Gruppo di lavoro LOD SBN

Premessa

Il Comitato tecnico scientifico per SBN, nel 2014 ha incaricato l'ICCU di costituire un gruppo di lavoro con il compito di verificare la fattibilità per l'accesso e la pubblicazione del patrimonio informativo delle biblioteche della rete SBN in modalità Linked Open Data (LOD).

Il Gruppo di lavoro, composto da rappresentati dell'ICCU e della struttura di ricerca VAST LAB - PIN (Polo Universitario della città di Prato - Servizi didattici e scientifici per l'Università di Firenze), ha realizzato una analisi per la sperimentazione della produzione e pubblicazione di un set di dati SBN strutturati in LOD.

Le attività del Gruppo si sono orientate sul lavoro di mappatura concettuale con le classi e le proprietà del modello FRBRoo, utilizzando il modello CIDOC CRM, di un set di record estratto dall'OPAC SBN in formato UNIMARC. Sulla base di tale mappatura è stata quindi sviluppata la conversione dei dati in formato RDF ed è stato ha creato un prototipo di interfaccia online per la gestione e la ricerca del set dei dati SBN in formato LOD.

Il Gruppo è così costituito: Margherita Aste, Francesco Gandolfi, Luca Martinelli, Patrizia Martini, Maria Cristina Mataloni, Elisa Sciotti, Carla Scognamiglio (ICCU); Achille Felicetti, Cinzia Luddi (VAST LAB – PIN).

Introduzione

Il Gruppo di lavoro ha inizialmente individuato un modello concettuale di riferimento per la rappresentazione dei dati in formato LOD, uniformato ai modelli e agli standard internazionali che consentisse l'integrazione e lo scambio di informazioni bibliografiche tra i differenti settori della cultura (biblioteche, musei e archivi).

Il Gruppo ha quindi selezionato il set di dati da estrarre da SBN, destinato ad essere prima convertito e poi consultato sotto forma di linked data.

Il lavoro è proseguito con la fase di mappatura dei campi SBN secondo il modello concettuale di riferimento prescelto per la loro strutturazione in RDF (Resource Description Framework)1. Infine, il set di dati, selezionato e convertito in formato RDF, è stato caricato all'interno di un prototipo di triple store, attualmente in fase di test.

I dati saranno resi accessibili al pubblico sia attraverso interfacce di browsing facilitato per consentire una navigazione user friendly, sia attraverso un apposito endpoint SPARQL per l'interrogazione diretta dei dati semantici, rivolto ad utenti più esperti.
I linked data saranno, inoltre, resi disponibili per il download.

Analisi

In fase di analisi, sono stati scelti:

- il CIDOC Conceptual Reference Model (CRM)[2], modello concettuale di riferimento, che costituisce la struttura formale per descrivere i concetti impliciti ed espliciti e le relazioni nella documentazione del patrimonio culturale;

- le FRBRoo (Functional Requirements for Bibliographic Records object oriented)[3], ontologia di riferimento che fornisce un sistema concettuale dotato di una complessa rete di classi e proprietà, in grado di descrivere le principali entità e relazioni dell'ambito bibliografico, permettendo l'integrazione e lo scambio di dati di diversa tipologia e formato nell'ambito dei beni culturali.

Per quanto riguarda la selezione dei record oggetto della mappatura e conversione in RDF, è stato individuato un dataset costituito da 300 record estratti in formato

[1] Il Resource Description Framework (RDF) è uno standard flessibile proposto dal consorzio W3C per la codifica, lo scambio e il riutilizzo di metadati. Il data model è formato da risorse, presenti nel web con un URI, proprietà ovvero relazioni e valori, anch'essi identificati da URI.

[2] Il CIDOC CRM è il risultato dell'attività svolta frutto di per oltre 10 anni di lavoro da parte dei dai gruppi di lavoro costituiti nel CIDOC all'interno del in seno al CIDOC ICOM's International Committee for Documentation. Dal 12 settembre 2006, il CIDOC CRM è riconosciuto come standard ISO 21127:2006 e nel dicembre 2014 è stato aggiornato nella nuova versione ISO 21127:2014.

[3] L'ontologia FRBRoo è nata nel 2003 dal gruppo di lavoro di armonizzazione FRBR/CIDOC CRM, che ha avuto lo scopo di allineare i due modelli di dati utilizzati in ambito bibliotecario (FRBR) e museale (CIDOC CRM) al fine di garantirne l'interoperabilità semantica.

