SBML2SHACL

Edoardo De Matteis
1746561

10 settembre 2020

Indice

1 Introduzione **2**
1.1 SBML 2
1.2 Logiche descrittive 2
1.3 RDF 3
1.4 SHACL 4
1.5 Obiettivi 4

2 Modellazione **5**

3 Parsing da SBML a SHACL **9**

4 Parsing da SHACL a SBML **10**

5 Esempio **11**

6 Commenti e critiche **12**

7 Fonti **13**

1 Introduzione

1.1 SBML

Systems Biology Markup Language (SBML) è un formato di rappresentazione basato su XML per la modellazione di processi biologici e chimici. Non è un linguaggio bensì una *lingua franca* tra tool che utilizzano formati di rappresentazione differenti e ad oggi è *de facto* uno standard per i modelli computazionali in biologia.

Ai fini di questo progetto distinguiamo due differenti SBML:

- **SBML Level 3**. La specifica del *Core* che offre le funzionalità base di SBML, in questo documento lo si indica come "SBML 3".
- **Hierarchical Model Composition**. Noto come "comp" (namespace del package e quindi parola chiave necessaria in un file XML) questo package offre la possibilità di include modelli e sottomodelli uno dentro l'altro permettendo allo sviluppatore ed eventuali tool di:

  1. Scomporre modelli troppo grandi in modelli più piccoli per ridurne la complessità.
  2. Avere istanze multiple di un modello - precedentemente definito - all'interno di altri modelli, evitando ripetizione di codice.
  3. Creare librerie di codice riusabile e modelli considerati corretti.

  In questo documento lo si indica come "Extended SBML".

1.2 Logiche descrittive

La logica del primo ordine (FOL) aggiunge espressività alla logica proposizionale introducendo predicati, funzioni e quantificatori e permettendo di modellare eventi dinamici nel tempo al prezzo di un'inferenza semidecidibile, data infatti una knowledge base KB e una formula α è sempre possibile scrivere un algoritmo che risponda affermativamente se $KB \models \alpha$ ma se $KB \not\models \alpha$ l'esecuzione potrebbe non terminare. Le **description logics** (DL) sono una famiglia di linguaggi formali per rappresentazione della conoscenza spesso più espressivi della logica proposizionale ma meno della logica del primo ordine e per alcuni di questi esistono algoritmi corretti e completi per l'inferenza, più il linguaggio diventa espressivo più è difficile che l'inferenza sia decidibile.

La sintassi delle DL è fatta in modo tale da rendere facile la modellazione di definizioni e proprietà di categorie (collezioni di "oggetti" utilizzate per la rappresentazione di conoscenza su larga scala) tramite operazioni insiemistiche piuttosto che con una caratterizzazione logica, tra FOL e DL cambia dunque la sintassi, ciò che in FOL è **costante**, **predicato unario** e **predicato binario** nelle DL diventa rispettivamente **individuo**, **concetto** e **ruolo**.

In tabella 2 $\Delta^{\mathcal{I}}$ è il dominio di interpretazione data una funzione di interpretazione $\mathcal{I}$; i simboli invece sono elementi di una nomenclatura convenzionale - ma informale - per cui un linguaggio viene chiamato in base a ciò che esprime.

Nome	**Sintassi**	**Semantica**	**Simbolo**
Top	$\top$	$\Delta^{\mathcal{I}}$	$\mathcal{AL}$
Bottom	$\bot$	$\emptyset$	$\mathcal{AL}$
Intersezione	$C \sqcap D$	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$	$\mathcal{AL}$
Unione	$C \sqcup D$	$C^{\mathcal{I}} \cup D^{\mathcal{I}}$	$\mathcal{U}$
Negazione	$\neg C$	$\Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$	$\mathcal{C}$
Restrizione	$\forall R.C$	$\{a \in \Delta^{\mathcal{I}} \mid \forall b.(a,b) \in R^{\mathcal{I}} \rightarrow b \in C^{\mathcal{I}}\}$	$\mathcal{AL}$
Esistenziale	$\exists R.C$	$\{a \in \Delta^{\mathcal{I}} \mid \exists b.(a,b) \in R^{\mathcal{I}} \wedge b \in C^{\mathcal{I}}\}$	$\mathcal{E}$

Tabella 2: Paronamica di alcuni elementi della sintassi nelle DL.

