Rappresentare Ricercatori e Istituti del Progetto RITMARE nel Web Semantico

Cristiano Fugazza, Alessandro Oggioni, Monica Pepe, Paola Carrara

Istituto per il Rilevamento Elettromagnetico dell'Ambiente
IREA – CNR
via Bassini 15 - 20133 Milano, tel +39 02 23699456
{fugazza.c, oggioni.a, pepe.m, carrara.p}@irea.cnr.it

Sommario Nell'organizzazione dell'Infrastruttura per Dati Spaziali (SDI) del Progetto Bandiera RITMARE, destinata a raccogliere i contributi delle diverse realtà di Ricerca coinvolte nel progetto, un ruolo centrale è costituito dalla rappresentazione di Ricercatori e Istituti. Al fine di poter facilmente integrare questa rappresentazione con la restante informazione di contesto utilizzata nell'infrastruttura (ad esempio, vocabolari controllati, gazetteer, metadati) è stato utilizzato il formato di dati Friend Of A Friend (FOAF), che è basato sul Resource Description Framework (RDF), il modello di dati sottostante il Web Semantico.
Questo tipo di rappresentazione consente, da un lato, di strutturare l'afferenza di Ricercatori a Istituti, dettagliare finemente ognuna di queste entità e rappresentare le loro mutue relazioni. Dall'altro, consente di integrare queste strutture dati nel contesto più generale del Web Semantico e, nello specifico, delle risorse rese disponibili a livello mondiale come Linked Data, la recente metodologia per la fruizione di dati indipendente dal protocollo di accesso.
In questo lavoro descriviamo la metodologia di rappresentazione utilizzata ed i vantaggi che essa comporta nella creazione degli strumenti web che il sottoprogetto 7 di RITMARE sta predisponendo, sia lato back-end che lato front-end.
Parole chiave: Web Semantico, Friend Of A Friend, FOAF, ricercatori, identificatori

Abstract The Spatial Data Infrastructure (SDI) of the RITMARE Flagship Project is aimed at conflating the distinct contributions to Italian marine research. In implementing the infrastructure, a pivotal element is the representation of Researchers and Institutes. For a seamless integration with the heterogeneous context information involved (controlled vocabularies, gazetteers, metadata), we employed the Friend Of A Friend (FOAF) data schema, which is based on the Resource Description Framework (RDF), the data format underlying the Semantic Web.
On the one hand, this kind of representation allows to relate Researchers with the respective Institutes, to detail these entities, and to express their mutual relationships. On the other, it is possible to integrate these data structures with the wider context of the Semantic Web, particularly with the resources made available as Linked Data, a recent paradigm for protocol-neutral access to resources. In this paper, we describe this representation methodology and outline its advantages for the creation of the web-based tools that the RITMARE sub-project 7 is developing for both front- and back-end functionalities.
Keywords: Semantic Web, Friend Of A Friend, FOAF, researchers, identifiers

1 – Introduzione:

Una descrizione esaustiva delle entità in gioco nel dominio di interesse è una caratteristica essenziale di tutti i sistemi informativi; tale descrizione tipicamente assume la forma di una base di dati, sia che questa risponda ad un modello relazionale, ad uno schema definito tramite l'eXtensible Markup Language[1] (XML) oppure, come nel caso delle strutture dati descritte in questo lavoro, al modello di dati costituito dal Resource Description Framework[2] (RDF), la "lingua franca" con la quale viene rappresentata l'informazione nel Web Semantico.

Quest'ultimo paradigma consente tuttavia una maggiore apertura rispetto alla fruizione libera dell'informazione. Gli schemi relazionali sono infatti strettamente proprietari (ogni realtà applicativa definisce cioè il proprio schema) e le relazioni tra le diverse entità da essi espresse risultano evidenti solo nelle interrogazioni (le *query* SQL) che vengono definite. Gli schemi XML risultano invece, come il nome suggerisce, pienamente estensibili ma sono nondimeno proprietari per quanto riguarda i singoli schemi che vengono combinati ed estesi; in questo caso, le relazioni tra le diverse entità restano evidenti ma implicano comunque la comprensione di come queste vengano rappresentate all'interno degli schemi utilizzati.