UNIMARC dall'OPAC SBN, composti da documenti relativi a monografie moderne, a monografie antiche e a periodici.

La scelta di questi tipi di materiale è stata dettata dalla volontà di mappare i dati comuni a tutti i documenti rimandando a una fase successiva l'analisi e la conversione dei documenti dotati di specificità (Grafica, Cartografia, Musica e Audiovisivi).

Mappatura

Il lavoro di mappatura è stato realizzato a partire dal documento di conversione *UNIMARC Bibliographic Format / SBN MARC*[4], con l'esclusione, come sopra detto, dei tag relativi alle specificità di Grafica, Cartografia, Musica, Audiovisivi e ai dati gestionali.

Nelle operazioni di mappatura, effettuate utilizzando il documento *FRBR object-oriented definition and mapping to FRBR-ER* (version 0.9 draft)[5], sono state create le corrispondenze fra campi, sottocampi e indicatori UNIMARC e uno o più "path" (catene di relazioni) FRBRoo, fornendo contestualmente le tabelle di decodifica utilizzate in SBN.

Ciò ha permesso di sviluppare le procedure di codifica in linguaggio RDF delle informazioni presenti nel catalogo SBN, rendendo i dati in formato "machine readable" e subito condivisibili e consultabili sul Web.

[4] Documento di lavoro interno

[5] http://www.ifla.org/files/assets/cataloguing/frbrrg/frbr-oo-v9.1_pr.pdf

Esempio di mappatura[6]

UNIMARC - FRBRoo MAPPING			
UNIMARC BIBLIOGRAPHIC FORMAT		FRBRoo Mapping	Note
DESCRIZIONE DEL CAMPO	LABEL		
GUIDA DEL RECORD			
Tipo record	**GUIDA**	**F2 Expression** **P2 has type** **E55 Type {Form}**	tabella geun
Livello bibliografico	**GUIDA**	**F2 Expression** **P2 has type** **E55 Type {Form}**	tabella nabi
DATI IDENTIFICATIVI	**0__**		
identificativo del record	**001**	**F3 Manifestation Product Type** **P1 is identified by** **F13 Identifier**	
identificativo di versione	**005**		
		F3 Manifestation Product Type **P129B is subject of** **E33 Linguistic Objecy** **P128B is carried by** **E84 Information Carrier** **P31B was modificed by** **E11 Modification** **P1 is identified by** **E42 Identifier**	
ISBN	**010**		
numero		**F3 Manifestation Product Type** **P1 is identified by** **F13 Identifier**	
qualificazioni		**F3 Manifestation Product Type** **P1 is identified by** **F13 Identifier**	
numero errato		**F3 Manifestation Product Type** **P1 is identified by** **F13 Identifier**	

[6] Sono state riportate solo alcune corrispondenze UNIMARC-FRBRoo

ISSN	**011**		
numero		**F3 Manifestation Product Type** **P1 is identified by** **F13 Identifier**	
qualificazioni		**F3 Manifestation Product Type** **P1 is identified by** **F13 Identifier**	
numero errato		**F3 Manifestation Product Type** **P1 is identified by** **F13 Identifier**	
IMPRONTA	**012**		
Numero		**F3 Manifestation Product Type** **P1 is identified by** **F13 Identifier**	

Normalizzazione, arricchimento e conversione dei dati

1. Normalizzazione dei dati

La conversione dei dati UNIMARC in triple RDF ha richiesto una fase di normalizzazione mirata all'analisi della coerenza e correttezza delle informazioni.

La qualità del dato in entrata rappresenta un aspetto particolarmente importante durante la fase di conversione, perché un dataset non "pulito" può rendere inefficienti o non realizzabili alcune operazioni di confronto, di similitudine e di allineamento dei dati, anche al fine di possibili link con database esterni.

Il problema della normalizzazione dei dati rappresenta un aspetto da analizzare e approfondire ulteriormente, sia ai fini dei collegamenti con risorse esterne, sia per consentire un miglior risultato nelle ricerche mirate dell'utente finale. In particolare, va sottolineato che la produzione di LOD di qualità dipende da un buon livello di pulizia dei dati del catalogo SBN.

