MUSEOLOGIA SCIENTIFICA nuova serie • 11: 114-118 • 2017 ISSN 1123-265X

Georiferimento di campioni museali nell'infrastruttura LifeWatch Italia: le nuove prospettive dal web semantico

Paolo Tagliolato
Istituto per le Scienze Marine (CNR-ISMAR). Arsenale, Castello 2737/F. I-30122 Venezia. E-mail: tagliolato.p@irea.cnr.it

Alessandro Oggioni
Cristiano Fugazza
Istituto per il Rilevamento Elettromagnetico dell'Ambiente (CNR-IREA). Via Bassini, 15. I-20133 Milano.
E-mail: oggioni.a@irea.cnr.it, fugazza.c@irea.cnr.it

Fabio Cianferoni
Stefano De Felici
Istituto di Biologia Agroambientale e Forestale (CNR-IBAF). Via Salaria km 29,300. I-00015 Monterotondo (Roma).
E-mail: fabio.cianferoni@unifi.it, stefano.de.felici@uniroma2.it

RIASSUNTO

Il georiferimento è una delle fasi più problematiche e delicate della digitalizzazione dei dati di occorrenza. Nei campioni museali i riferimenti spaziali sono di norma forniti attraverso una descrizione testuale di località, una tecnica "informale" da trattare opportunamente per la riconduzione a una rappresentazione "formale". Nell'ambito di LifeWatch-Italia è in corso la sperimentazione di un approccio innovativo al trattamento degli oggetti geografici che, seguendo l'ontologia GeoNames, considera la località come "concetto", svincolandola da legami con i nomi o con rappresentazioni geografiche statiche e inserendola esplicitamente nella rete di relazioni spaziali e semantiche con altre entità. Si è scelto di formalizzare così i toponimi italiani tratti dalle tavolette IGM che rappresentano una fonte autorevole di nomi di località, così che essi possano essere utilizzabili come fonte semantica nell'ambito museale e non solo.

Parole chiave:
occorrenza, toponimi, geonames, musei, biodiversità.

ABSTRACT

Georeferencing museological specimens in the LifeWatch Italy infrastructure: The new perspectives enabled by the Semantic Web

Geo-referencing is one of the most problematic and troublesome steps in digitization of occurrence data. In museum samples, spatial references are typically provided as textual description of localities, an "informal" technique that requires appropriate processing in order to become a "formal" representation. The e-Biodiversity Research Institute LifeWatch-Italy is proposing an innovative approach for processing of geographic objects. This approach, which follows the GeoNames ontology, considers localities as "concepts", thus enabling: a) to loosen their connections with names and with static geographic representations and b) to explicitly insert them in the network of spatial and semantic relationships with other entities.
It was decided to formalize the Italian place names taken from the IGM plates ("tavolette"), an authoritative source of place names, according to these principles and in order to use these in the museum or biodiversity domain.

Key words:
occurrence, locality names, geonames, museum, biodiversity.

INTRODUZIONE

I dati delle collezioni museali rappresentano un patrimonio insostituibile per lo studio della diversità biologica, dei processi evolutivi, della distribuzione degli organismi e delle sue variazioni spaziali e temporali (Holmes et al., 2016; Page et al., 2015). Con lo sviluppo di grandi infrastrutture digitali di ricerca, la disponibilità di tali informazioni in formato elettronico ha aperto la strada a scenari di e-science inediti, consentendo analisi a scale molto ampie precedentemente assai difficili da effettuare o addirittura impossibili (Brooks et al., 2011).