RDF consente invece di definire strutture dati eterogenee, di associare loro identificativi univoci, Uniform Resource Identifiers[3] (URI) del tutto analoghi ai tradizionali indirizzi Internet, mantenendo escplicite le relazioni tra le entità definite. Schemi proprietari sono definibili tramite opportune estensioni ma il loro legame con i costrutti-base definiti da RDF garantisce comunque un livello minimo di comprensione a prescindere dalla conoscenza degli schemi utilizzati. La filosofia di accesso nota come Linked Data rappresenta il culmine di questo processo di apertura in quanto l'informazione diviene direttamente accessible attraverso il suo identificatore, lo URI, tramite il protocollo comunemente utilizzato per le pagine Web.

L'infrastruttura di dati sviluppata dal sottoprogetto 7 del Progetto RITMARE (il Progetto bandiera finanziato dal *Ministero dell'Istruzione, dell'Università e della Ricerca* che mira ad integrare i risultati della ricerca marina italiana), utilizza RDF come formato di rappresentazione dei *metadati*, ovvero dei descrittori delle risorse geospaziali che ci si propone di aggregare e che risultano essenziali per il reperimento e l'indicizzazione delle risorse stesse da parte di motori di ricerca specializzati denominati *geoportali*. Tale approccio è in accordo con le *Linee guida per l'interoperabilità semantica attraverso i Linked Open Data* espresse nel documento (Commissione di Coordinamento SPC, 2012).

Nello specifico, questo lavoro descrive la metodologia di rappresentazione che è stata adottata per quanto riguarda l'anagrafica dei Ricercatori e degli Istituti coinvolti nel progetto. Sebbene quest'informazione sia in principio distinta dai metadati sopracitati, disporre di tale rappresentazione in formato RDF consente di semplificare il processo di creazione dei metadati, che debbono essere confacenti gli obblighi di Legge dettati dalla normativa *Repertorio Nazionale dei Dati Territoriali*[4] (RNDT) (Agenzia per l'Italia Digitale, 2014), il recepimento italiano della Direttiva INSPIRE[5] vigente nell'Unione Europea (Commissione Europea, 2007).

Inoltre, tale rappresentazione orientata alla semantica del panorama di fornitori di dati e utenti RITMARE consente di realizzare paradigmi di accesso e fruizione delle risorse maggiormente evoluti rispetto allo stato dell'arte. Ad esempio, queste strutture dati consentono la creazione di un portale di accesso ai dati altamente personalizzabile sulla base dei profili degli utenti finali. Lo stesso accesso ai dati (aperto, secondo la *data policy* vigente fino alla fine del progetto) potrà facilmente essere condizionato sulla base di specifiche relazioni intercorrenti tra Ricercatori e Istituti.

1 eXtensible Markup Language (XML): http://www.w3.org/XML/
2 Resource Description Framework (RDF): http://www.w3.org/RDF/
3 Uniform Resource Identifier (URI): http://tools.ietf.org/html/rfc3986
4 Repertorio Nazionale dei Dati Territoriali (RNDT): http://www.rndt.gov.it/RNDT/home/index.php
5 INSPIRE: http://inspire.ec.europa.eu/

Questo lavoro è organizzato come segue: la Sezione 2 introduce lo schema di dati utilizzato nella rappresentazione dell'anagrafica del progetto RITMARE; la Sezione 3 descrive la struttura dati creata per l'anagrafica di progetto; la Sezione 4 presenta alcuni esempi di utilizzo di tale informazione per la metadatazione e per l'accesso alle risorse; infine, la Sezione 5 trae le conclusioni e presenta le attività future.