2. Arricchimento dei dati

I LOD permettono di arricchire i propri dati e aumentare la fruibilità, la qualità e la quantità delle informazioni associate, attraverso il collegamento a fonti esterne autorevoli che possiedono già una descrizione accurata di numerose entità. Perché ciò avvenga, tuttavia, è necessario che le risorse di una base dati siano facilmente identificabili e collegabili.

RDF definisce una risorsa come un qualsiasi oggetto che sia identificabile univocamente mediante un Uniform Resource Identifier (URI). Nel nostro caso, è stato definito un criterio di generazione delle URI secondo la seguente strutturazione:

- il dominio http://sbn.it come radice della URI
- il riferimento al tipo di risorsa a livello di manifestazione
- l'identificativo univoco del record

La costruzione di URI per le risorse prevede l'implementazione di un sistema di mantenimento di URI permanenti che possano essere facilmente raggiungibili

dall'utente mediante protocollo HTTP. Tali specifiche saranno concordate e implementate in fase di rilascio finale.

Il subset di dati UNIMARC forniti per lo sviluppo del prototipo, già consente di collegare le informazioni presenti in SBN con altre fonti esterne (ad es. il Nuovo Soggettario di Firenze, la Classificazione Decimale Dewey, l'Anagrafe delle biblioteche italiane, l'ISBN, etc.). È stato inoltre deciso di ampliare il numero di collegamenti aggiuntivi, in particolare con i repository di Geonames, Edit16, VIAF e OPAC SBN (per quanto concerne le schede di autorità). Questa decisione ha permesso di arricchire i dati dell'Indice SBN con ulteriori informazioni, ad es. applicando le coordinate geografiche all'entità "Luogo normalizzato" tramite il collegamento a Geonames.

I legami alle entità dei suddetti repository sono stati introdotti in modo puntuale all'interno del prototipo a titolo esemplificativo e sono stati tradotti in triple RDF, memorizzati e consultabili all'interno del triple store.

3. Conversione dei dati

Il set di record bibliografici, una volta normalizzato e arricchito nei campi dimostrativi per il prototipo, è stato convertito in triple RDF attraverso la mappatura FRBRoo, applicando strumenti e procedure sviluppati *ad hoc*.

Pubblicazione

I dati, convertiti in RDF, sono stati caricati all'interno di un triple store per l'accesso ai dati. Per la loro pubblicazione sono state analizzate piattaforme che, oltre al repository RDF e l'endpoint di interrogazione SPARQL, avessero a disposizione interfacce aggiuntive per la costruzione di servizi.

Le piattaforme analizzate sono state Aduna Sesame e OpenLink Virtuoso. Aduna Sesame è stato scelto in quanto framework open source e per l'ottima efficienza nella gestione di database in-memory, considerando anche il numero limitato di record processati dal prototipo.

Le triple RDF disponibili nel triple store Sesame sono accessibili mediante:

- consultazione e download delle risorse sotto forma di file RDF/XML;
- endpoint SPARQL, mediante il quale è possibile ottenere informazioni a seguito dell'inserimento di una query;
- API (Application Programming Interface) che permettono la creazione di servizi aggiuntivi per facilitare le query sul triple store.

Interfaccia di consultazione delle triple

Il prototipo è stato arricchito utilizzando le API di interrogazione delle triple RDF ed è stata realizzata un'interfaccia utente che permette la consultazione delle triple RDF e/o delle query SPARQL anche a utenti non esperti.

Per ogni opera è possibile visualizzare:

- tutte le informazioni bibliografiche disponibili nei campi UNIMARC;
- link esterno alla CDD;
- lista delle biblioteche che possiedono una copia dell'opera;
- mappa relativa alla posizione delle biblioteche, realizzata mediante link al repository Geonames.

L'interfaccia permette la ricerca a partire dall'inserimento di una o più parole chiave, effettuando query semantiche sulle triple RDF, fino a raggiungere le informazioni riguardanti "autori", "titoli" e "soggetti".

A seguito della digitazione dei primi tre caratteri, le tre colonne di ricerca si popolano con tutti i possibili risultati. Continuando la digitazione, l'elenco dei risultati si raffina ulteriormente. Ogni risultato è cliccabile e porta alla relativa pagina informativa.

Dalla pagina "risultato di ricerca" di un **autore** è possibile consultare il collegamento alla scheda dell'autore presente in repository esterni (tra cui, quando possibile, VIAF e OPAC) e la lista di tutte le opere di cui è autore o co-autore presenti all'interno dei dati forniti.

