Tecnologie Semantiche per la Produzione e Pubblicazione di Open Data presso l'ACI - Automobile Club d'Italia

Daniela Caltabiano[1], **Emanuela Catoni**[1], **Antonio Fabrizi**[1], **Domenico Lembo**[2], **Mauro Minenna**[1], **Mario Punchina**[1], **Valerio Santarelli**[2,3]

[1]ACI Informatica, [2]Sapienza Università di Roma, [3]OBDA Systems

[1]*{d.caltabiano;e.catoni;a.fabrizi;m.minenna;m.punchina}*@informatica.aci.it,
[2]*lastname*@diag.uniroma1.it, [3]santarelli@obdasystems.com

1 Introduzione

Le tecnologie semantiche combinano tecniche di rappresentazione della conoscenza ed intelligenza artificiale al fine di ottenere una più efficace gestione del patrimonio informativo di una organizzazione, grazie alla loro abilità di separare il livello concettuale delle applicazioni da quelli logico e fisico, e per la possibilità che offrono di sfruttare servizi di ragionamento automatico per l'accesso ed il controllo dei dati. In questo contesto, l'*Ontology-based Data Management* (OBDM) [Lenzerini, 2018] si è affermato come paradigma per la gestione dei dati, basato su una architettura a tre livelli, composta da un'ontologia, dalle sorgenti di dati, e dai collegamenti, o *mapping*, tra i due. L'ontologia costituisce una concettualizzazione formale del dominio di interesse, i cui elementi principali (o predicati) sono *concetti*, che denotano insiemi di oggetti dalle caratteristiche comuni, *relazioni* (binarie) tra concetti, che denotano legami tra oggetti, ed *attributi*, che denotano associazioni fra oggetti e valori. Il livello delle sorgenti è costituito dalle banche dati utilizzate dai sistemi operazionali dell'organizzazione. Il mapping invece definisce in modo dichiarativo i collegamenti tra gli elementi dell'ontologia e le sorgenti di dati, attraverso la definizione di opportune interrogazioni sui database sottostanti, senza che sia necessario modificarli o materializzare i dati in un nuovo repository. In un sistema basato su questa architettura, la possibilità di sfruttare servizi di ragionamento sull'ontologia semplifica enormemente l'accesso ai dati, che avviene prevalentemente attraverso interrogazioni sull'ontologia stessa, che vengono processate automaticamente sulla base dell'ontologia, del mapping e dei dati alle sorgenti [Poggi *et al.*, 2008; Kontchakov *et al.*, 2014]. Inoltre, diversi servizi di ragionamento intensionale supportano il progettista nella definizione e gestione dell'ontologia e del mapping.

In questo lavoro presentiamo un progetto svolto in collaborazione tra la Sapienza Università di Roma, l'Automobile Club d'Italia (ACI) ed OKKAM S.r.l.[1], spin off dell'Università degli Studi di Trento. Nel progetto sono state utilizzate tecnologie semantiche per l'accesso a dati relativi al Pubblico Registro Automobilistico (PRA) ed alle tasse automobilistiche gestite da ACI, ed è stato realizzato un portale per la pubblicazione in formato *Linked Open* di dati riguardanti il parco veicolare italiano.

[1]http://www.okkam.it/

2 Ontologia e Mapping

Nell'ambito del progetto è stata realizzata una ontologia che descrive gli aspetti rilevanti del contesto operativo di ACI, ed in particolare si concentra sulle caratteristiche tecniche dei veicoli, sui tributi a questi relativi, e sulle formalità presentate al PRA. L'ontologia fa uso di alcuni avanzati pattern di modellazione, usati ad esempio per specificare proprietà di concetti che evolvono nel tempo e per gestire diverse rappresentazioni dello stesso concetto da parte di più agenti[2]. L'ontologia è espressa in OWL 2, ed è stata realizzata con l'ambiente di sviluppo per ontologie Eddy[3]. Eddy consente di definire l'ontologia usando Graphol [Lembo *et al.*, 2016], un linguaggio grafico per la creazione di ontologie OWL 2 tramite diagrammi che hanno una struttura simile a quella adottata nei modelli ER o in UML. Eddy offre tra l'altro funzionalità per garantire la correttezza sintattica e semantica dell'ontologia, ed in particolare consente di verificare la consistenza dell'ontologia e dei suoi predicati, ed ottenere spiegazioni (explanation) su eventuali inconsistenze, sfruttando i servizi del ragionatore per Logiche Descrittive HermiT [Glimm *et al.*, 2014].

