



**UNIVERSITÀ
DI TORINO**

di.unito.it

**DIPARTIMENTO
DI INFORMATICA**

UNIVERSITÀ DI TORINO DIPARTIMENTO DI INFORMATICA
CORSO DI LAUREA TRIENNALE IN INFORMATICA

Editing di ontologie tramite il linguaggio di programmazione funzionale CDuce

Relatore
Professoressa Viviana Bono

Laureando
Davide Camino
Matricola: 897753

ANNO ACCADEMICO 2022-2023

Data di laurea Novembre 14, 2023

A tutti quelli che mi vogliono bene.

*L'informatica riguarda i computer
non più di quanto l'astronomia riguardi i telescopi.
E. W. Dijkstra*

Dichiaro di essere responsabile del contenuto dell'elaborato che presento al fine del conseguimento del titolo, di non avere plagiato in tutto o in parte il lavoro prodotto da altri e di aver citato le fonti originali in modo congruente alle normative vigenti in materia di plagio e di diritto d'autore. Sono inoltre consapevole che nel caso la mia dichiarazione risultasse mendace, potrei incorrere nelle sanzioni previste dalla legge e la mia ammissione alla prova finale potrebbe essere negata.

ABSTRACT

Questo lavoro illustra lo sviluppo di strumenti per l'editing di ontologie tramite il linguaggio di programmazione funzionale CDuce, in particolare la realizzazione di programmi per il refactoring, il merge e la traduzione da tesauri a ontologie. Lo scopo dello studio è quello di valutare se, e in che condizioni, lo sviluppo di strumenti più o meno ad hoc per l'editing di ontologie sia vantaggioso rispetto all'uso di strumenti grafici tradizionali come Protégé. Nella valutazione si tengono in considerazione principalmente la difficoltà tecnica e il tempo di sviluppo degli strumenti creati con CDuce e il tempo necessario e la ripetitività per fare le stesse modifiche con strumenti grafici.

Indice

ABSTRACT	iii
1 Concetti di base	1
1.1 Basi di conoscenza	1
1.1.1 Ontologie	1
1.1.2 Tesauri	3
1.2 Strumenti di editing	4
1.2.1 Protégé	4
1.2.2 CDuce	4
1.2.3 Feature	5
1.3 Metalinguaggi	6
1.3.1 RDF/RDFS	6
1.3.2 OWL	7
1.3.3 SKOS	7
2 Strumenti offerti da CDuce	8
2.1 Parsing di documenti XML	8
2.1.1 Esempio	8
2.2 Manipolare documenti XML	9
2.2.1 Esempio	10
2.3 Query	10
2.3.1 Esempio	11
3 Trasformare un tesauro in un'ontologia	13
3.1 Vantaggi del passaggio da tesauro ad ontologia	13
3.1.1 Europeana Fashion Thesaurus: capturing imagination	13
3.1.2 Svantaggi del tesauro	14
3.2 Struttura del tesauro	14
3.3 Struttura dell'ontologia	15
3.4 Da concetto SKOS a classe OWL	15
3.4.1 Trasformare gli attributi	16
3.4.2 Trasformare una singola classe	16
3.5 Costruire la nuova ontologia	17
3.6 Versione compatta	18
3.7 Aumentare l'espressività	18
3.8 In Protégé	19
3.9 Conclusioni	19

4	Merge di ontologie	20
4.1	Obiettivo del merge	20
4.1.1	Ontologie di partenza	20
4.1.2	Ontologia di arrivo	21
4.2	Struttura generica di un'ontologia	21
4.3	Merge	22
4.3.1	Funzioni utili	22
4.3.2	Selezione degli abiti	24
4.3.3	Costruzione dell'ontologia	25
4.3.4	Risultato finale	27
4.4	In Protégé	27
4.5	Conclusioni	27
5	Conclusioni	29
5.1	Maturità di CDuce	29
5.1.1	Il progetto CDuce	29
5.1.2	Stato di sviluppo attuale	29
5.1.3	Difficoltà incontrate	30
5.1.4	Sviluppi futuri	31
5.2	Punti di forza di CDuce	31
5.2.1	Sistema di tipi	32
5.2.2	Funzioni di ordine superiore	32
5.2.3	Pattern matching	32
5.3	Spunti per paragoni futuri	32
5.3.1	Confronto con un linguaggio imperativo	33
5.3.2	Confronto con un linguaggio funzionale	33
5.3.3	Confronti possibili	33
5.4	Sviluppo di nuovi strumenti	34
5.4.1	Criticità di CDuce	34
5.4.2	Interfaccia grafica	34
5.5	Conclusioni	35

Capitolo 1

Concetti di base

Introduzione

Qui illustriamo alcuni dei concetti di base che serviranno per comprendere il resto della discussione, daremo una definizione di ontologia e tesaurus, descriveremo brevemente gli strumenti utilizzati e i linguaggi con cui si descrivono le basi di conoscenza che tratteremo.