Dal momento che nelle DL si eseguono operazioni insiemistiche non si parla propriamente di inferenza ma di sussunzione (subsumption): $C \sqsubseteq D$ equivale in FOL a $C \rightarrow D$ similmente al model checking in logica proposizionale dato che $\alpha \models \beta \longleftrightarrow M(\alpha) \subseteq M(\beta)$ - la funzione $M(\phi)$ restituisce un modello della formula ϕ - e per il teorema di deduzione $\alpha \models \beta$ equivale a $\alpha \rightarrow \beta$.

Le DL si basano su un'interfaccia "tell and ask" e si fanno sussunzioni analizzando TBox e ABox della KB: il primo definisce una conoscenza intensionale ed è composto quindi da definizioni di concetti, nell'ABox invece vi sono affermazioni su individui (membership assertions) e si definsice quindi una conoscenza estensionale del dominio di interesse.

1.3 RDF

Il **Resource Description Framework (RDF)** è una famiglia di specifiche definite dal World Wide Web Consortium (W3C) utilizzate per la modellazione di informazioni implementate come risorse web. RDF è simile ad altri approcci classici come database relazionali o logiche descrittive nei quali si hanno affermazioni sulle risorse sotto forma di triple soggetto-predicato-oggetto; il soggetto è la risorsa stessa e il predicato esprime una relazione tra soggetto e oggetto.

Se ad esempio volessimo esprimere il concetto "il cielo è blu" avremmo:

soggetto: "il cielo"

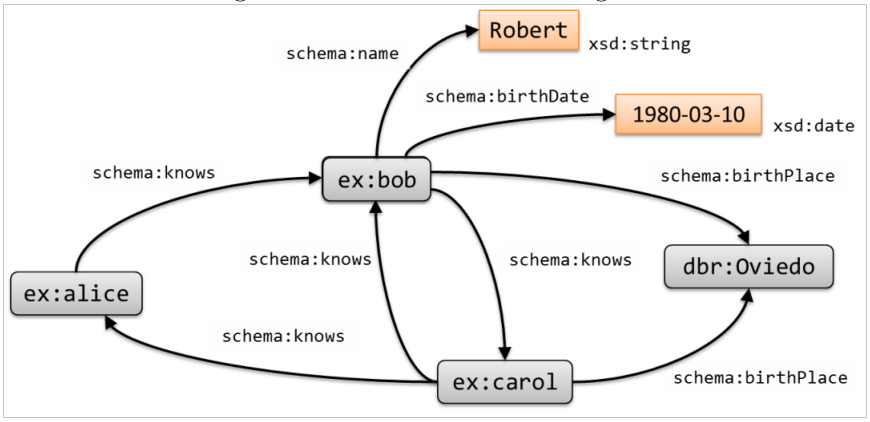
predicato: "è di colore"

oggetto: "blu"

Contrariamente all'approccio object oriented entità-attributo-valore in cui avremmo un oggetto "cielo" con attributo "colore" di valore "blu". Come si può vedere nell'immagine 1 un insieme di triple rappresenta intrinsecamente un grafo: una rappresentazione decisamente conveniente ma spesso nella pratica un modello RDF è mantenuto come database relazionale (chiamato Triplestore o Quad store).

Esistono vari formati per esprimere il modello astratto RDF, in questo progetto ho scelto di utilizzare Turtle (`.ttl`).

Figura 1: Un modello RDF è un grafo.



1.4 SHACL

Shapes Constraint Language (SHACL) è una specifica del W3C per la verifica di informazioni sotto forma di grafi dato un insieme di vincoli. Il processo di verifica prende in input un grafo contenente le dichiarazioni delle shape ovvero le definizioni cui i nodi, le istanze, dovranno essere verificati; in input si può avere qulsiasi formato RDF, in questo progetto si userà Turtle.