2 – Il formato Friend Of A Friend e l'anagrafica di RITMARE

Sin dalla sua iniziale formulazione, risalente all'anno 2000, il formato di dati Friend Of A Friend[6] si propone di consentire l'integrazione di descrittori decentralizzati che rappresentano persone, loro aggregazioni multilivello e istituzioni. Tali descrittori possiedono infatti un identificatore univoco, uno URI, che identifica la sequenza di *domini Internet* (ad esempio, `sp7.irea.cnr.it`) e lo specifico *percorso* (ad esempio, `/rdfdata/project/AlessandroOggioniIREA`) mediante il quale è possibile raggiungere l'informazione rappresentata dallo URI mediante la giustapposizione delle sue componenti. Ad esempio, il Ricercatore Alessandro Oggioni dispone del proprio descrittore FOAF all'indirizzo:

```
(1)    http://sp7.irea.cnr.it/rdfdata/project/AlessandroOggioniIREA
```

Agenti umani o automatizzati possono quindi accedere all'informazione corrispondente in modo diretto attraverso il protocollo HTTP sul quale si fonda la tradizionale navigazione del Web, talvolta specificando il formato di dati che si desidera ricevere. Infatti, questa informazione può essere codificata in molteplici modi, quali ad esempio una normale pagina HTML destinata alla consultazione (a volte contenente la medesima informazione in modalità *machine-friendly* attraverso opportune annotazioni trasparenti all'utente umano) piuttosto che in uno dei molteplici formati nei quali è possibile esprimere il modello di dati RDF, come ad esempio il Terse RDF Triple Language[7] (Turtle) utilizzato in questo articolo. Il Listato 1 mostra parte dell'informazione associata ad Alessandro Oggioni e reperibile all'indirizzo sopracitato.

```
01 @prefix rdf:         <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
02 @prefix xsd:         <http://www.w3.org/2001/XMLSchema#> .
03 @prefix foaf:                <http://xmlns.com/foaf/0.1/> .
04 @prefix vcard2006:   <http://www.w3.org/2006/vcard/ns#> .
05 @prefix ns1:         <http://sp7.irea.cnr.it/rdfdata/project/> .
06 @prefix ns2:         <http://ritmare.it/rdfdata/thesauri/domains#> .
07 ns1:AlessandroOggioniIREA    rdf:type        foaf:Person ;
08      foaf:firstName  "Alessandro"^^xsd:string ;
09      foaf:familyName "Oggioni"^^xsd:string ;
10      vcard2006:email <mailto:oggioni.a@irea.cnr.it> ;
11      vcard2006:org           ns1:IREA ;
12      foaf:knows              ns1:CristianoFugazzaIREA ;
13      foaf:topic_interest ns2:Ecosistemi .
14 ns1:SP7_WP1_AZ1_U002 foaf:member     ns1:AlessandroOggioniIREA .
15 ...
```

Listato 1 □Descrittore FOAF associato ad uno degli autori.

RDF (su cui FOAF si basa) suddivide l'informazione in sequenze di *triple*, ovvero di terne soggetto-predicato-oggetto; è bene osservare quanto una rappresentazione così semplificata dell'informazione riesca in realtà ad esprimere strutture dati molto complesse. Nel codice Turtle riportato nel Listato 1, una tripla termina con il punto ". ", come alla linea 14. Quando invece il soggetto della tripla

6 Friend Of A Friend (FOAF): http://www.foaf-project.org/
7 Terse RDF Triple Language (Turtle): http://www.w3.org/TeamSubmission/turtle/

resta lo stesso in una serie di triple, si utilizza il punto e virgola ";", come alle linee 7-12 (in questi casi, per facilitare la lettura, nel listato si è anche utilizzata l'indentazione).
Il significato del Listato 1 è il seguente. Le righe 1-6 definiscono *prefissi* che vengono utilizzati nel seguito del listato per accorciare la rappresentazione delle entità in gioco; ad esempio, per accorciare lo URI in (1) in `ns1:AlessandroOggioniIREA`. La riga 7 definisce a quale classe di entità appartiene quest'ultimo, in questo caso la classe `Person` definita da FOAF. Le righe 8 e 9 forniscono una rappresentazione letterale dell'URI che, in principio, potrebbe contenere una serie di identificativi numerici anziché la rappresentazione maggiormente mnemonica che è stata scelta. Infine, le righe 10-14 realizzano quello che è considerato da molti lo scopo ultimo di una rappresentazione dei dati in RDF, ovvero la correlazione del dato in esame con altre entità del Web Semantico che lo circonda.
Nello specifico, la riga 10 associa il Ricercatore con il rispettivo indirizzo e-mail; la riga 11 lo associa all'istituto di appartenenza; la riga 12 rappresenta il legame tra due Ricercatori nella rete sociale che l'insieme di descrittori FOAF del progetto sottende; la riga 13 definisce invece quale sia il dominio di ricerca dell'individuo descritto. Infine, la riga 14 associa il ricercatore ad una specifica *unità operativa*, di una determinata *azione* all'interno di uno specifico *work package* del *sottoprogetto* 7 del *progetto* RITMARE. Non è superfluo notare come, sebbene nelle ultime quattro triple vengano riferite entità appartenenti al medesimo "spazio dei nomi" (*namespace*), e cioè quello definito dal sottoprogetto 7 per i propri dati, queste potrebbero virtualmente riferire entità all'esterno del progetto e non necessariamente sotto il controllo di quest'ultimo. Questo esemplifica uno dei gradi di libertà che consente al Web Semantico di crescere in maniera indipendente, decentralizzata e pienamente esplicita.