L'ontologia realizzata è composta da 8 moduli, che riflettono la partizione del dominio di interesse in altrettante aree logicamente connesse. Fra i moduli realizzati, quello relativo ai *veicoli* comprende la caratterizzazione e classificazione dei veicoli e la gestione dei loro stati rilevanti, per catturare le variazione delle loro proprietà nel tempo. Il modulo relativo alle *formalità* e quello riguardante il *possesso* descrivono rispettivamente le registrazioni al PRA di modifiche delle caratteristiche di un veicolo durante il suo ciclo di vita ed il suo contesto di uso. Inoltre, il modulo *statistico* si focalizza su aspetti riguardanti i dati statistici pubblicati da ACI sotto forma di Linked Open Data. In quest'ultimo, sono stati importati elementi della *Data Cube ontology*[4], un vocabolario standard del W3C per la rappresentazione di dati statistici multidimensionali, mentre per modellare aspetti del territorio sono stati utilizzati alcuni termini dell'ontologia *GeoNames*[5].

Il mapping definisce il legame semantico fra la porzione di interesse delle banche dati dei sistemi informativi di ACI e

[2]Per un approfondimento si rimanda al tutorial su "Methods and Tools for Developing OBDM Solutions" (https://goo.gl/9UtkLn)

[3]http://www.obdasystems.com/eddy

[4]https://www.w3.org/TR/vocab-data-cube/

[5]http://www.geonames.org/ontology/documentation.html

l'ontologia. In generale, un mapping è composto da asserzioni che associano gli elementi dell'ontologia ad interrogazioni sugli schemi delle sorgenti. Intuitivamente, tali asserzioni definiscono come i predicati dell'ontologia possono essere popolati a partire dai dati restituiti dalle interrogazioni. Nel progetto, il mapping è stato realizzato utilizzando un plug-in per Protégé prodotto da Sapienza e O.B.D.A. Systems[6]. Le asserzioni generate sono compatibili con lo standard R2RML[7] e sono esportabili in questo formato tramite le funzionalità del plug-in. Durante la definizione del mapping sono state verificate alcune sue proprietà semantiche, per prevenire tipiche anomalie di progettazione, quali l'inconsistenza (dovuta ad asserzioni che popolano l'ontologia in modo sempre contraddittorio, indipendentemente dai dati alle sorgenti) e la ridondanza (dovuta alla presenza di asserzioni logicamente implicate da altre asserzioni e dall'ontologia), che può risultare problematica in fase di manutenzione [Lembo *et al.*, 2015].

3 L'accesso ai dati ed i Linked Open Data

Il sistema realizzato è in grado di processare automaticamente interrogazioni espresse in SPARQL sull'ontologia, sfruttando i servizi del ragionatore per Mastro [De Giacomo *et al.*, 2012]. L'approccio di Mastro per tale scopo è puramente intensionale: l'interrogazione utente viene riformulata rispetto all'ontologia ed al mapping, in modo da produrre una nuova interrogazione che codifica il ragionamento ed è direttamente eseguibile sui sistemi relazionali alle sorgenti. Si noti che gli algoritmi di riscrittura usati da Mastro richiedono che l'ontologia sia specificata in una logica *DL-Lite* [Poggi *et al.*, 2008]. Per questo, ai fini del calcolo della risposta alle interrogazioni, Mastro preliminarmente approssima l'ontologia ACI da OWL 2 a *DL-Lite* [Console *et al.*, 2014].

Nella soluzione attualmente in uso presso ACI, i dataset prodotti con il sistema di interrogazione precedentemente descritto, o estratti direttamente dalle sorgenti relazionali, vengono successivamente annotati semanticamente rispetto ai predicati dell'ontologia e codificati in formato RDF. In un possibile sviluppo futuro, il processo di annotazione e creazione del dataset RDF potrà essere gestito direttamente da Mastro durante l'elaborazione delle interrogazioni.