La digitalizzazione dei dati primari di biodiversità (*occurrence*) ha visto nel tempo la proposta di numerosi strumenti (es. Geolocate o Georeferencing Calculator) e procedure per (semi-) automatizzarne il processo e ottenere risultati di elevato livello qualitativo (Wieczorek et al., 2004; Guralnick et al., 2006; Moncla et al., 2014; Blom, 2016). Benché neppure i dati recenti siano immuni da problemi, le maggiori difficoltà riguardano, nei dati museali, il "georiferimento", l'individuazione di oggetti geografici (Hackeloeer et al., 2014) e la valutazione di una stima di incertezza (Guo et al., 2008; Liu et al., 2009), che classicamente avvengono a partire dalle descrizioni testuali di località presenti sui supporti originali (etichette, cartellini, diari di raccolta etc.). Tali modalità di georiferimento di tipo "informale" (Chapman et al., 2006) sono soggette a problematiche quali: presenza di nomi "storici" obsoleti, vaghezza dei riferimenti spaziali ("nei dintorni di…", "vicino a…" etc.), uso di nomi strettamente locali, errori di compitazione ed altro, che richiedono un opportuno trattamento per la riconduzione a una rappresentazione "formale" (Hill, 2006).

Il gruppo di lavoro della Commissione Europea sui dati scientifici (European Commission High Level Expert Group on Scientific Data, 2010) e una recente revisione riguardante le necessità di unificare le pratiche dell'informatica per la biodiversità in Europa (Koureas et al., 2016) hanno evidenziato, quale punto nodale per lo sviluppo di infrastrutture scientifiche, la capacità di rendere comprensibili a livello globale dati di natura estremamente varia per forma, contenuto e pratiche di comunità un tempo separate. Componenti chiave in questo senso, utili anche allo scopo di favorire l'interoperabilità (Haslhofer & Klas, 2010), sono l'impiego di sintassi condivise e di autorevoli fonti di riferimento per le terminologie utilizzate nei dati e nelle loro descrizioni (metadati).

LifeWatch, la nascente infrastruttura virtuale di ricerca europea di promozione e supporto tecnologico agli studi sulla biodiversità e gli ecosistemi, ha dato il via alla costituzione di gruppi di lavoro sullo sviluppo di "vocabolari controllati" nell'ambito della biodiversità (Rosati et al., 2015; Tagliolato et al., 2015; Bergami et al., 2016, Rosati et al., 2017). In questa sede viene presentata una linea di sviluppo che si colloca nell'alveo della strategia atta a favorire l'interoperabilità semantica dei dati ecologici, che si concentra sui nomi e sulla rappresentazione delle entità geografiche. I toponimi vengono utilizzati in modo trasversale in svariate discipline e il loro supporto tecnologico appare ancora per certi versi inadeguato a uno scenario di infrastruttura a livello internazionale.

VERSO LA COSTRUZIONE DI GAZETTEER AUTOREVOLI NELLA PROSPETTIVA DEL WEB SEMANTICO

Nell'ambito dei sistemi informativi la conversione da un georiferimento "informale", basato cioè su riferimento a nomi, ad un sistema "formale" avviene attraverso l'uso di "gazetteer", i.e. indici geografici, che permettono di ottenere, dato un nome geografico la corrispondente entità spaziale (puntuale, lineare o superficiale).

Se la rappresentazione formale tradizionalmente codificata e gestita in sistemi GIS e in Spatial Data Infrastructures (SDI) ha un'assodata maturità per le caratteristiche geometriche e topologiche, risulta invece meno sviluppata la possibilità di trattamento di aspetti *semantici*, in genere modellati come attributi aggiuntivi degli oggetti spaziali rappresentati.

I nomi geografici rappresentano, è vero, luoghi, ma nel concetto di "luogo" la delimitazione spaziale può non essere l'unico aspetto di interesse. Il modello GIS tradizionale può non catturare parti rilevanti di informazione: si pensi alle reti di relazione, non solo relazioni topologiche, ma anche amministrative, politiche, economiche o ecologiche che un "luogo" può avere con altri "luoghi".

Tali relazioni, come i nomi stessi e i concetti di luogo, possono permanere anche al mutare della loro rispondenza geografica, come può avvenire nella confinistica quando le linee di confine tra due stati siano assegnate a elementi naturali caratterizzati da incertezza, variabili nel tempo, come ad esempio fiumi o linee spartiacque. In questo senso poter introdurre un modello formale "semantico" a supporto di quello esistente appare rilevante.