SHACL corrisponde ad una logica descrittiva molto estesa - $\mathcal{ALCOIQ}(\circ)$ - che è corretta, non completa, senza garanzie di terminazione e NEXPTIME-hard; le premesse non sembrano ottimistiche ma è possibile rendere meno espressivo il linguaggio per ottenerere inferenze decidibili. Ad esempio escludendo la composizione di ruoli $\circ$ si ha $\mathcal{SROIQ}$ ($\mathcal{S}$ sta per $\mathcal{ALC}$ con ruoli transitivi) per la quale esiste un algoritmo d'inferenza corretto, non completo, decidibile e NEXPTIME-hard; escludendo oltre la path concatenation (equivalentemente composozione di ruoli data la corrispondenza tra modello e grafo in RDF) anche la possibilità di esprimere ruoli inversi si ottiene la DL $\mathcal{ALCOQ}$ che è sound, completa, decidibile e PSPACE-complete e dal momento che NP $\subseteq$ PSPACE $\subseteq$ EXPTIME sicuramente non può esserci un peggioramento rispetto a prima.

Nello svolgimento di questo progetto ho preferito dividere in due file differenzi shape e istanze, questi corrispondono a TBox e ABox e ogniqualvolta si verifica la correttezza di un modello con `shacl_verifier.py` si eseguono inferenze, ad essere più precisi nella TBox si dovrebbe parlare di classificazione e instance checking ma sono entrambi ricondubili a casi specifici di sussuzione.

SHACL include anche il linguaggio di query **SPARQL Protocol and RDF Query Language (SPARQL)**.

1.5 Obiettivi

Lo scopo di questo progetto, sotto la guida del Professor Tronci, è quello di convertire automaticamente una specifica SBML in SHACL. A tal fine il problema

è stato diviso in tre fasi:

1. Modellare in SHACL un sottoinsieme di costrutti che definiscono SBML .
2. Scrivere un parser che traduca una specifica SBML in SHACL.
3. Scrivere un parser che traduca una specifica SHACL in SBML.

Si assume a priori che i file SBML in input siano corretti, la verifica è eseguibile online. Per il corretto funzionamento del codice inoltre è necessario il download del package Python *rdflib*.

2 Modellazione

In questa prima fase si è scelto un sottoinsieme di SBML 3.2 più i costrutti in Extended SBML e ne è stato formalizzato un modello. Nella tabella 3 vengono descritti i principali costrutti, sono state volontariamente omesse shape quali `listOf*` il cui significato è intuitivo e avrebbero solo reso la tabella meno leggibile, in ogni caso è possibile consultare ulteriormente il diagramma `diagram.png`, mostrato anche in figura 2, e il file delle shape `shapes.ttl`. È possibile verificare la consistenza di un file di nodi rispetto al modello tramite lo script `shacl_verifier.py`.

