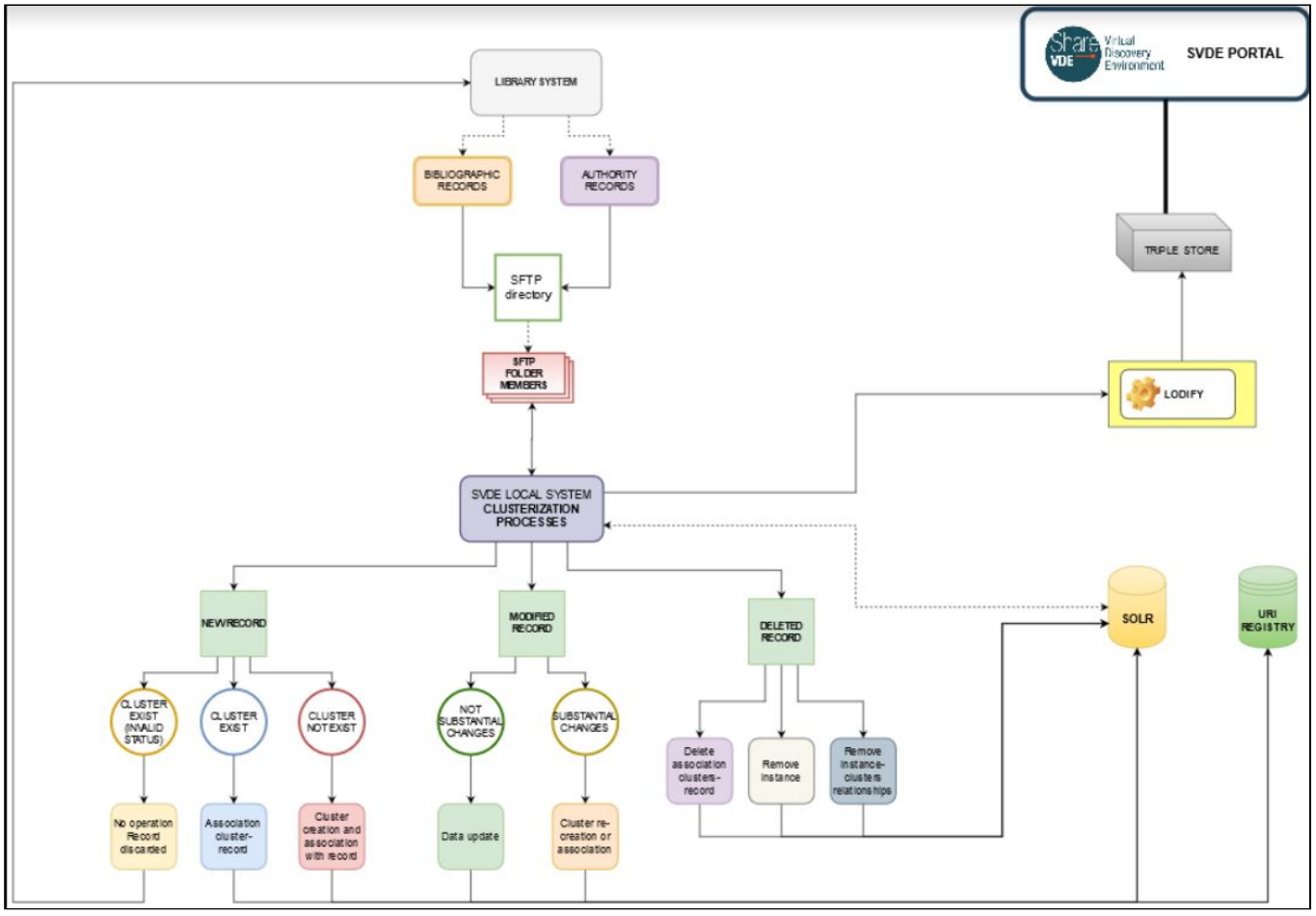
Passaggi del processo:

1. la biblioteca fornisce record bibliografici e di autorità alla directory SFTP di Share-VDE, nel percorso dedicato alla propria istituzione. Share-VDE si aspetta di ricevere da ogni biblioteca solo il delta dei record, cioè solo quei record che sono stati modificati o aggiunti o cancellati, rispetto al precedente invio;
2. lo script in esecuzione lato Share-VDE elabora i record in ordine sequenziale, secondo il nome del file, e accetta in input file .mrc (per record nuovi e modificati) e .txt (per record cancellati);
3. download dei record MARC della biblioteca nel sistema locale Share-VDE: dopo che i record MARC sono stati caricati dalla biblioteca sul server SFTP, uno script automatico collega il sistema interno di Share-VDE alle singole cartelle SFTP delle biblioteche, controlla se un nuovo il file è stato caricato sull'SFTP e scarica i record MARC nel sistema locale di Share-VDE. I file inviati dalle biblioteche di Share-VDE vengono automaticamente trasferiti dalla sottodirectory SFTP dell'istituzione che ha caricato i file nella corrispondente sottocartella del repository interno di Share-VDE;
4. elaborazione dei record MARC: gli aggiornamenti delta dei record MARC vengono elaborati secondo le procedure SVDE e arricchiti (ovvero viene aggiunto l'URI originale di Share-VDE e altri URI da fonti esterne come ISNI, VIAF, vengono arricchiti altri tag etc.). I dati vengono salvati nel database Postgres;
5. upload su Solr: i record elaborati vengono caricati sulla piattaforma Solr per l'indicizzazione, prima di popolare il portale Share-VDE. Tra i processi coinvolti, i dati dei record della biblioteca vengono elaborati e indicizzati in modo che la funzione di completamento automatico nei campi di ricerca su https://share-vde.org/ visualizzi i dati indicizzati (ad esempio autore, titolo) come risultato suggerito all'utente che esegue una ricerca per una risorsa;

6. dati aggiornati online: dopo la fase di indicizzazione, le informazioni elaborate da Share-VDE sono pronte per essere pubblicate su https://share-vde.org/.

Il processo di aggiornamento delta attiva: l'aggiornamento dei cluster di entità sul portale Share-VDE; l'aggiornamento dei dati disponibili sul triplestore Stardog; la consegna alle biblioteche dei record MARC arricchiti.

Segue un diagramma che mostra il flusso di dati per l'elaborazione dei record di aggiornamento delta nel sistema di Share-VDE.



9. Interfaccia utente

Qui di seguito sono raffigurati alcuni dettagli dell'interfaccia utente del portale di discovery attraverso il quale consultare le entità linked data convertite dalla LOD Platform (nell'esempio di seguito, le entità Agent e Work).