Una seconda considerazione di particolare interesse riguarda la possibilità di avere definizioni globali ed autorevoli dei nomi geografici. Da un punto di vista informatico nell'ambito dello sviluppo di una infrastruttura distribuita, risulta infatti essenziale che ad oggetti geografici distinti corrispondano identificativi digitali univoci (URI). In questo scenario ad un identificativo univoco corrisponde il concetto di un particolare luogo, modellato semanticamente nella sua rete di relazioni.

L'impiego di URI per l'individuazione virtuale di risorse a livello globale, congiunta a tecnologie quali le ontologie informatiche è la strategia di integrazione di sistemi distribuiti legata alla visione del web semantico (Allemang & Hendler, 2011).

I TOPONIMI ITALIANI E LA LORO MAPPATURA IN STRUTTURE RDF

Con l'obiettivo della costruzione di un "gazetteer semantico" per i nomi geografici italiani è stata effettuata la mappatura delle 114 categorie dei toponimi IGM sulle analoghe 677 categorie ("Feature codes" e "Feature classes") definite nell'ontologia GeoNames (Vedi Siti web n. 1) e la trasformazione di tutti i toponimi (in totale 716707 liberamente accessibili come servizio WFS dal Geoportale Nazionale) (Vedi Siti web n. 2) (fig. 1) in istanze di oggetti geografici dell'ontologia ("GeographicFeatures"). Queste nuove risorse sono state collegate con altre già esistenti a livello istituzionale nell'ambito del Linked Open Data Cloud (LOD) (Berners-Lee, 2006) (Vedi Siti web n. 3). In particolare si sono collegati i toponimi alle risorse semantiche rappresentative delle entità amministrative presenti nelle basi di dati di ISPRA e di ISTAT. Nella modellazione semantica i 6 attributi (tab. 1) che nella fonte del geoportale definiscono l'appartenenza di un toponimo a comune, provincia e regione, sono stati sostituiti dal predicato geonames. In questo modo le risorse RDF disponibili nel sistema globale sono state arricchite, mantenendo per ognuna di esse il fornitore di servizio più autorevole per la sua manutenzione. Le trasformazioni sono state codificate in linguaggio XSLT (eXtensible Stylesheet Language Transformation) (W3C XSL Working Group, 2007). Un esempio di un'istanza della modellazione dell'ontologia GeoNames ottenuta, in rappresentazione RDF Turtle viene mostrato di seguito:

```
<http://rdfdata.get-it.it/.../1418> a gn:Feature;
        skos:inScheme <http:// rdfdata.get-it.it/
        /IGM_ Toponymes>;
        gn:name "PUNTA SOTTILE";
        rdfs:label "PUNTA SOTTILE";
        gn:countryCode "IT";
        gn:featureCode <http:// rdfdata.get-it.it/...
        /FC/ area_geografica>;
        dc:source "Tavoletta: 53A14XE";
        dc:create "19 0";
        gn:parentADM3 <http://dati.isprambiente
        .it/id/place/32003>;
        wgs84_pos:long "13.717225";
        wgs84_pos:lat "45.60645".
```