3 – La struttura dati per l'anagrafica del progetto RITMARE

La categorizzazione delle entità coinvolte in RITMARE è stata ricavata a partire dalla documentazione fornita dall'amministrazione del progetto come foglio di calcolo; tale documentazione è stata esportata come documento XML ed elaborata attraverso un foglio di stile espresso come eXtensible Stylesheet Language Transformations[8] (XSLT). Questo ha consentito di ricavare strutture dati grezze nel formato RDF/XML, cioè un formato di serializzazione di RDF, alternativo al formato utilizzato nel Listato 1, che utilizza XML per la sua rappresentazione.
Ricercatori e Istituti coinvolti nel progetto sono stati così rappresentati, rispettivamente, attraverso istanze delle classi `Person` e `Organization` introdotte da FOAF. Inoltre, la gerarchia a quattro livelli del progetto, la quale lo categorizza progressivamente in sottoprogetti, work package, azioni e unità operative, è stata rappresentata con istanze della classe `Group`, anch'essa definita da FOAF. I Ricercatori sono quindi stati messi in relazione sia con gli Istituti di afferenza sia con le branche del progetto di loro competenza.
A questo punto, l'intervento umano si è reso necessario per normalizzare i dati da entità duplicate e per inserire quelle che non fossero presenti nel documento originale (o che non fossero state "catturate" dalla logica di programmazione espressa dal foglio di stile). In aggiunta a questa fase di validazione del dato grezzo, un'altra importante attività è consistita nell'associare alcuni utenti ai rispettivi domini di interesse, una informazione che non era presente nel foglio di calcolo originale. Questa fase del flusso di lavoro prevede due possibili metodologie:

1. allo scopo di testare le funzionalità orientate all'utente che si vogliono realizzare, una apposita applicazione di *backoffice* è stata creata per consentire l'associazione tra Ricercatori e uno dei 12 domini di ricerca del progetto;
2. dal momento che questa informazione appartiene al profilo associato ad ognuno degli utenti, questi saranno in grado di modificarli a piacimento attraverso il portale centralizzato in via di sviluppo.

8 eXtensible Stylesheet Language (XSL): http://www.w3.org/Style/XSL/

Il risultato finale è una struttura dati che raccoglie le informazioni relative a un insieme di 1197 Ricercatori afferenti a 31 Istituti e suddivisi in 563 gruppi di lavoro organizzati in una gerarchia a quattro livelli.
La Sezione che segue mostrerà come questa struttura dati, insieme alle altre realizzate o raccolte nel sottoprogetto 7, consenta la realizzazione di funzionalità innovative nell'implementazione dell'Infrastruttura di Dati Spaziali (SDI) di RITMARE.