1.1 Basi di conoscenza

Una base di conoscenza (Knowledge Base KB) è un insieme di affermazioni, ognuna delle quali espressa attraverso un linguaggio di rappresentazione della conoscenza. Le affermazioni esprimono concetti riguardanti il dominio di interesse.

Perché una base di conoscenza possa essere utile deve essere possibile aggiungere conoscenza (fare affermazioni) e interrogare la KB; queste operazioni possono coinvolgere meccanismi di inferenza che permettano di ricavare nuove affermazioni da quelle già note. L'inferenza sfrutta una logica formale per ricavare nuove informazioni partendo da quelle che sono già presenti nella KB.[\[13\]](#)

1.1.1 Ontologie

Definizione

In accordo con una definizione ampiamente accettata [\[7\]](#) un'ontologia ha lo scopo di rappresentare un vocabolario per definire i concetti di un particolare dominio di interesse condiviso, ed è costituita da definizioni di classi, relazioni, funzioni e altri oggetti utili a rappresentare la conoscenza[\[6\]](#).

La definizione è ancora un po' vaga, più precisamente un'ontologia è una base di conoscenza che permette di descrivere concetti e relazioni tra di essi specificando questi oggetti tramite un linguaggio di rappresentazione della conoscenza basato su una logica formale.

Scopo delle ontologie

Nel contesto del web semantico¹ le ontologie sono il mezzo principale per condividere, integrare e scoprire dati[\[7\]](#).

¹Un'estensione del World Wide Web in cui le informazioni siano comprensibili ad un automa[\[4\]](#)

Manipolare ontologie

Dato lo scopo delle ontologie risulta essere un argomento centrale la possibilità di riutilizzare, modificare e ampliare ontologie esistenti. Spesso le basi di conoscenza sono sì strutturate, ma sono eterogenee e non permettono interoperabilità. Scopo di questo lavoro è quello di presentare degli strumenti per manipolare basi di conoscenza (ci concentriamo su tesauri e ontologie) operando in modo tale da rendere compatibili informazioni tratte da fonti differenti.

Il testo [16] presenta una trattazione teorica approfondita sulle metodologie per sviluppare un'ontologia, in particolare il framework NeOn prevede vari scenari tipici in cui ci si può trovare quando si voglia costruire un'ontologia, si fa particolare attenzione ai casi nei quali siano già presenti informazioni, ma queste vanno riorganizzate.

Esempio

Consideriamo una semplice ontologia che rappresenta persone con legami di parentela genitore-figlio; le persone hanno uno o più nomi salvati nel tag `comment`. Modelliamo questa ontologia con una classe `Persone` e una sottoclasse `Genitori` (i cui individui sono `Persone` che realizzano la relazione `genitoreDi`). Creiamo la relazione `genitoreDi`. Infine popoliamo l'ontologia con alcuni individui. Il risultato ottenuto con Protégé è un documento XML di questo tipo:

Listato 1.1: persone.rdf

```
1  <?xml version="1.0"?>
2  <rdf:RDF xmlns="http://www.persone/"
3      xml:base="http://www.persone/"
4      xmlns:owl="http://www.w3.org/2002/07/owl#"
5      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
6      xmlns:www="http://www.persone#"
7      xmlns:xml="http://www.w3.org/XML/1998/namespace"
8      xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
9      xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
10     <owl:Ontology rdf:about="http://www.persone"/>
11     <!-- http://www.persone#genitoreDi -->
12     <owl:ObjectProperty rdf:about="http://www.persone#genitoreDi"/>
13     <!-- http://www.persone#Genitori -->
14     <owl:Class rdf:about="http://www.persone#Genitori">
15         <owl:equivalentClass>
16             <owl:Restriction>
17                 <owl:onProperty>
18                     ↪ rdf:resource="http://www.persone#genitoreDi"/>
19                 <owl:someValuesFrom>
20                     ↪ rdf:resource="http://www.persone#Persone"/>
21             </owl:Restriction>
22         </owl:equivalentClass>
23         <rdfs:subClassOf rdf:resource="http://www.persone#Persone"/>
24     </owl:Class>
25     <!-- http://www.persone#Persone -->
26     <owl:Class rdf:about="http://www.persone#Persone">
27         <!-- http://www.persone#Bruto -->
```

```

26 <owl:NamedIndividual rdf:about="http://www.persone#Bruto">
27   <rdf:type rdf:resource="http://www.persone#Persone"/>
28   <rdfs:comment>Bruto</rdfs:comment>
29 </owl:NamedIndividual>
30 <!-- http://www.persone#Cesare -->
31 <owl:NamedIndividual rdf:about="http://www.persone#Cesare">
32   <rdf:type rdf:resource="http://www.persone#Genitori"/>
33   <www:genitoreDi rdf:resource="http://www.persone#Bruto"/>
34   <rdfs:comment>Caio</rdfs:comment>
35   <rdfs:comment>Augusto</rdfs:comment>
36   <rdfs:comment>Giulio</rdfs:comment>
37   <rdfs:comment>Cesare</rdfs:comment>
38 </owl:NamedIndividual>
39 </rdf:RDF>

```

Per quanto non sia impossibile leggere la struttura e i dati dal listato precedente un modo naturale per rappresentare le ontologie è sotto forma di grafi. Qui vediamo il grafo² che mostra la struttura dell'ontologia e possiamo apprezzare quanto sia semplice la sua struttura rispetto a quello che avremmo potuto immaginare dal listato 1.1. Nel grafo non sono rappresentati gli individui, possiamo comunque leggere quanti ve ne sono per ogni classe (in questo caso un individuo di tipo **Persone** e uno di tipo **Genitori**)

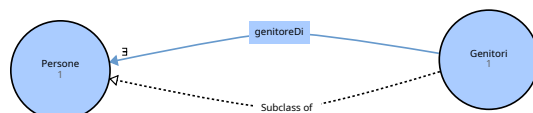


Figura 1.1: Grafo dell'ontologia persone

1.1.2 Tesauri

Definizione

Nella loro accezione più generale possibile i tesauri sono risorse nelle quali termini affini sono raggruppati assieme[10]. In particolare un tesoro fornisce un vocabolario preciso e controllato rispetto ad un particolare dominio di interesse.[15] Queste strutture possono aiutare il ricercatore a riformulare le strategie di ricerca fornendo una serie di sinonimi, contrari, definizioni e traduzioni in altre lingue del termine cercato.

Esistono diversi tipi di tesoro in base alle modalità di costruzione e fruizione [10], nel nostro caso il tesoro sarà costituito da un vocabolario tassonomico in cui le relazioni tra oggetti sono di tipo BT (broader term), cioè ogni concetto può avere un riferimento ad un concetto più generale, formando in questo modo una struttura ad albero.

Espressività

Come si può immaginare il potere espressivo di un tesoro è inferiore a quello di un'ontologia che non pone alcun limite alle relazioni definibili tra individui. In un tesoro inoltre ogni concetto ha al più un genitore, mentre in un ontologia possiamo creare una classe che erediti le caratteristiche da più classi distinte.

²tutti i grafi presenti un questo elaborato sono stati ottenuti grazie a <http://vowl.visualdataweb.org/webvowl.html>

1.2 Strumenti di editing

1.2.1 Protégé

Protégé³ è uno strumento per la modellazione della conoscenza molto conosciuto; è un progetto open-source sviluppato all'università di Stanford e permette la manipolazione interattiva di ontologie e KB attraverso un'interfaccia grafica e delle API java.

Le funzionalità di Protégé possono essere aumentate grazie a dei componenti plug-in il cui numero è in continua crescita. Questi plug-in offrono nuovi metodi per la gestione delle ontologie, supporto per dati multimediali, engines per il ragionamento automatico e per l'interrogazione delle basi di conoscenza[14].

Descrizione del software

Protégé presenta un'interfaccia con numerose schede, ogni scheda permette all'utente di accedere ad una differente funzionalità del software. Le schede base consentono di aggiungere nuova conoscenza e di effettuare ricerche nella KB.