Dalla pagina "risultato di ricerca" di un **titolo** è possibile consultare la lista di tutte le edizioni dell'opera per il titolo selezionato.

Dalla pagina "risultato di ricerca" di un **soggetto** è possibile consultare la lista di tutte le opere che fanno riferimento al soggetto di ricerca.

Ogni risultato è corredato da tutte le informazioni relative alla descrizione bibliografica.

La demo del prototipo per la consultazione e la codifica semantica dei dati bibliografici è stata installata sul server dell'ICCU. Il Gruppo di lavoro sta completando i test, individuando le criticità della mappatura e analizzando i possibili sviluppi che includano ulteriori arricchimenti con link esterni.

La demo sarà pubblicata, terminata la fase di test.

Un interessante sviluppo potrebbe riguardare la mappatura per la sperimentazione in LOD di altre base dati gestite dall'ICCU (es. EDIT16).

Aspetti da definire

Nel corso delle attività sono emerse alcune criticità di diversa natura che riguardano sia i dati del catalogo, sia il modello del CIDOC CRM/FRBRoo che si possono ricondurre a:

- Presenza di dati non normalizzati
- numero limitato di schede di autorità relative agli Autori e, in alcuni casi, mancanza di completezza delle informazioni in esse contenute
- Assenza d'informazioni relative agli altri archivi di autorità (titolo uniforme, luoghi, marche, etc.) le cui schede non vengono attualmente esportate in OPAC
- Insufficiente rappresentatività in SBN di tutte le entità FRBR
- Impossibilità di riferirsi a identificativi persistenti (es. BID e VID) per la natura stessa della catalogazione partecipata che prevede fusioni e cancellazioni dei record da parte delle biblioteche.

Le criticità del modello FRBRoo riguardano essenzialmente l'impossibilità di tradurre la totalità dei dati e dei legami presenti in SBN.

Nel caso dei periodici, ad esempio, non è stato possibile completare la mappatura fra SBN e FRBRoo in quanto nel modello non sono rappresentati i dati relativi ad alcuni legami (tag 423; 431; 434; 447, rispettivamente "Pubblicato con", "Continuazione in parte di", "Ha assorbito", "Si fonde con").

A questo riguardo si attende il consolidamento del lavoro del gruppo costituto in ambito ISSN, che ha elaborato l'evoluzione e l'estensione di FRBRoo alle risorse in continuazione [7] . Il documento prodotto, PRESSoo (versione 1.0) [8], non è ancora uno standard IFLA ma può essere considerato un valido strumento per la mappatura delle relazioni tra risorse in continuazione.

[7] Il gruppo è costituito da: ISSN, International Centre ISSN IC, ISSN Review Group e Bibliotéque National de France, http://www.issn.org/the-centre-and-the-network/our-partners-and-projects/pressoo/

[8] http://www.issn.org/wp-content/uploads/2014/02/PRESSoo_1-0.pdf

Il Gruppo LOD SBN sta inoltre seguendo a livello internazionale altre iniziative quale il progetto di sviluppo della rappresentazione di UNIMARC in RDF presentato nel 2013 al convegno IFLA con il documento "*The UNIMARC in RDF project: namespaces and linked data"* [9].

Gruppo di lavoro LOD SBN

(Gennaio 2015)

[9] http://library.ifla.org/156/1/222-willer-en.pdf

Allegato

Interfaccia di consultazione delle triple

L'interfaccia di consultazione è stata realizzata, a titolo puramente dimostrativo, per provare il possibile funzionamento di un portale per l'accesso ai dati. Tale interfaccia non presuppone la conoscenza specifica, da parte degli utenti utilizzatori, dei metodi di consultazione delle triple RDF e/o delle query SPARQL.