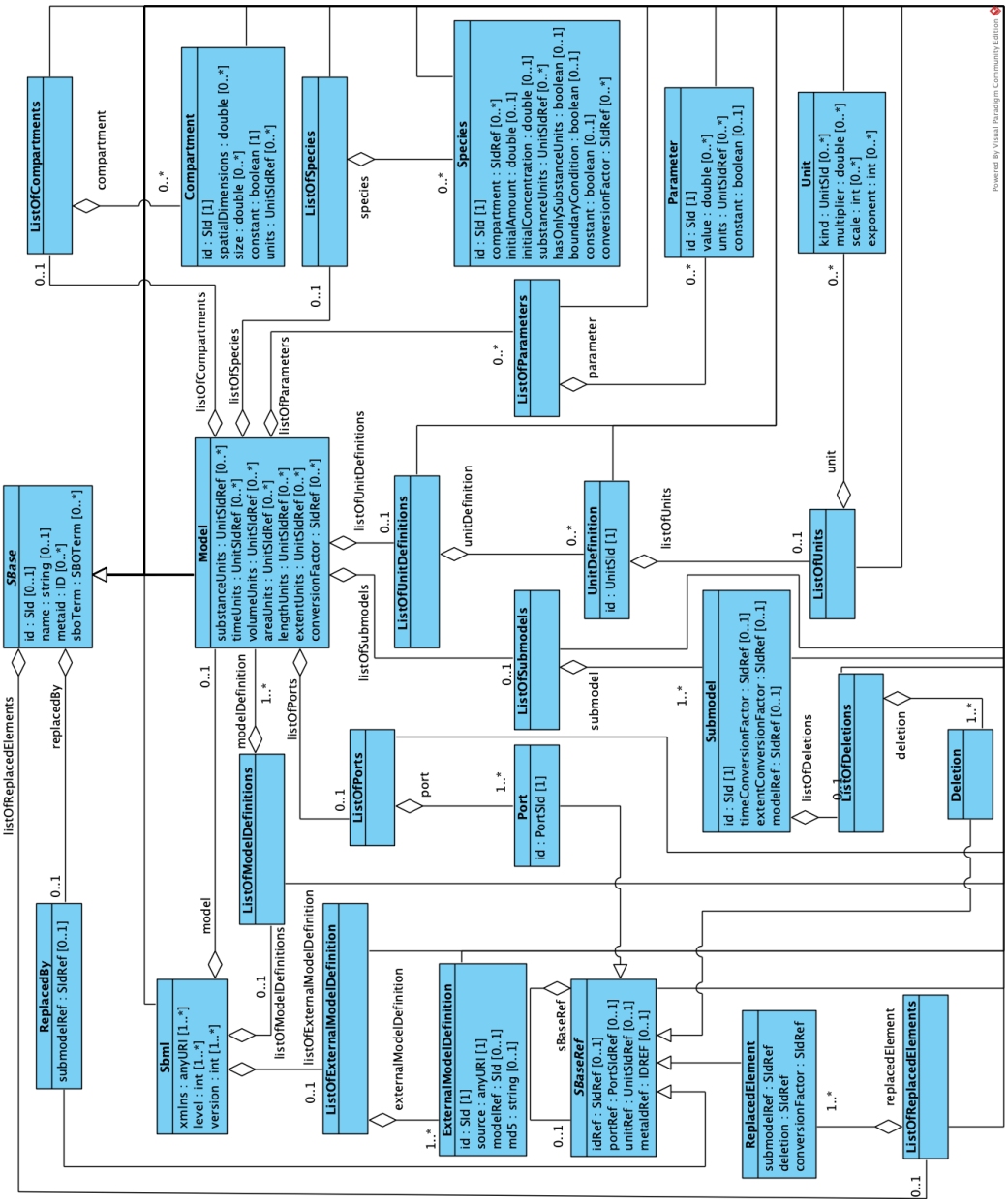
Entità	**Descrizione**
	SBML 3.2
SBase	Classe astratta superclasse di ogni altra classe. Per quanto in SBML sia considerata a tutti gli effetti un tipo, poiché non esistono attributi di tipo SBase a questa classe si riserva un trattamento differente rispetto ai tipi composti quali ID, SId, etc . . . che sono state omesse dal diagramma ma sono comunque presenti nel file `shapes.ttl`.
Sbml	Ogni file SBML ha un'etichetta con tag `sbml`, data la sua obbligatorietà ed unicità è possibile costruire grafi SHACL composti da multipli modelli SBML ed esplorarli radicando l'albero in Sbml, è importante specificare che questo nuovo modello composto non sarà legale in SBML e non si dà alcuna garanzia nella fase di riconversione in SBML. Il costrutto Sbml ammette altri attributi oltre quelli noti, siccome in SHACL ho necessità di conoscerne il tipo è stata persa questa possibilità.