Uno degli obiettivi fondamentali di Protégé è quello di rendere l'inserimento e la ricerca in una base di conoscenza il più semplice possibile: mentre il sistema genera uno strumento per l'acquisizione della conoscenza l'utente aggiunge informazioni riempiendo form intuitivi, selezionando elementi da delle liste e disegnando diagrammi.

Protégé permette di salvare le ontologie in numerosi formati, tra cui UML, XML, RDF e OWL.

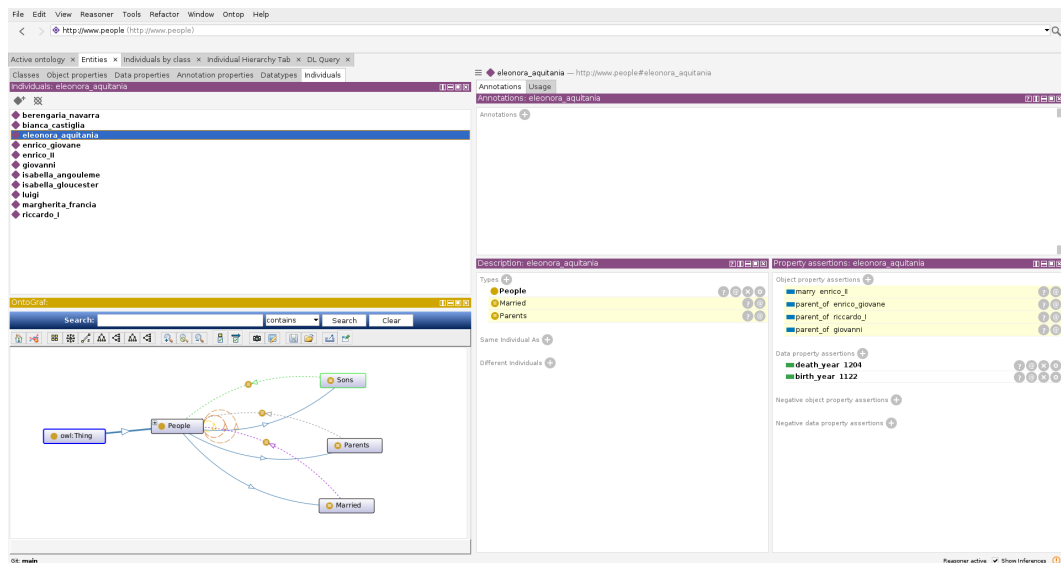


Figura 1.2: Interfaccia di Protégé

1.2.2 CDuce

CDuce è un linguaggio di programmazione funzionale, staticamente tipato e orientato allo sviluppo di applicazioni che lavorano su documenti XML[3], nasce con l'intento di migliorare XDuce: un altro linguaggio di programmazione per il processing di file XML[8].

³<https://protege.stanford.edu/>

XDuce permette di manipolare unicamente documenti XML, questa sua specificità mostra come la presenza di feature mirate possano aiutare e semplificare lo sviluppo di applicazioni XML sacrificando, però, la possibilità di scrivere applicazioni complesse in cui non si facciano solo semplici trasformazioni di elementi XML.

CDuce si pone l'obiettivo di integrare specifiche funzioni orientate alla manipolazione XML all'interno di un linguaggio potenzialmente general-purpose. Per fare questo CDuce usa un'algebra general-purpose, con costruttori di tipo standard mantenendo l'espressività e la potenza delle espressioni regolari di XDuce attraverso l'uso di tipi ricorsivi e combinatori logici (unione, intersezione, differenza).

In CDuce è possibile creare complesse strutture dati, modellare il tipo di documenti XML, e interfacciarsi in modo relativamente semplice con altri linguaggi di programmazione[3].

1.2.3 Feature

Vediamo brevemente le principali feature di CDuce, non entreremo nel dettaglio essendo presente una guida dettagliata sul sito del progetto⁴

Verifica statica dei tipi

Il sistema di tipi è un componente centrale di CDuce; dal punto di vista pratico un tipo è un insieme di valori identificati da un'espressione sintattica, questa descrizione di un tipo permette di definire in modo naturale (come insiemi matematici) relazioni di sottotipo, connettivi logici nell'algebra dei tipi ed equivalenza di tipi.