Ricerca delle informazioni riguardanti autori, titoli e soggetti

Fig.1

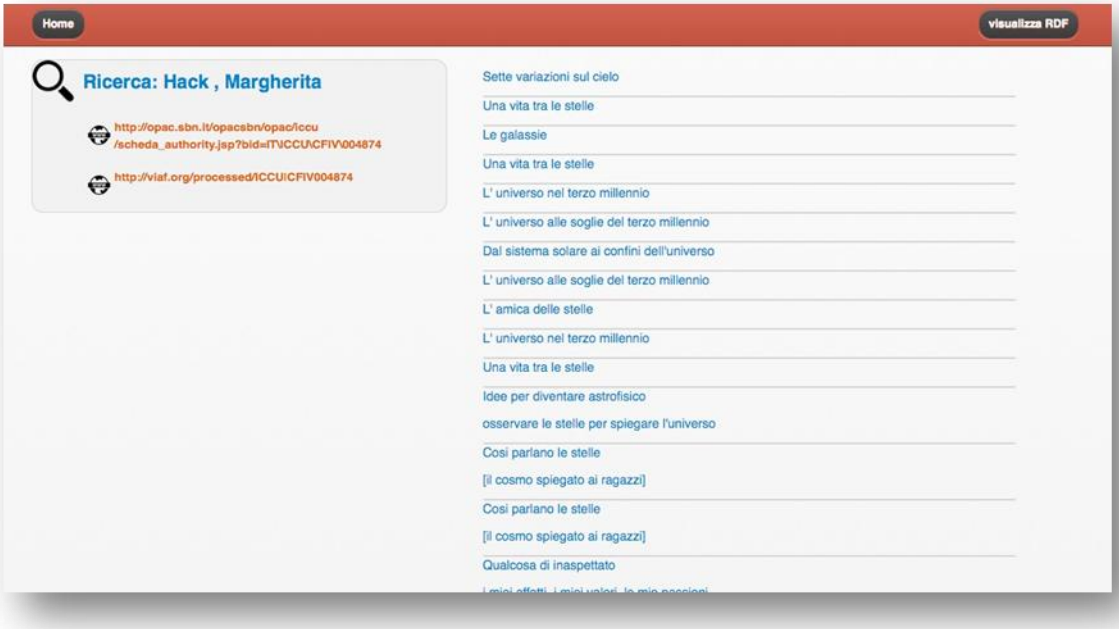

Fig.2

A seguito della digitazione dei primi tre caratteri, le tre colonne di ricerca si popolano con tutti i possibili risultati. Continuando la digitazione l'elenco dei risultati si raffina ulteriormente. Ogni risultato è cliccabile e porta alla relativa pagina informativa.

Tutti i dati visualizzati sono output di query semantiche costruite dall'interfaccia ed effettuate sulle triple RDF contenute nel repository Sesame.

È possibile visualizzare e scaricare l'intera rappresentazione dei dati di esempio in formato RDF tramite l'apposito link "Visualizza RDF", posto nell'angolo in alto a destra dell'interfaccia.

Dalla pagina risultato di ricerca di un autore è possibile consultare (fig.2):

- Colonna sinistra: collegamenti alle schede dell'autore presente in repository esterni (tra cui, quando possibile, VIAF e OPAC). I link esterni a VIAF e OPAC sono stati creati mediante procedura automatizzata sulla base del BID autore presente nei dati UNIMARC originari.

- Colonna destra: lista di tutte le opere di cui il soggetto è autore o co-autore, presenti all'interno dei dati UNIMARC.

