Archeologia e Calcolatori
Supplemento 8, 2016, 211-216

LE API DI EUROPEANA COME ESEMPIO DI INTEGRAZIONE E RAPPRESENTAZIONE DELLE RISORSE CULTURALI

1. Introduzione

Il dominio dei beni culturali è diventato uno dei più proficui e completi esempi di applicazione delle tecnologie legate ai linked data e al semantic web. Attualmente, è possibile individuare la "cartina tornasole" di questo trend nello sviluppo dei numerosi portali e digital libraries sia a livello nazionale che internazionale e nello sviluppo di repositories orientati alla gestione di grandi quantitativi di metadati. Lo scopo del presente lavoro ha una duplice finalità: da una parte dare un quadro generale su come i principi del semantic web vengano applicati nel campo del cultural heritage per la pubblicazione e fruizione delle collezioni digitali, dall'altra esaminare come le terze parti possano a loro volta utilizzare i dati e metadati messi a disposizione dalle digital libraries, integrandoli in applicazioni web sviluppate utilizzando API messe a disposizione dalle digital libraries stesse. Si prenderà come caso d'uso la digital library Europeana, sviluppando un framework che utilizza l'API key messa a disposizione per la ricerca semantica categorizzata per topics. Si andrà così a definire un'indicizzazione trasversale dei metadati, in modo da permettere, anche a chi non ha oggetti digitali nella digital library o ne ha provveduto indirettamente, tramite un content provider, di integrare i contenuti in terze parti.

2. Web 3.0 e beni culturali

Il web è diventato negli ultimi anni uno dei mezzi più importanti per la pubblicazione e fruizione delle risorse digitali all'interno del dominio dei beni culturali. Ne sono un esempio le biblioteche, gli archivi, i musei che, digitalizzando il proprio patrimonio, lo rendono fruibile attraverso le digital libraries sia nazionali che internazionali che, oltre ad assolvere alla funzione di preservare il cultural object, fungono anche da hub centrale per i diversi repositories. Grazie alle best practices suggerite dal W3C e dall'adozione di sistemi standard in materia di catalogazione e gestione delle risorse, attualmente è possibile riscontrare una completa interoperabilità tra i diversi archivi che si vanno man mano sviluppando, rendendo i contenuti e le risorse human readable nel dominio di interesse dell'utente finale. Questo cambiamento ha comportato lo sviluppo di una nuova architettura basata su relazioni semantiche e risorse relazionate tra di loro, che si arricchiscono utilizzando diversi

tipi di informazioni. Pertanto oggi il web 3.0, rappresenta una fondamentale area di ricerca e sviluppo multidisciplinare applicabile in diversi contesti. Il dominio dei beni culturali è particolarmente incline a sfruttarne le potenzialità di codifica grazie all'eterogeneità dei dati e delle informazioni da gestire.

Al momento, i formalismi propri del semantic web trovano applicazione in:

– sviluppo delle digital libraries, dove, coniugando l'aspetto legato alla fruizione e alla conservazione del patrimonio culturale digitale, affrontano il problema della gestione standardizzata e interoperabile di milioni di dati seguendo precisi data schema e formalismi;

– sviluppo di piattaforme Linked Open Data (LOD), dove, attraverso un end-point, devono garantire libero accesso e riutilizzo semantico e non dei dataset archeologici messi a disposizione. I LOD in archeologia trovano la loro ragione nel fatto che un dataset ristretto ad un contesto preciso può essere arricchito con le altre risorse afferenti allo stesso dominio e presenti in altri end-point.

3. Le API nelle digital Humanities

Al fine di garantire l'integrazione in terze parti delle risorse gestite sia dalle digital libraries che dagli SPARQL end-point, uno dei metodi attualmente più utilizzati è quello dell'Application Programming Interface (API), interfacce per esprimere i contenuti on line sia su piattaforme desktop che mobili. Contestualmente al dominio dei beni culturali, assistiamo a due diversi utilizzi delle API, uno passivo ed uno attivo. Quello passivo si verifica nel caso in cui vengano integrate delle API sviluppate da terzi in altre applicazioni web, siti o portali. Ne sono un esempio i siti di musei, archivi o digital libraries che integrano le API per contribuire alla condivisione dei contenuti.

L'utilizzo attivo si verifica, invece, nel caso in cui progetti relativi all'utilizzo e allo sviluppo di risorse culturali e digitali provvedano essi stessi allo sviluppo di end-point raggiungibili utilizzando il protocollo REST per integrare i contenuti in terze parti. In questo senso ne sono un esempio soprattutto le digital libraries che provvedono all'output dei dati e metadati in formati interoperabili, come ad esempio JSON, XML o RDF. Possiamo trovare alcuni esempi in:

– Cultura Italia che, utilizzando l'end-point http://dati.culturaitalia.it/, permette l'interrogazione SPARQL e il riutilizzo dei metadati caricati sul portale.

– Amsterdam Museum che, utilizzando l'end-point http://semanticweb.cs.vu.nl/europeana/user/query/, mette a disposizione, tramite il formalismo SPARQL, i metadati caricati dalla digital library Europeana arricchiti semanticamente con altri vocabolari specializzati, come ad esempio quelli esposti dal Getty Museum.

