

Ottimizzare i Metadati Geospaziali per il Web Semantico

Cristiano Fugazza (*), Monica Pepe (*), Alessandro Oggioni (*), Paolo Tagliolato (*, **, ***),
Fabio Pavesi (*), Anna Basoni (*), Stefano Menegon (**), Paola Carrara (*)

(*) CNR - IREA, via Bassini 15 - 20133 Milano, tel +39 0223699456

{fugazza.c, pepe.m, oggioni.a, tagliolato.p, pavesi.f, basoni.a, carrara.p}@irea.cnr.it

(**) CNR ISMAR Arsenal - Tesa 104, Castello 2737/F, 30122 Venezia, tel. +39 041 2407927

stefano.menegon@ismar.cnr.it

(***) LifeWatch Italy

Abstract

In the geospatial realm, data annotation and discovery rely on a number of ad-hoc formats and protocols. Metadata are at the heart of the discovery process and nevertheless they are often neglected or encoded in formats that either are not aimed at efficient retrieval of resources or are plainly outdated. Instead, our workflow for metadata management envisages i) editing via customizable web-based forms, ii) encoding of records in any XML application profile, iii) translation into RDF (involving the semantic lift of metadata records), and finally iv) publication as Linked Data and back-translation into the original XML format with added discovery and internationalization capabilities.

Sommario esteso

Nel dominio dei dati geospaziali, l'annotazione e la ricerca (*discovery*) delle risorse si avvalgono di formati e protocolli appositamente concepiti. Essi sono progettati al fine di abilitare il reperimento delle risorse per casi d'uso dominio-specifici, per i quali i meccanismi di ricerca generica non sono praticabili. La metadatatazione è ancora sostanzialmente dominata da codifiche XML, secondo diversi profili: in particolare, per i dati geografici ci si riferisce agli standard ISO 19115/19119, ripresi ed adattati all'interno della normativa INSPIRE e, per quanto riguarda le informazioni relative ai sensori, al SensorML. Tuttavia, sebbene i metadati rappresentino il cardine stesso dei processi di *discovery*, essi sono spesso trascurati e i formati in cui sono espressi, rappresentano soluzioni sub-ottimali.

In particolare, i formati attualmente in uso non abilitano una ricerca fine, in grado di superare le barriere di internazionalizzazione o di eterogeneità semantica, anche se alcune tecnologie lo renderebbero possibile. Purtroppo, infatti, il salto quantico rappresentato dall'introduzione delle pratiche denominate *Linked Open Data* (LOD) non ha sinora prodotto un riferimento coerente e ben interconnesso nel dominio geospaziale. A titolo esemplificativo, si pensi come i dati sperimentali, la letteratura scientifica basata su di essi ed i ricercatori responsabili o detentori di questi prodotti siano fra loro solo debolmente correlati all'interno dei metadati. Questi ultimi sono spesso espressi in formati intelligibili unicamente ad un operatore umano o duplicati in sistemi distinti e raramente fra loro normalizzati.

Qui proponiamo, nel contesto delle *Spatial Data Infrastructure* (SDI), un flusso di lavoro per la creazione di metadati (MD) che contempli anche informazioni e formati in grado di abilitare una annotazione più raffinata ed espressiva, attraverso un arricchimento semantico applicabile sia a layer geografici tradizionali che ad osservazioni raccolte da sensori. Riassunto in breve, tale flusso di lavoro per la gestione dei MD prevede: i) la compilazione del MD attraverso una interfaccia Web personalizzabile per un dato *application profile*; ii) la sua codifica XML secondo il profilo stesso; iii) una traduzione in *Resource Description Framework* (RDF), atta ad includere l'arricchimento

semantico dei MD; e infine iv) la loro pubblicazione come *Linked Data* e serializzazione da RDF nel formato XML originale con funzionalità aggiuntive per la *discovery* e l'internazionalizzazione. In particolare, la fase iii) consiste nel mettere in relazione i MD con strutture dati RDF che rappresentino parole-chiave da vocabolari controllati, toponimi, ma anche ricercatori, istituti e, in linea di principio, una qualsiasi rappresentazione che si può ritrovare (o direttamente pubblicare) nella nuvola dei LOD.

Per esemplificare la procedura e spiegarne l'utilità, l'abbiamo applicata ad un caso d'uso specifico realizzato all'interno di un catalogo di metadati tradizionale. Il cambio di afferenza di un ricercatore riferito nei MD rende questi ultimi inconsistenti fintantoché opportune (ed onerose) operazioni di aggiornamento non vengono eseguite. Al contrario, una rappresentazione dei MD orientata alla semantica consente di semplificare, e in alcuni casi automatizzare del tutto, queste attività dal momento che il MD può essere generato all'occorrenza, recuperando le informazioni più aggiornate riguardo le entità che sono riferite.

Nel flusso di lavoro per la creazione di MD, si utilizza uno strumento *open-source*, composto di un servizio web di *back-end* (EDI-NG server) e di un cliente HTML5/javascript (EDI-NG client) appositamente sviluppati nel Progetto RITMARE (<http://www.ritmare.it/sp7>). Il servizio EDI-NG server è guidato da un *template*, ovvero un modello che rappresenta il profilo di MD richiesto per il relativo *encoding* XML e per la sua traduzione in RDF. Il *template*, una volta richiamato attraverso il client, crea dinamicamente un'interfaccia utente a *form*, semplificata rispetto alla complessità del formato di MD. EDI-NG è stato concepito per connettersi a fonti di dati esterne eterogenee, tipicamente vocabolari controllati o altre strutture dati RDF (Fugazza et al., 2014a), che forniscono definizioni autorevoli e consistenti per alcuni elementi di MD, come ad esempio le parole chiave ed i punti di contatto. Grazie a queste connessioni e a metodi formalizzati nel *template*, EDI-NG client supporta da un lato l'utente nella compilazione facilitata dei campi di MD e, dall'altro, è in grado di supportare sia la codifica XML conforme al profilo INSPIRE che la sua traduzione in formato RDF. In particolare, quest'ultima contiene gli identificatori univoci (URI) alle entità complesse riferite nel MD, quali le parole chiave provenienti da vocabolari controllati, espressi come thesauri SKOS, ed i punti di contatto, espressi come strutture dati FOAF (Fugazza et al., 2014b).

In questa visione innovativa, abilitata dalla rappresentazione semantica, è più facile garantire la consistenza del MD attraverso la generazione *on demand* del metadato nello specifico formato XML prescritto. Il fine ultimo è rappresentato da una gestione distribuita del MD, in linea con l'architettura distribuita del Web e, in particolare, del Web Semantico.

Ringraziamenti

Questo lavoro è stato finanziato nell'ambito del Progetto bandiera del MIUR RITMARE e di LifeWatch Italia.

Riferimenti bibliografici

- Fugazza C., Oggioni A., Pepe M., Pavesi F., Carrara P. (2014a), "A holistic, semantics-aware approach to Spatial Data Infrastructures". In *Proceedings of DATA 2014, Vienna (Austria), 29-31 August 2014*, 349-356.
- Fugazza C., Basoni A., Menegon S., Oggioni A., Pavesi F., Pepe M., Sarretta A., Carrara P. (2014b), "RITMARE: Semantics-Aware harmonisation of data in Italian marine research", In *Proceedings CRIS2014, Rome 13-15 May 2014 Procedia computer science, Elsevier, Amsterdam*, 33: 261 - 265 (DOI: 10.1016/j.procs.2014.06.041).