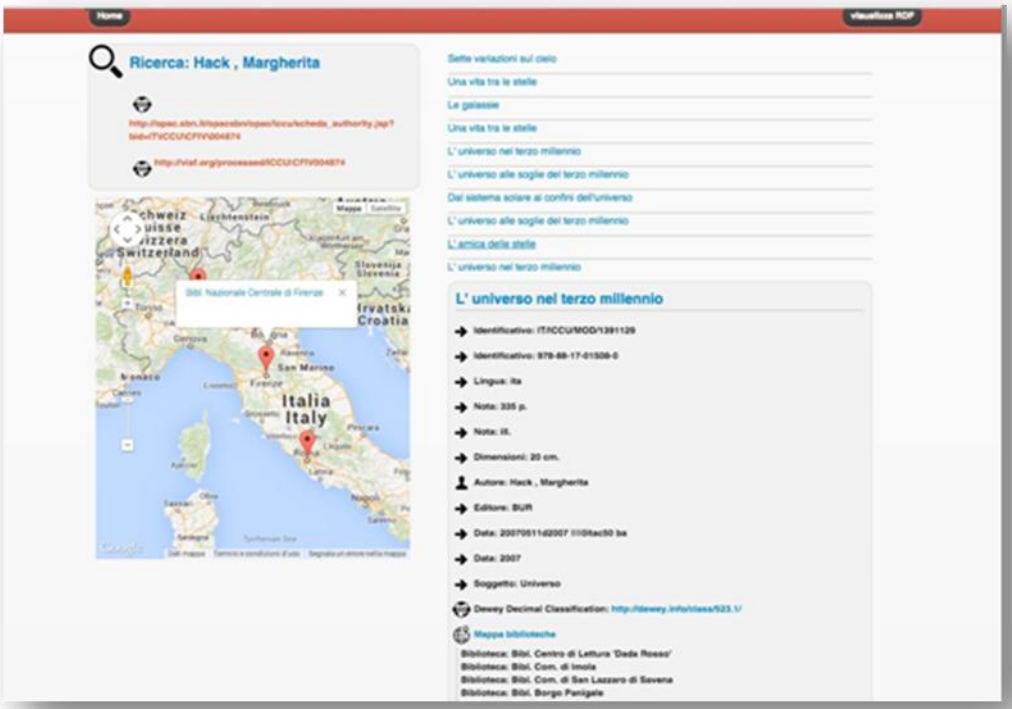

Fig.3

Ogni opera è corredata dalle proprie informazioni bibliografiche, visualizzabili dall'apertura a tendina dei dettagli (ad es. cliccando sul titolo "L'universo nel terzo millennio") (fig.3).

- Soggetto/i, arricchiti automaticamente mediante script ad hoc dai relativi link esterni alla CDD;

- Cliccando sul link "mappa biblioteche" è possibile visualizzare una mappa dimostrativa in cui sono georeferenziate alcune delle città di riferimento (vedi lista dei luoghi georeferenziati).