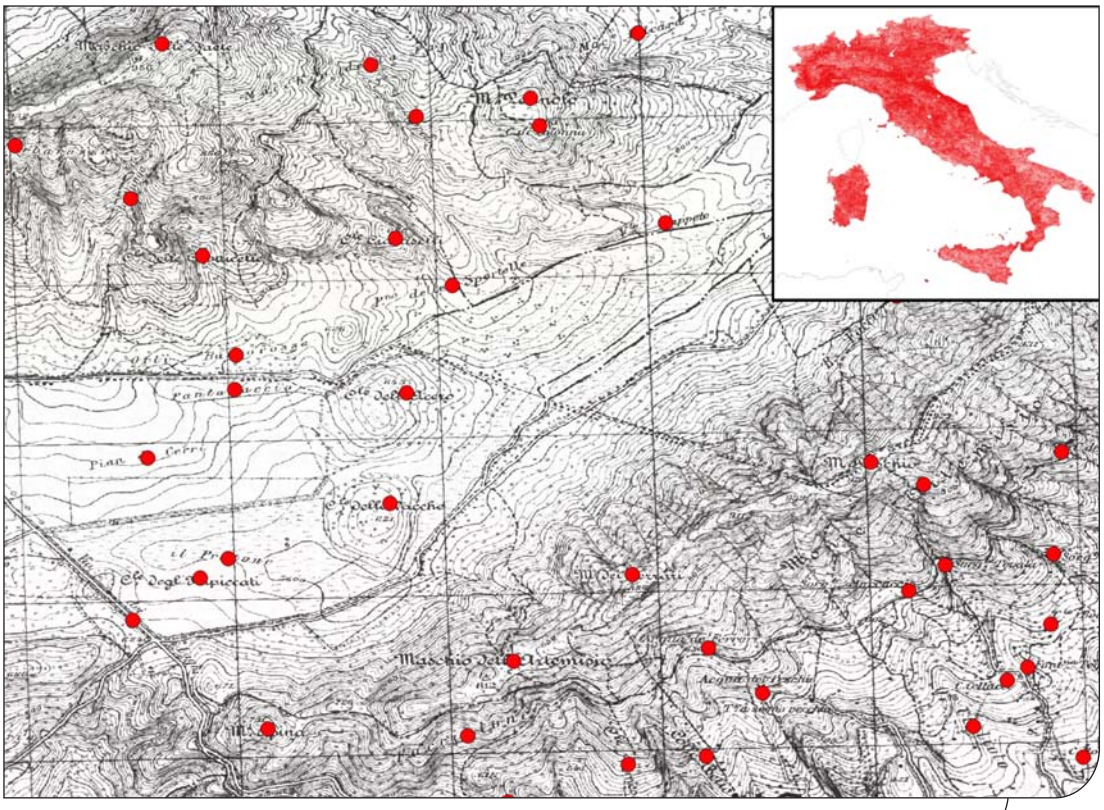

Fig. 1. La densità media dei toponimi, calcolata sull'intera superficie nazionale, è pari a 2,38 nomi/Km^2, valore che assicura un'eccellente copertura spaziale dei nomi di località. Vertici dell'immagine NO: 12,72283; 41,75083; SE 12,78890; 41,70478; SRS WGS84

Attributo	Descrizione
toponimo	nome del toponimo;
secondo_nome	eventuale secondo nome;
tipo	tipo di toponimo;
oggetto_toponimo	oggetto del toponimo; (categoria)
tavoletta	tavoletta da cui è stato digitalizzato il toponimo;
edizione	edizione della tavoletta;
data	data di ricognizione della tavoletta;
codice_istat	codice ISTAT aggiornato al 2011;
comune	nome del comune aggiornato al 2011;
provincia	nome della provincia aggiornato al 2011;
regione	nome della regione aggiornato al 2011;
cod_comune	codice del comune aggiornato al 2011;
cod_provincia	codice della provincia aggiornato al 2011;
cod_regione	codice della regione aggiornato al 2011.

Tab. 1. Attributi e relativa descrizione dei toponimi italiani così come rilasciati dal Geoportale Nazionale.