– Europeana, la digital library europea che, utilizzando sia l'end-point REST che SPARQL, mette a disposizione i metadati riferiti agli oggetti digitali in un formato standard e interoperabile, utilizzando l'ontologia Europeana Data Model (EDM).

4. La digital library EUROPEANA

Europeana rappresenta forse uno dei casi più conosciuti di integrazione ed esposizione di metadati in un formato strutturato semanticamente[1]. In questo caso, il processo di arricchimento semantico avviene direttamente sul repository della digital library dove i metadati, strutturati seguendo un profilo Dublin Core, vengono semanticamente definiti con l'ontologia Europeana Data Model (EDM), profilo applicativo di CIDOC-CRM. Segue principalmente una struttura RDF, sviluppata sulla base di triple e arricchita con classi derivate direttamente dal Dublin Core e dall'Open Archives Initiative Object Reuse and Exchange (OAI-ORE) e dall'utilizzo del formalismo SKOS per l'indicizzazione e la rappresentazione dei metadati secondo le quattro classi: who (edm:Agent); where (edm:Place); when (edm:TimeSpan); what (skos:Concept). I metadati sono resi disponibili utilizzando le API http rest o lo SPARQL end-point. Nel primo caso si avrà come output un file .json, mentre nel caso dell'end-point SPARQL un file .RDF.

4.1 *Ragionando sull'EDM*

Nonostante la flessibilità e l'interoperabilità del sistema, l'EDM presenta ancora alcune carenze dal punto di vista sia della concettualizzazione dei dati sia della definizione dei metadati che, talvolta, sono troppo sintetici e poco esplicativi.

4.1.1 Edm:type

La prima fonte di confusione è rappresentata dal concetto di media type (edm:type) che discrimina la maggior parte delle queries all'interno di una qualsiasi digital library. La logica imposta dall'EDM è quella di distinguere tra descrizione dell'oggetto culturale e la sua rappresentazione digitale. Se, ad esempio, facciamo una ricerca per "Colosseo", come risposta avremo diversi record: alcuni riferiti alle immagini che raffigurano il Colosseo, altri che riportano risorse testuali relative al Colosseo, etc. Tuttavia, se esaminiamo le risorse testuali, ci possiamo rendere conto che in realtà non sono risorse testuali pure, ma immagini del testo relative al Colosseo.

[1] Il progetto è attualmente è caratterizzato da 120 aggregatori di cui fanno parte 2300 tra musei, biblioteche, archivi e collezioni audio-visive e da 29.000.000 di oggetti digitali caricati.

4.1.2 Definizione multi livello

Tale definizione interessa la distinzione che viene fatta tra risorsa web e oggetto culturale che può essere specificata nella distinzione tra oggetto e soggetto del record descritto e come questo viene mappato. Facendo sempre riferimento al Colosseo, facendo una ricerca libera per "Piranesi" e "Colosseo", il risultato saranno 38 record (immagini). In alcuni casi, il record viene mappato e descritto prendendo come soggetto l'immagine digitale, in altre viene preso come soggetto l'oggetto ritrattato nell'immagine.

Come è possibile vedere dalla tabella, nel primo caso ci troviamo davanti ad un oggetto fisico, dove non è specificato il "creator" e come data di creazione è riportato il 1930, anno in cui è stata riprodotta fotograficamente l'incisione del Piranesi. Nel secondo esempio, invece, l'oggetto viene indicato come Stampa, fatta da Giovanni Piranesi nel 1776. Infine, nell'ultimo caso, il type viene definito come anfiteatro con il "creator" omesso e come data viene indicata quella del 70/80 d.C. Si tratta di diversi livelli di concettualizzazione e astrazione dell'oggetto digitale che, pur definendo lo stesso item (Colosseo), usano una definizione dei metadati diversa portando, spesso, a delle interpretazioni fuorvianti.

4.1.3 Arricchimento dei metadati

Uno degli scopi principali dell'EDM e del flusso di gestione dati di Europeana è quello di arricchire semanticamente i metadati forniti dai content provider e permettere l'accesso al content provider in modo da averli indietro secondo la nuova definizione EDM. L'operazione, tuttavia, ha in sé due problemi di diverso ordine, ma collegati. Esaminiamo come scenario un ipotetico Museo che vuole recuperare i dati arricchiti in EDM. Si rileva una difficoltà nell'acquisizione dei dati, in quanto Europeana ancora non ha provveduto ad un sistema che permetta l'harvesting inverso, ovvero da Europeana a content provider. L'API SPARQL-end-point potrebbe essere una soluzione. Inoltre, si nota anche poca trasparenza nel processo di arricchimento, non tanto dal punto di vista tecnico-funzionale, ma piuttosto da quello concettuale. Tale problematica emerge particolarmente quando il Museo in questione, che ipoteticamente ha fornito un solo record ad Europeana, voglia estendere il data model a tutti gli altri record del repository.