I dataset RDF così annotati sono distribuibili in modalità aperta con il massimo livello di qualità per gli Open Data: le 5 stelle[8]. Essi sono infatti in forma strutturata e in formato non proprietario, ma codificati con linguaggi standard del W3C (RDF), annotati semanticamente tramite l'ontologia, ed infine collegati (*linked*) a dati di altre organizzazioni, in particolare a quelli pubblicati da GeoNames e da ISPRA[9]. Per costruire questi link semantici è stato utilizzato l'*Entity Name System* (ENS) di OKKAM, un sistema per gestire identificativi univoci e persistenti di entità in sistemi di dati distribuiti.

4 Il portale di ACI per i Linked Open Data

Uno dei principali risultati del progetto è stata la realizzazione di un portale in cui ACI espone l'ontologia ed alcuni dati in formato aperto[10]. Il portale è stato sviluppato attraverso l'utilizzo del sistema Mastro Studio[11]. Questo strumento fornisce un ambiente software in cui l'utente finale può ispezionare l'ontologia, sia nella versione Graphol che OWL, leggere la sua documentazione in formato *wiki* (quindi arricchita di collegamento ipertestuali per facilitarne la navigazione), ed accedere ai dati pubblicati. Il sistema ha una sezione dedicata a quest'ultima funzionalità, in cui ogni statistica è disponibile come dataset, in formato RDF o CSV. Accedendo alla pagina di un dataset, l'utente può visualizzare la sua descrizione, il tipo di licenza sotto la quale il dataset è distribuito ed i file relativi ai vari formati in cui è disponibile. Inoltre, ogni risorsa pubblicata nei dataset di ACI è associata ad una URI, che corrisponde ad una pagina di descrizione della risorsa che la URI identifica; nel portale è quindi possibile ispezionare gli open data attraverso una navigazione web guidata da tali URI.

Mastro Studio è una applicazione web basata sul CMS (Content Management System) per Open Data DKAN[12], e comprende i moduli core di DKAN ed una suite di moduli *custom*, i.e., estensioni del CMS.Per la realizzazione delle sue funzionalità di accesso semantico ai dati, Mastro Studio si affida a Mastro attraverso una interfaccia web-service.

Riferimenti bibliografici

[Console *et al.*, 2014] M. Console, J. Mora, R. Rosati, V. Santarelli, e D. F. Savo. Effective computation of maximal sound approximations of description logic ontologies. In *Proc. of ISWC*, pages 164–179, 2014.

[De Giacomo *et al.*, 2012] G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, R. Rosati, M. Ruzzi, e D. F. Savo. MASTRO: A reasoner for effective ontology-based data access. In *Proc. of ORE*, 2012.

[Glimm *et al.*, 2014] B. Glimm, I. Horrocks, B. Motik, G. Stoilos, e Z. Wang. Hermit: An OWL 2 reasoner. *JAR*, 53(3):245–269, 2014.

[Kontchakov *et al.*, 2014] R. Kontchakov, M. Rezk, M. Rodriguez-Muro, G. Xiao, e M. Zakharyaschev. Answering SPARQL queries over databases under OWL 2 QL entailment regime. In *Proc. of ISWC*, pages 552–567, 2014.

[Lembo *et al.*, 2015] D. Lembo, J. Mora, R. Rosati, D. F. Savo, e E. Thorstensen. Mapping analysis in ontology-based data access: Algorithms and complexity. In *Proc. of ISWC*, pages 217–234, 2015.

[Lembo *et al.*, 2016] D. Lembo, D. Pantaleone, V. Santarelli, e D. F. Savo. Easy OWL drawing with the Graphol visual ontology language. In *Proc. of KR*, pages 573–576, 2016.

[Lenzerini, 2018] M. Lenzerini. Managing data through the lens of an ontology. *AI Magazine*, 39(2):65–74, 2018.

[Poggi *et al.*, 2008] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, e R. Rosati. Linking data to ontologies. *J. on Data Semantics*, X:133–173, 2008.

[6] http://obdasystems.com/mastro-protege-plugin

[7] https://www.w3.org/TR/r2rml/

[8] https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data

[9] http://dati.isprambiente.it/

[10] lod.aci.it

[11] https://www.obdasystems.com/mastrostudio

[12] https://getdkan.org/