CONCLUSIONE E SVILUPPI FUTURI

In questo lavoro si è proposta una modellazione di risorse geografiche da fonti autorevoli sotto forma di risorse per il web semantico. Il vantaggio, già anticipato nell'introduzione, è quello di fornire uno strumento che faciliti il passaggio da rappresentazioni informali delle entità geografiche a rappresentazioni formali, gestite mediante strumenti informatici e collocate in reti di relazioni semantiche con altre risorse. Le risorse, così come descritte nel testo, sono disponibili allo sparql endpoint (Vedi Siti web n. 4) e navigabili come Linked Data a partire per esempio dall'indirizzo (Vedi Siti web n. 5). Mentre una prima sperimentazione in corso riguarda l'abilitazione del software di metadatazione EDI (Pavesi et al., 2016) (Vedi Siti web n. 6) al formato Darwin Core. La peculiarità di EDI è facilitare la produzione di metadati sfruttando le risorse disponibili nel web semantico, fornendo agli utenti suggerimenti circa i valori da inserire nel metadato e producendo un'informazione arricchita da questo tipo di fonti.
In ambito museale soluzioni di questo genere facilitano le modalità di digitalizzazione delle collezioni e le perfezionano, abilitando l'inclusione in grandi infrastrutture di ricerca.

RINGRAZIAMENTI

L'attività descritta in questo articolo è stata supportata da LifeWatch-Italia.

BIBLIOGRAFIA

ALLEMANG D., HENDLER J.A., 2011. *Semantic Web for the working ontologist: effective modeling in RDFS and OWL*, 2nd ed. Morgan Kaufmann/Elsevier, Waltham, Massachusetts, 53 pp.

BERGAMI C., FIORE N., OGGIONI A., ROSATI I., TAGLIOLATO P., 2016. *Thesauri and Semantics in the Ecological Domain Report.* Università del Salento, 9-10 Giugno 2016, Lecce.

BERNERS-LEE T., 2006. Linked Data. *International Journal on Semantic Web and Information Systems, 4: 1.* doi:10.4018/jswis.2009081901

BLOM J., 2016. *Onderzoeksverslag. How to georeference primary specimendata in a (semi-) automated process and which tools are available and most useful? Haagse Hogeschool Informatie dienstverlening en informatie management.* Naturalis Biodiversity Center, Leiden, 61 pp.

BROOKS S. J. AND CLIMATE CHANGE RESOURCES GROUP, 2011. Natural history collections as sources of long-term datasets. *Trends in Ecology and Evolution, 26(4):* 153-154.
https://doi.org/10.1016/j.tree.2010.12.009

CHAPMAN A.D., WIECZOREK J., BIOGEOMANCER CONSORTIUM, 2006. *Guide to best practices for georeferencing*. Global Biodiversity Information Facility, Copenhagen, Denmark, 30 pp.

EUROPEAN COMMISSION HIGH LEVEL EXPERT GROUP ON SCIENTIFIC DATA, 2010. *Riding the wave How Europe can gain from the rising tide of scientific data*. Final report of the High Level Expert Group on Scientific Data. *Communication, 21:* 40. doi:10.1016/j.cub.2011.03.009

Guo Q., Liu Y., Wieczorek J., 2008. Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. *International Journal of Geographical Information Science*, 22(10): 1067-1090. https://doi.org/10.1080/13658810701851420

Guralnick R., Wieczorek J., Beaman R.S., Hijmans R., 2006. BioGeomancer: automated georeferencing to map the world's biodiversity data. *PLoS Biology*, *4(11)*: e381. https://doi.org/10.1371/journal.pbio.0040381

Hackeloeer A., Klasing K., Krisp J.M., Meng L., 2014. "Georeferencing: a review of methods and applications". *Annals of GIS*, *20(1)*: 61-69.

Haslhofer B. & Klas W., 2010. A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys*, *42(2)*: 1-37. http://dl.acm.org/citation.cfm?id=1667062.1667064http://portal.acm.org/citation.cfm?doid=1667062.1667064

Hill L.L., 2006. *Georeferencing: the geographic associations of information*. MIT Press, Cambridge, Massachusetts, 70 pp.