Model	Rappresenta il modello, se ne può avere più di uno come detto sopra nonostante SBML non lo permetta.
Unit	Definisce un'unità di misura definita dall'utente, in SBML sono definite delle unità base (i.e. le unità del SI e altre scelte dagli sviluppatori di SBML) e combinandole opportunamente tra loro è possibile definire nuove unità di misura (e.g l'accelerazione $\frac{m}{s^2}$ che altro non è se non $m^1 s^{-2}$). Nel modello in `shapes.ttl` le unità base (kind) sono trattate come normalissime unità di misura, in SBML non è concesso avere come attributo kind un'unità di misura che non sia base e questa libertà non genera problemi.
Compartment	Rappresenta un insieme di entità biologiche.
Species	Rappresenta un'entità biologica.
Parameter	In SBML si possono definire parametri sia locali che globali, il modello presentato in queso progetto non implementa la località quindi i parametri sono sempre globali e in Model l'attributo `ListOfParameters` ha moltiplicità $[0, n]$ piuttosto che $[0, 1]$.
Extended SBML	
ExternalModelDefinition	Necessario per importare modelli esterni.
ModelDefinition	Questo costrutto fornisce la definizione di un modello.
Submodel	L'istanza di una ModelDefinition è rappresentata da un Submodel, si ha quindi un modello dentro ad un altro modello. Durante la fase di test per extended SBML non ho usato esempi scaricati da Biomodels perché la gerarchia non veniva rappresentata tramite il costrutto Submodel ma con un attributo `outer` in Compartment, con il Professor Tronci si è ritenuto fosse meglio attenersi allo standard W3C.

Port	Una port permette allo sviluppatore di definire come ci si deve interfacciare con un modello, per quanto non siano vincolanti di norma è preferibile seguire le indicazioni dello sviluppatore.
Deletion	Non è detto che i modelli importati abbiano solo ed esclusivamente componenti desiderabili e con una deletion è possibile ignorare quelli indesiderati.
Replacement	Tramite questo costrutto è possibile sostituire una componente con un'altra, ogni riferimento alla prima ora punta alla seconda. A causa di Replacement sono presenti più parser: il primo `parser.py` esplora il file XML come una lista ed `extended_parser.py` come un albero.
SBaseRef	Port, Deletion, Replacement e Submodel utilizzano dei riferimenti, SBaseRef similmente a SBase offre alle proprie sottoclassi degli attributi comuni a tutte.

Tabella 3: Modellazione SHACL

Figura 2: Modellazione SBML



3 Parsing da SBML a SHACL

In questa seconda fase dato uno o più file SBML in input li si traduce in SHACL tramite un parser e si verifica la correttezza dell'output tramite il file `shacl_verifier.py`. Come già detto si hanno due parser: `parser.py` è la prima versione e non supporta Extended SBML contrariamente a `extended_parser.py` che è la versione finale.

In `parser.py` si riescono a tradurre senza problemi file in SBML 3.2 ma poiché il file XML viene letto come una lista di etichette risulta macchinoso (ammesso e non concesso sia possibile) fare riferimento all'etichetta padre di una sostituzione o una cancellazione. Il problema non si presenta in `extended_parser.py` dato che il file XML viene esplorato come un albero, questa seconda versione è inoltre più breve (444 righe anziché 932) ed è più facilmente estendibile. Questo parser infatti legge qualsiasi tag del file XML e vi associa i propri attributi - noti - al rispettivo valore, ho comunque imposto dei controlli per evitare questo comportamento dato che avrebbe solo rallentato la fase di verifica. Inoltre è più veloce come si può vedere in tabella 5: il calcolo delle performance è basato sull'esecuzione di `test.sh`, sono stati scaricati da Biomodels i 18 modelli SBML presenti nella cartella `examples/input/biomodel` e di seguito elencati, in `test.sh` i parser prendono in input due file così da verificare la correttezza della costruzione di un grafo a partire da più modelli e allo stesso tempo si ha un numero più consistente di test ovvero 324.

1. BIOMD0000000087.xml
2. BIOMD0000000105.xml
3. BIOMD0000000399.xml
4. BIOMD0000000474.xml
5. BIOMD0000000476.xml
6. BIOMD0000000559.xml
7. BIOMD0000000562.xml
8. BIOMD0000000619.xml
9. BIOMD0000000624.xml
10. BIOMD0000000705.xml
11. BIOMD0000000706.xml
12. MODEL1012110001.xml
13. MODEL1012220002.xml
14. MODEL1012220003.xml

15. MODEL1012220004.xml
16. MODEL1112260002.xml
17. MODEL1812100001.xml
18. MODEL3632127506.xml

File	**Time (m)**
parser.py	42:54.54
extended_parser.py	17:35.19

Tabella 5: Risultati ottenuti eseguendo `time ./test.sh` su macOS Catalina 10.15.6 con processore Intel Core i5 1.6 GHz Dual-Core.