4 – Esempi di utilizzo dell'informazione

Il primo e più immediato vantaggio della struttura dati per l'anagrafica di progetto descritta precedentemente consiste nella semplificazione della procedura di metadatazione delle risorse, ovvero della specifica delle informazioni che consente la loro ricerca all'interno della SDI. Ad esempio, il formato di metadatazione per i dati geografici ambientali previsto dalla normativa italiana richiede la specifica di ben quattro punti di contatto, per ognuno dei quali si devono specificare altrettanti frammenti di informazione per indicare l'Istituto di appartenenza, il sito web di quest'ultimo, l'e-mail della persona e il ruolo ricoperto rispetto alla risorsa (es., fornitore, distributore, ecc.). Un'attività sicuramente tediosa e scoraggiante per l'utente. Questo diviene ancor più evidente nel caso della metadazione di sensori utilizzati nelle osservazioni di RITMARE, per i quali lo standard utilizzato, il Sensor Observation Service[9], richiede di fornire ulteriori specifiche rispetto ai punti di contatto, quali ad esempio l'indirizzo completo del fornitore.
La struttura dati per l'anagrafica di progetto è utile in questi casi perché consente di suggerire all'utente l'indirizzo e-mail della persona che si intende specificare come punto di contatto (si veda l'informazione alla linea 10 del Listato 1), mentre l'Istituto di appartenenza (specificato alla riga 11 del Listato 1) consente di recuperare i dati relativi a quest'ultimo, parte dei quali sono riportati nel Listato 2. In questo modo è possibile sviluppare applicazioni per l'autocompletamento della maggior parte dei campi richiesti.

```
01 @prefix rdf:         <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
02 @prefix xsd:         <http://www.w3.org/2001/XMLSchema#> .
03 @prefix foaf:                <http://xmlns.com/foaf/0.1/> .
04 @prefix vcard2006:   <http://www.w3.org/2006/vcard/ns#> .
05 @prefix ns1:         <http://sp7.irea.cnr.it/rdfdata/project/> .
06 @prefix ns3:         <http://wymiwyg.org/ontologies/foaf/postaddress#> .
07 ns1:IREA     rdf:type        foaf:Organization ;
08      rdf:label                       "IREA" ^^xsd:string ;
09      foaf:homepage                   <http://irea.cnr.it/> ;
10      ns3:thoroughfareName    "Bassini"^^xsd:string ;
11      ns3:thoroughfareType    "Via"^^xsd:string ;
12      ns3:streetNr                    "15"^^xsd:string ;
13      ns3:town                        "Milano"^^xsd:string ;
14      ns3:postcode                    "20133"^^xsd:string ;
15      ns3:country                     "IT"^^xsd:string .
16 ...
```

Listato 2 □Descrittore FOAF associato all'istituto IREA.

Un vantaggio ancora più fine dell'anagrafica consiste nell'utilizzare le informazioni riguardanti l'afferenza o la rete sociale associate ad un Ricercatore per stabilire le politiche di accesso alle risorse che questi ha fornito. Infatti, una volta che un utente si sia autenticato presso il portale RITMARE, è possibile associarlo ad un proprio descrittore FOAF e, in base alle informazioni in esso contenute, ricavare le relazioni intercorrenti tra questi e le altre entità conosciute al sistema, attribuendogli permessi di accesso relativi alla sua posizione riguardo a RITMARE. E' interessante notare che è possibile integrare i descrittori FOAF con meccanismi di autenticazione forniti da