Holmes M.W., Hammond T.T., Wogan G.O.U., Walsh R.E., Labarbera K., Wommack E.A., Martins F.M., Crawford J.C., Mack K.L., Bloch L.M. Nachman, M.W., 2016. Natural history collections as windows on evolutionary processes. *Molecular Ecology*, *25(4)*: 864-881. https://doi.org/10.1111/mec.13529

Koureas D., Hardisty A., Vos R., Agosti D., Arvanitidis C., Bogatencov P., Buttigieg P., de Jong Y., Horvath F., Gkoutos G., Groom Q., Kliment T., Kõljalg U., Manakos I., Marcer A., Marhold K., Morse D., Mergen P., Penev L., Pettersson L., Svenning J., van de Putte A., Smith V., 2016. Unifying European Biodiversity Informatics (BioUnify). *Research Ideas and Outcomes*, *2*: e7787. https://doi.org/10.3897/rio.2.e7787

Liu Y., Guo Q., Wieczorek J., Goodchild M., 2009. Positioning localities based on spatial assertions. *International Journal of Geographical Information Science*, *23(11)*: 1471-1501. https://doi.org/10.1080/13658810802247114

Moncla L., Renteria-agualimpia W., Nogueras-iso J., Moncla L., Renteria-agualimpia W., Nogueras-iso J., Geocoding M. G., 2014. Geocoding for texts with fine-grain toponyms : an experiment on a geoparsed hiking descriptions corpus. *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Dallas, Texas.

Page, L.M., Macfadden B.J., Fortes J.A., Soltis P.S., Riccardi G., 2015. Digitization of Biodiversity Collections Reveals Biggest Data on Biodiversity. *BioScience*, *65(9)*: 841-842. https://doi.org/10.1093/biosci/biv104

Pavesi F., Basoni A., Fugazza C., Menegon S., Oggioni A., Pepe M., Tagliolato P., Carrara P., 2016. EDI – A Template-Driven Metadata Editor for Research Data. *Journal of Open Research Software - JORS 4*. doi: 10.5334/jors.106.

Rosati I., Boggero A., Fiore N., Franzoi P., Fugazza C., Oggioni A., Pugnetti A., Stanca R.L., Tagliolato P., Vizzini S., Zingone A., Basset A., 2015. Semantic tools for functional trait-based approaches: development and applicability. *Ecology at the Interface - 13th European Ecological Federation Congress (EEF 2015)*. European Ecological Federation, September 21-25, 2015.

Rosati I., Bergami C., Stanca E., Roselli E., Tagliolato P., Oggioni A., Fiore N., Pugnetti A., Zingone A., Boggero A., Basset A., 2017. A thesaurus for phytoplankton trait-based approaches: Development and applicability. Ecological Informatics, *42*: 129–138. https://doi.org/10.1016/j.ecoinf.2017.10.014

Tagliolato P., Oggioni A., Fugazza C., De Felici S., Cianferoni F., Rosati I., Fiore N., Carrara P., Basset A., 2015. Leveraging biodiversity interoperability through Lifewatch semantic resources. *Ecology at the Interface - 13th European Ecological Federation Congress (EEF 2015)*. European Ecological Federation, September 21-25, 2015.

W3C XSL Working Group, 2007. *Xsl transformations (xslt) version 2. W3C Recommendation 23 January 2007*, World Wide Web Consortium.

Wieczorek J., Guo Q., Hijmans R., 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, *18(8)*: 745-767. https://doi.org/10.1080/13658810412331280211

Siti Web (accessed 27.02.17)

1) Ontologia GeoNames
http://www.geonames.org

2) Geoportale Nazionale Ministero Ambiente
http://www.pcn.minambiente.it/GN/

3) Risorse RDF relative a comuni, province e regioni
http://dati.isprambiente.it/id/place

4) Distribuzione toponimi RDF (SPARQL endpoint)
http://fuseki1.get-it.it/LWItaToponyms/

5) Risorse RDF relative alle categorizzazioni e ai toponimi con accesso Linked Data
http://rdfdata.get-it.it/LWItaToponyms/FC

6) Ed 0itor metadati EDI
http://edidemo.get-it.it/dist/DwC.html

Submitted: February 28th, 2017 - Accepted: November 14th, 2017
Published: December 18th, 2017