5. Il framework

Considerando quanto esposto sopra, è possibile trovare una parziale soluzione impiegando il linguaggio di programmazione Python e, in particolare, la libreria Flask[2]. Europeana espone i metadati utilizzando due protocolli: l'HTTP REST e lo SPARQ end-point. Nel caso di HTTP REST i metadati

[2] Flask è un framework basato su Werkzeug e Jinja2; il primo è una libreria per la gestione del protocollo WSGI, mentre il secondo è un motore di templating molto utilizzato e performante.

vengono esposti con un formato JSON e rispecchiano, sia a livello di fruizione che a livello di rappresentazione, quella che è la struttura standard dei metadati di Europeana, mentre nel caso dell'end-point SPARQL vengono esposti arricchiti semanticamente in EDM. Grazie alla scalabilità e all'interoperabilità offerta da Python, è stato possibile sviluppare un web framework in Flask che, partendo da una chiamata REST sull'HTTP offerto da Europeana (http://www.europeana.eu/api/v2/), è in grado di integrare, utilizzando la libreria rdflib, altri end-point SPARQL esterni, come ad esempio DbPEDIA; di indicizzarne i dati all'interno del repository in modo da provvedere ad un allineamento con l'EDM; e infine di normalizzare i contenuti con altri dello stesso dominio, ma provenienti da fonti dati diverse, nonché di arricchire il contenuto dei metadati.

Trattandosi di un proof of concept, l'architettura logica e il codice dell'applicativo sono stati resi disponibili attraverso la piattaforma GitHub (http://www.github.com/matteoLorenzini/), con l'idea di rendere il progetto collaborativo e in costante implementazione.

6. Conclusioni

La digitalizzazione e la catalagazione degli oggetti culturali sono ormai diventate una costante in quanto strettamente legate allo sviluppo di digital libraries. Tuttavia, tale approccio comporta una serie di problemi legati alla codifica e all'utilizzo dei dati che viene fatto sia da parte degli utenti che da parte degli sviluppatori nelle fasi di produzione e sviluppo. Il presente lavoro ha analizzato in particolare la digital library Europeana e i problemi ad essa legati e per lo più dipendenti dalla complessità del modello dati che viene usato. Concentrandosi sull'integrazione dei dati, si è evidenziato come, utilizzando le API fornite, i contenuti di Europeana, possano essere fuiti in maniera più proficua provvedendo ad una consultazione che vada al di là della semplice visualizzazione desktop based.

Rimangono invece aperti i problemi relativi alla strutturazione dei metodi di ricerca all'interno della DL. L'attuale implementazione prevede queries strutturate solo all'interno dello schema dati usato per l'indicizzazione. La soluzione, a mio parere, è rappresentata dall'utilizzo dello SPARQL end-point lanciato in versione stabile a novembre del 2013 che, strutturando l'informazione in RDF/EDM, permette una sofisticazione delle queries maggiore rispetto a quella classica fornita dal sistema di indicizzazione, come è stato possibile osservare integrando il framework sviluppato.

Matteo Lorenzini
Österreichische Akademie der Wissenschaften
Austrian Center for Digital Humanitites (ACDH)
matteo.lorenzini@oeaw.ac.at

BIBLIOGRAFIA

Aloia N., Debole F., Gavrillis D., Meghini C., Papatheodoru C. 2014, *Describing research data: A case study for archaeology*, in R. Meersman, H. Panetto, T. Dillon, M. Missikoff, L. Lin, O. Pastor *et al.* (eds.), *OTM 2014 Conferences*, Berlin, Springer-Verlag, 768-775.

Doerr M., Gradmann S., Hennicke S., Isaac A., Meghini C., Van de Sompel H. 2010, *The Europeana Data Model (EDM)*, in *World Library and Information Congress*: *76th IFLA General Conference and Assembly*, Gothenburg, IFLA Publications, 10-15.

Haslhofer B., Isaac A. 2011, *Data.europeana.eu: The Europeana linked Open Data pilot*, in *International Conference on Dublin Core and Metadata Applications*, Amsterdam, IOS Press, 94-104.

Schröttner M., Havemann S., Theodoridou M., Doerr M., D.W. Fellner 2012, *A generic approach for generating cultural heritage metadata*, in M. Ioannides, D. Fritsch, J. Leissner, R. Davies, F. Remondino, R. Caffo (eds.), *Euromed*, Berlin, Springer-Verlag, 231-240.

Theodoridou M., Tzitzikas Y., Doerr M., Marketakis Y., Melessanakis V. 2010, *Modeling and querying provenance by extending CIDOC CRM*, «International Journal Distributed and Parallel Databases», 169-210.

ABSTRACT

During the last years, the production of digital datasets in the Cultural Heritage domain, has seen an exponential increase and the structured repositories have become the most used infrastructure for knowledge management and consultation with different kinds of systems and platforms ensuring a complete interoperability and reach ability of data. Furthermore, the dataset products represent a great resource from which it is possible to extract and deduce new knowledge. In cultural heritage, thanks to the technology related to semantic web, we are able to manage and enrich our data using formalisms and data standards: digital libraries and digital archives, SPARQL-endpoint are some examples. This work, starting from the analysis of Europeana's data model, discusses the integration and use of semantic data in third parts using the API system as a framework.