9 Sensor Observation Service (SOS): http://www.opengeospatial.org/standards/sos

terze parti, quali ad esempio il paradigma OpenID[10] supportato da un crescente numero di fornitori di servizi.
Ad esempio, la politica di accesso più banale, secondo la quale un Ricercatore è in grado di accedere alle risorse che egli stesso ha fornito, è immediatamente realizzabile facendo riferimento all'identificatore di uno dei punti di contatto che sono stati definiti per la risorsa stessa all'interno dei metadati associati a quest'ultima. Allo stesso modo, l'accesso può essere esteso ai Ricercatori che siano a lui correlati attraverso la proprietà `knows` (come alla riga 12 del Listato 1), a Ricercatori afferenti allo stesso Istituto, a quelli coinvolti nello stesso sottoprogetto, work package, azione o unità operativa all'interno di RITMARE, ecc. Una flessibilità ancora maggiore può essere ottenuta facendo riferimento a strutture dati definite esternamente al progetto.
Infine, l'utilizzo più generalizzato di questa categoria di strutture dati lo si ottiene nella *discovery* di risorse, ovvero nella loro ricerca attraverso la specifica di un numero di parametri che trovino riscontro nei metadati ad esse associati. Supponiamo ad esempio che si voglia presentare all'utente, attraverso il portale RITMARE, una lista di risorse di recente inserimento all'interno dell'infrastruttura (informazione ricavabile dalla data di creazione dei metadati) o che hanno riscosso un notevole interesse da parte degli utenti (informazione deducibile dal numero di accessi ai metadati o alla risorsa stessa). Nel fare ciò, si vorrebbe tuttavia dare maggiore risalto alle risorse che si suppone possano essere d'interesse per l'utente, che rispecchino cioè il dominio di interesse di quest'ultimo.
Nel Listato 1 abbiamo visto come un Ricercatore possa essere associato ad uno o più argomenti di interesse attraverso la proprietà `topic_interest` di FOAF. Non è però detto che questo stesso termine sia stato utilizzato nella metadatazione delle risorse, tanto più che il termine riferito alla linea 12 del listato appartiene alla lista delle aree di ricerche RITMARE che, data la granularità grossolana, consta di pochi elementi. Ciononostante, il formato utilizzato per la specifica dei vocabolari controllati utilizzati in RITMARE, e cioè Simple Knowledge Organization System[11] (SKOS), dispone delle primitive necessarie ad includere questo termine nel più vasto panorama di vocabolari controllati e thesauri disponibili in rete. Ad esempio, il termine indicato nel Listato 1 come area di ricerca del Ricercatore Alessandro Oggioni (`Ecosistemi`) è attinente ai quattro *temi* di INSPIRE "Zone a rischio naturale", "Regioni biogeografiche", "Habitat e biotopi" e "Distribuzione delle specie".
Contrariamente al termine che esprime l'area di ricerca RITMARE di interesse per Alessandro Oggioni, i temi di INSPIRE hanno un'alta probabilità di essere presenti nei record di metadati che siano confacenti la Direttiva, in quanto la specifica di almeno uno di essi è obbligatoria. In questo modo è possibile selezionare il sottoinsieme di risorse che con maggiore probabilità saranno di interesse per lo specifico utente che usufruisce del portale RITMARE. Inoltre vale la pena sottolineare come ognuno dei termini INSPIRE disponga di una traduzione in tutte le lingue ufficiali della Unione Europea: in questo modo, è possibile abilitare anche la ricerca multilingue delle risorse.

5 – Conclusioni

In questo lavoro abbiamo descritto la strategia seguita, nel contesto del Progetto Bandiera RITMARE, per la rappresentazione orientata alla semantica dell'insieme di Ricercatori e Istituti coinvolti; tale rappresentazione è in grado di dettagliare finemente queste entità e la loro collocazione all'interno della suddivisione multilivello del progetto.
Abbiamo fornito esempi pratici di come le strutture dati create possano essere utilizzate in tre contesti differenti e complementari nell'organizzazione, fruizione e reperimento dell'informazione geografica: precisamente, a supporto della metadatazione, del controllo dell'accesso alla risorsa e della ricerca.

10 OpenID: http://openid.net/
11 Simple Knowledge Organization System (SKOS): http://www.w3.org/2004/02/skos/

Per una descrizione esaustiva delle variegate strutture dati che sono state integrate a supporto dell'implementazione dell'Infrastruttura di Dati Spaziali di RITMARE, si rimanda agli articoli (Fugazza et al., 2014a) e (Fugazza et al., 2014b).

Ringraziamenti: Questo lavoro è stato finanziato nell'ambito del Progetto bandiera del MIUR RITMARE.

Riferimenti bibliografici

Commissione Europea (2007), "Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)" Directive 2007/2/EC.

Agenzia per l'Italia Digitale (2014), "Guida operativa metadati sui DATI - v. 2.0", http://www.rndt.gov.it/RNDT/home/images/RNDT_guida_operativa_dati_v2.0_20140725.pdf

Commissione di Coordinamento SPC (2012), "Linee guida per l'interoperabilità semantica attraverso i Linked Open Data", http://www.digitpa.gov.it/notizie/pubblicate-le-linee-guida-l-interoperabilita-semantica-attraverso-i-linked-open-data

Fugazza C., Pepe M., Oggioni A., Pavesi F., Carrara P. (2014a), "A holistic, semantics-aware approach to Spatial Data Infrastructures", 3rd International Conference on Data Management Technologies and Applications (DATA)

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", Current Research Information Systems (CRIS)