Ricerca titolo

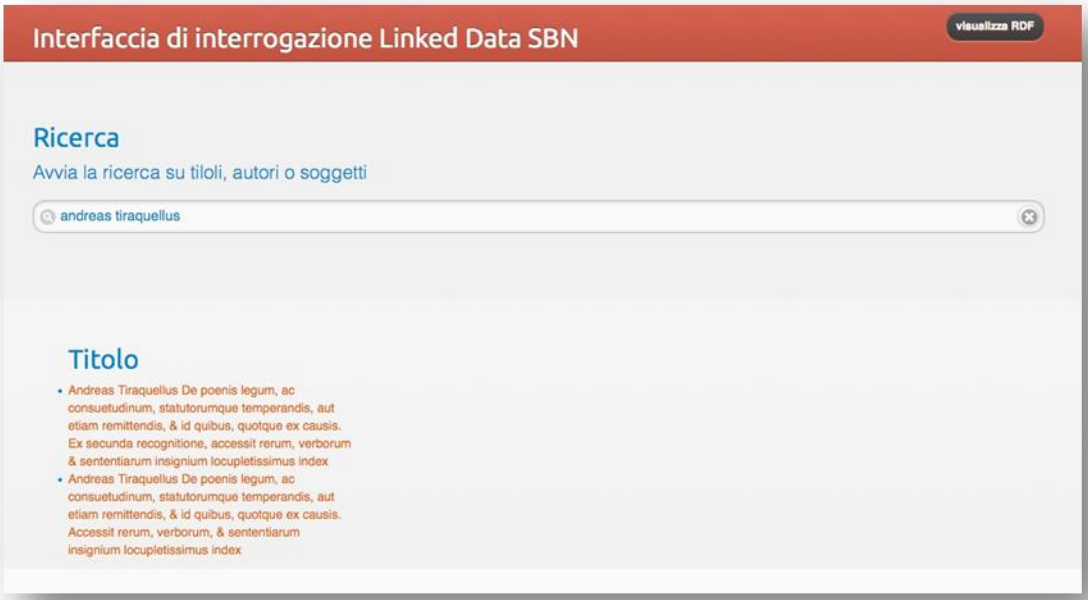

Fig.4

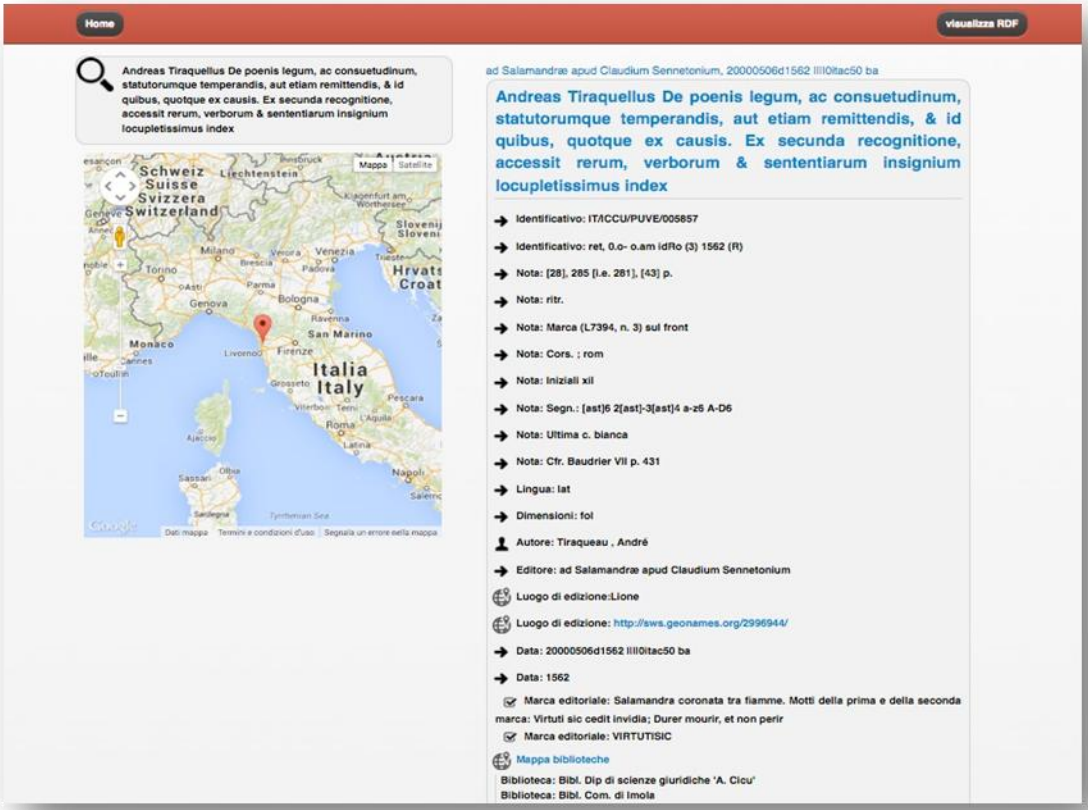

Fig.5

Dalla pagina risultato di ricerca di un titolo è possibile consultare (Fig.5):

- lista di tutte le edizioni dell'opera per lo specifico titolo. La lista delle edizioni è identificata da "editore e data di edizione".

Il campo "data di edizione" è stato ricavato dal campo 100 UNIMARC, indicato come codificato e normalizzato secondo determinate specifiche.

- Cliccando sul ognuno dei record è possibile consultare, all'interno della tendina, le informazioni specifiche della singola "expression".

Lista dei luoghi georefernziati

La lista contente alcuni dei luoghi presenti nell'archivio e georeferenziati, è stata effettuata in modo puntuale, a titolo esemplificativo, su alcuni singoli record.

Alle seguenti Biblioteche, è stato associato il link al servizio offerto dal portale Geonames (http://www.geonames.org/) relativo al centro della città di appartenenza.

Nella pianificazione di un successivo intervento massivo sull'intera base dati, farà riferimento al link Geonames riferito all'esatta posizione della Biblioteca all'interno della città di appartenenza.

Milano Centro città = http://sws.geonames.org/3173435/ = Bibl. Nazionale Braidense;

Bologna Centro città = http://sws.geonames.org/3181928/ = Bibl. Universitaria (MBAC) Bologna;

Roma Centro città = http://sws.geonames.org/3169071/ = Bibl. Nazionale Centrale Vittorio Emanuele II;

Ferrara Centro città = http://sws.geonames.org/6299592/ = Bibl. Fac. di Architettura - Univ. Ferrara;

Venezia Centro città = http://sws.geonames.org/3164603/ = Bibl. centrale IUAV;

Firenze Centro città = http://sws.geonames.org/3176959/ = Bibl. Nazionale Centrale di Firenze;

Rimini Centro città = http://sws.geonames.org/3169361/ = Bibl. Civica Gambalunga di Rimini;

Padova Centro città = http://sws.geonames.org/3171728/ = Bibl. Civica di Padova;

Pisa Centro città = http://sws.geonames.org/3170647/ = Bibl. Universitaria MBAC Pisa;

Palermo Centro città = http://sws.geonames.org/2523920/ = Bibl. Centrale della Regione Sicilia;

Modena Centro città = http://sws.geonames.org/3173331 = Bibl. Civ. A. Delfini;