I tipi ricoprono un ruolo importante anche nella parte statica del linguaggio: la correttezza di ogni trasformazione definita in CDuce è garantita staticamente. Inoltre in CDuce il pattern matching viene effettuato attraverso un'esatta inferenza di tipo, nel senso che viene assegnata ad ogni variabile legata esattamente il set di tutti i valori che questa può assumere.

In questo modo è possibile costruire in sistema di tipi statico e molto preciso, che permette una migliore descrizione del comportamento dinamico del programma sviluppato

Pattern matching

È un'operazione fondamentale in CDuce ed ha la forma:

```
match e with
| p1 -> e1
...
| pn -> en
```

Si cerca di fare il match tra la valutazione di un'espressione *e* e vari pattern *pi*. Il primo pattern che fa il match con *e* attiva la corrispondente espressione sulla destra che può usare le variabili legate dal pattern.

Il controllo statico dei tipi si assicura che il pattern matching sia esaustivo, il tipo valutato per *e* deve essere sottotipo dell'unione dei tipi accettati dai pattern definiti sotto.

Funzioni

La forma generale di una funzione è

⁴<https://www.cduce.org/>

```

fun myfunc (t1 -> s1; ...; tn -> sn)
| p1 -> e1
...
| pm -> em

```

In cui la prima riga è l'interfaccia della funzione, il rimanente è detto corpo:

- interfaccia: rappresenta il modo in cui la funzione si comporta: quando riceve un elemento di tipo **ti** la funzione restituirà un elemento di tipo **si**, per ogni clausola **ti -> si** il sistema verifica che la funzione trasformi correttamente gli elementi;
- corpo della funzione: le funzioni operano per pattern matching degli argomenti: viene eseguita la prima trasformazione tale che l'argomento ricevuto come parametro faccia il match col pattern **pi**.

In CDuce è possibile definire anche applicazioni parziali (curried functions⁵) e funzioni di ordine superiore⁶

1.3 Metalinguaggi

Un metalinguaggio è in generale un linguaggio, o sistema di segni, naturale o artificiale, adottato per la descrizione della struttura formale di dati linguaggi⁷. In questo elaborato parliamo di linguaggi su tre differenti livelli:

- linguaggi di programmazione: sono i linguaggi con i quali sviluppiamo strumenti più o meno automatici per la manipolazione di basi di conoscenza;
- ontologie e tesauri: esprimono della conoscenza in modo formale, rispettano certe logiche per poter essere usate e restituiscono all'interrogatore una risposta che quest'ultimo sia in grado di comprendere
- linguaggi per descrivere le basi di conoscenza: abbiamo bisogno di un linguaggio per descrivere la struttura di una KB, per definire quali relazioni possono esserci e in generale per parlare le KB stessa; chiameremo questi ultimi metalinguaggi⁸ per distinguerli dai linguaggi di programmazione e dalla conoscenza che stiamo rappresentando.

Presentiamo alcuni metalinguaggi importanti che useremo per descrivere ontologie e tesauri.

1.3.1 RDF/RDFS

RDF (Resource Description Framework) è un framework⁹ per il processing di metadati¹⁰. RDF permette l'interoperabilità tra applicazioni che si scambiano informazioni comprensibili ad un automa.

Lo scopo di RDF è quello di automatizzare il processing di risorse Web; trova quindi applicazione nella ricerca di informazioni permettendo un motore di ricerca migliore, in applicazioni che catalogano la conoscenza, per descrivere una collezione di pagine che rappresentano un unico "documento" logico, ecc...[5]

⁵<https://en.wikipedia.org/wiki/Currying>

⁶https://en.wikipedia.org/wiki/Higher-order_function

⁷<https://www.treccani.it/vocabolario/metalinguaggio>

⁸questi linguaggi non ci permettono di descrivere il dominio di interesse di per sè ma descrivono come descrivere il dominio

⁹asd

¹⁰esattamente come i metalinguaggi i metadati sono dati che descrivono altri dati

1.3.2 OWL

L'espressività di RDF e RDFS è limitata, potendo descrivere solo predicati binari, una struttura delle classi gerarchica e proprietà anch'esse gerarchiche con un dominio e un codominio.

Sono state individuate dal gruppo di ricerca "Web Ontology Working Group" una serie di casi in cui le ontologie devono poter essere descritte con linguaggi più espressivi rispetto a RDF e RDFS^[1].