Nella sottocartella `custom` invece vi sono file d'esempio forniti direttamente dal W3C per Extended SBML, potranno essere usati solo per verificare `extended_parser.py`. In Biomodels per esprimere una gerarchia tra compartment si fa uso di un attributo `outer` e non di sottomodelli come nella specifica del W3C.

4 Parsing da SHACL a SBML

Nella terza e ultima fase si esegue la controprova della precedente ovvero si traduce il file di istanze SHACL ottenuto in un file SBML e ci si attende che venga modellata la stessa conoscenza, il file non sarà uguale perché privo di eventuali costrutti non modellati. Nella cartella `query` sono presenti tre file:

- `query.py`: Permette di interrogare il modello con query SPARQL, la query deve essere scritta direttamente nel codice perché gli argomenti della print variano in base alla query.
- `ttl2sbml.py`: Il parser da SHACL a SBML.
- `ttl2xml`: Converte un file di istanze Turtle in XML, una query e un file XML possono esprimere la stessa conoscenza: una tripla `<subject> <predicate> <object>` si può tradurre in XML nelle etichette `<subject> <predicate> <object/> </predicate> </subject>`.

Nella scrittura del parser è risultato necessario poter riottenere la struttura annidata dell'XML che era andata persa nella scrittura delle istanze SHACL, per farlo ho definito delle classi Tree e Node, nella fase di lettura del file l'albero viene costruito gradualmente e una volta ottenuto basta esplorarlo ricorsivamente per costruire il testo XML, ricordandosi chiaramente di chiudere le etichette una volta terminate le chiamate ai figli.

L'algoritmo genera la stessa conoscenza dei file di partenza, per SBML 3.2 non ci sono problemi invece per i file nella cartella `custom` tutti tranne `MANUAL_1.xml` non superano la verifica SBML perché trattandosi di file d'esempio i riferimenti non sono reali; chiaramente anche l'output di `ttl2sbml.py` dà risultato negativo, per `MANUAL_1.txt` invece si ha lo stesso esito ovvero numerosi warning.

5 Esempio

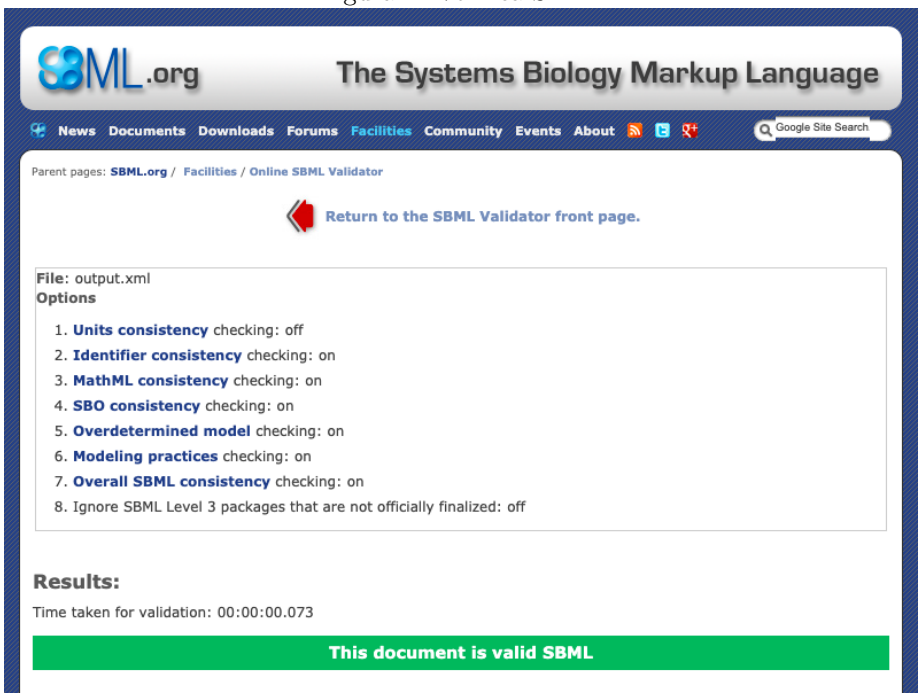
Un esempio di esecuzione è quello nella cartella `examples/output`, è stato ottenuto eseguendo i comandi in figura 3 e in figura 4 si può vedere come SBML verifichi la traduzione da SHACL a SBML.