Presentiamo le funzionalità aggiunte da OWL Full: la versione più espressiva^[11] del linguaggio OWL; per una descrizione più approfondita del linguaggio si rimanda al documento prodotto da W3C a riguardo^[12].

OWL Full

Presenta la compatibilità massima coi documenti RDF/RDFS, ha la massima espressività, ma non è un sistema decidibile e completo. Descriviamo questo con il solo scopo di fornire l'elenco più completo possibile di funzionalità aggiunte allo scopo di aumentare l'espressività di RDF/RDFS.

OWL Full^[13] permette di:

- definire restrizioni sui domini delle relazioni in base alla classe (si può limitare ad esempio un individuo di classe mucca a mangiare solo piante, senza applicare la stessa restrizione ad altri animali);
- esprimere disgiunzione tra classi;
- definire classi tramite operazioni insiemistiche tra altre classi;
- definire restrizioni sulla cardinalità di una proprietà
- definire caratteristiche particolari di una proprietà come transitività, riflessività, proprietà inversa, ecc...

In OWL Full è, inoltre, possibile utilizzare ogni primitiva del linguaggio OWL combinandola con le primitive del linguaggio RDF/RDFS in modo da alterarne il significato originario (applicando anche una primitiva all'altra); quest'ultima caratteristica permette a OWL Full di essere completamente compatibile con RDF/RDFS e avere espressività massima, ma lo rende anche un sistema non decidibile.

1.3.3 SKOS

¹¹e anche più pesante computazionalmente

¹²<https://www.w3.org/TR/2004/REC-owl-features-20040210>

¹³come anche OWL DL

Capitolo 2

Strumenti offerti da CDuce

Introduzione

In questo capitolo analizziamo brevemente gli strumenti principali offerti da CDuce per l'editing di documenti XML. La trattazione vale per un qualsiasi documento XML, ma negli esempi ci concentreremo sulle ontologie vedendo dei primi comandi per trattare ontologie dalla struttura semplice.

2.1 Parsing di documenti XML

Un documento XML non è altro che una struttura ad albero: ogni nodo rappresenta un concetto che può essere meglio definito nei figli del nodo stesso.

In CDuce possiamo ricostruire tale struttura definendo dei tipi. La forma generale di un elemento XML è `<(tag) (attr)> content` dove `tag`, `attr` e `content` sono espressioni su cui è possibile fare pattern matching. Si può quindi creare un tipo generale che faccia match con il tag root del documento XML e che, come `content`, abbia un array eterogeneo che conterrà tutti i figli del tag root. Ad ogni elemento del vettore sarà associato un tipo che avrà la stessa struttura del tipo generale e che permetterà di descrivere la struttura del documento XML discendendo fino alle foglie.

2.1.1 Esempio

Torniamo all'esempio dell'ontologia di persone definita nel listato 1.1 e descriviamo in CDuce la sua struttura.

Listato 2.1: persone.cd

```
1 namespace owl="http://www.w3.org/2002/07/owl#"
2 namespace rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3 namespace www="http://www.persone#"
4 namespace xml="http://www.w3.org/XML/1998/namespace"
5 namespace xsd="http://www.w3.org/2001/XMLSchema#"
6 namespace rdfs="http://www.w3.org/2000/01/rdf-schema#";
7
8 type Ontology = <rdf:RDF xml:base=String> [ Thing * ]
9 type Thing = Ont | Property | Class | Individual
```



```

10
11 type Ont = <owl:Ontology rdf:about=String> []
12
13 type Property = <owl:ObjectProperty rdf:about=String> []
14
15 type Class = <owl:Class rdf:about=String> [ ClassAttr * ]
16 type ClassAttr = EqClass | SubClass
17 type EqClass = <owl:equivalentClass> [ <owl:Restriction> [ AnyXml * ] ]
18 type SubClass = <rdfs:subClassOf rdf:resource=String> []
19
20 type Individual = <owl:NamedIndividual rdf:about=String> [ IndAttr * ]
21 type IndAttr = TypeInd | PropInd | Name
22 type Name = <rdfs:comment> String
23 type TypeInd = <rdf:type rdf:resource=String> []
24 type PropInd = <_ rdf:resource=String> [];;
25
26 let ontology :? Ontology = load_xml "persone.rdf";;