Figura 3: Comandi

```
bash-3.2$ python ./project/parser/extended_parser.py -f ./examples/input/biomodel/BIOMD0000000087.xml -o ./examples/output/output.ttl
bash-3.2$ python ./project/model/shacl_verifier.py -s ./project/model/shapes.ttl -d ./examples/output/output.ttl
True
Validation Report
Conforms: True

bash-3.2$ python ./project/query/ttl2sbml.py -f ./examples/output/output.ttl -o ./examples/output/output.xml
bash-3.2$ python ./project/query/query.py -f ./examples/output/output.ttl
http://example.org/ns#compartment_10 a schema:Compartment
http://example.org/ns#compartment_12 a schema:Compartment
bash-3.2$
```

Figura 4: Verifica SBML



6 Commenti e critiche

Durante lo sviluppo del progetto si è tenuto conto dei seguenti commenti e critiche da parte del Professor Tronci:

Data	**Descrizione**
22/07/20	Videochiamata con specifica del problema da parte del Professor Tronci.
27/07/20	Prima stesura di una modellazione dei costrutti in SBML, avendo dimenticato di modellare extended SBML mi è stato fatto notare dal Professore e ho risolto, mi è stato anche consegnato il materiale su extended SBML da consultare.
28/07/20	Corretto il punto precedente non avevo implementato deletion e replacement, il Professore ha inoltre consigliato la strategia di testing.
31/07/20	Dopo aver aggiunto replacement e deletion e dopo aver eseguito dei test con risultati positivi è seguita una videochiamata con il Professor Tronci in cui mi è stato indicato di modificare il parser in maniera tale da poter ricevere in input più modelli SBML e di capire se il risultato di una query SPARQL rappresenti lo stesso tipo di conoscenza di un file XML/SBML, così da poter sfruttare questa corrispondenza per eseguire dei controtest. Questa corrispondenza esiste ed è usata spesso sotto il nome di "SPARQL query results xml format".
06/08/2020	Videochiamata con il Professor Tronci in cui si è deciso di comprendere se fosse possibile scrivere un parser da SHACL a SBML e, in caso di risposta affermativa, farlo.
19/08/2020	Correzioni su questa relazione ampliandola affinché i concetti risultino più chiari durante l'esposizione.

7 Fonti

- The systems biology markup language.
- Sparql query language for rdf, 2008.
- Sparql query results xml format (second edition), 2013.
- Rdf 1.1 concepts and abstract syntax, 2014.
- Shapes constraint language (shacl), 2017.
- Michael Hucka, Frank T. Bergmann, Claudine Chaouiya, Andreas Dräger, Stefan Hoops, Sarah M. Keating, Matthias König, Nicolas Le Novère,Chris J. Myers, Brett G. Olivier, Sven Sahle, James C. Schaff, RahumanSheriff, Lucian P. Smith, Dagmar Waltemath, Darren J. Wilkinson, andFengkai Zhang. The systems biology markup language (sbml): Languagespecification for level 3 version 2 core, 2019.
- Lucian P. Smith, Stefan Hoops, Martin Ginkel, Ion Moraru, Michael Hucka,Andrew Finney, Chris J. Myers, and Wolfram Liebermeister. Hierarchicalmodel composition, 2013.
- Deciding SHACL Shape Containment through Description Logics Reasoning (Extended Version)
- SHACL Satisfiability and Containment
- The Description Logic Handbook: Theory, Implementation, and Applications
- Complexity of reasoning in Description Logics
- Description Logic Terminology