```

Si definiscono (linea 1 a 6) i Namespace in modo che CDuce li possa correttamente interpretare nel resto del documento. Si passa poi alla struttura vera e propria del file: tutte le informazioni sono contenute nell'elemento di tipo `Ontology` che è costituito da un tag e da una lista di elementi di tipo `Thing` che a sua volta può rappresentare un elemento di tipo `Ont`, `Property`, `Class` o `Individual`. Ognuno di questi elementi fa il match con un nodo figlio del nodo root del documento XML; a loro volta questi elementi hanno una struttura interna che può essere più o meno dettagliata a seconda di come ci interessa operare sul documento. È interessante notare che, dato che in questo caso le specifiche proprietà della restrizione di un elemento `EqClass` non ci interessano, abbiamo potuto fare il match di queste con un vettore di lunghezza arbitraria (anche nulla) di generici elementi XML. Se avessimo voluto specificare che il tag restrizione contiene esattamente 2 elementi senza specificare quali avremmo potuto scrivere `<owl>Restriction> [AnyXml AnyXml]`.

Quando andremo a caricare il documento con il comando alla riga 26, CDuce carica il file ed esegue il controllo di tipo verificando che la struttura del file XML sia effettivamente quella descritta dall'elemento `Ontology`; se il controllo va a buon fine otterremo l'elemento chiamato `ontology` di tipo `Ontology` contenente tutto il file XML.

2.2 Manipolare documenti XML

Una volta definita la struttura del documento possiamo definire delle funzioni che mappano un elemento in un altro elemento. Gli strumenti fondamentali sono:

- pattern matching come descritto in [1.2.3](#)
- **map**: permette di applicare una funzione a tutti gli elementi di una lista e restituisce una nuova lista di elementi trasformati della stessa lunghezza della lista iniziale;
- **transform**: permette di applicare una funzione a ogni elemento di una lista, restituendo per ogni elemento una lista di lunghezza arbitraria (anche nulla) e concatenando infine il risultato. Grazie alla funzione **transform** si può ottenere una lista di lunghezza diversa rispetto alla lista di partenza.

2.2.1 Esempio

Consideriamo nuovamente l'ontologia padri-figli con struttura definita in 2.1 e costruiamo 2 funzioni, la prima (chiamata **name**) che permetta di estrarre da un individuo tutti i nomi, la seconda (chiamata **names**) che usi la prima per costruire una lista con tutti i nomi contenuti nell'ontologia.

Listato 2.2: basic functions

```
1 (*first function*)
2 let fun name (Individual -> [String *])
3   <owl:NamedIndividual ..> [ (n::Name | _) * ] -> map n with
4   <rdfs:comment> s -> s;;
5 (*second function*)
6 let fun names (Ontology -> [String *])
7   <rdf:RDF ..> x -> transform x with
8   | (x & Individual) -> name x
9   | _ -> [];
```

Nella prima funzione, usando il pattern matching, lego la variabile **n** alla sequenza di tutti gli elementi di tipo **Name** associati a quell'individuo, poi uso **map** per estrarre da ogni elemento di questa lista solo la stringa col nome.

Nella seconda funzione uso **transform** per selezionare dalla lista **x** (lista di elementi tipo **Thing**) solo gli elementi **Individual**, a questi applico la funzione **name** sopra definita. La seconda clausola di **transform**, che serve a buttare via tutti gli elementi di cui non si è ancora fatto il match è implicita e può essere quindi omessa.

La prima funzione ci ha permesso di passare da una lista **n** di elementi **Name** ad una lista della stessa lunghezza di stringhe. La seconda ci ha permesso di passare da una lista di elementi **Thing** a una lista di elementi **Individual** di differente lunghezza.

2.3 Query

Un punto di forza di CDuce sono le Query. Le useremo profusamente nei capitoli 3 e 4 per creare liste e selezionare elementi in modo rapido e leggibile al posto della funzione **transform**. La forma generale di una Query è molto simile alla stessa espressa in SQL

```
select e
from p1 in e1,
     p2 in e2,
     :
     pn in en
where c
```

Dove **e** ed **ei** sono espressioni, **pi** sono pattern e **c** è un'espressione booleana. Il risultato finale è la lista di tutti i valori ottenuti valutando **e** nella sequenza di possibili combinazioni in cui le variabili libere di **e** sono legate facendo il match dei pattern **pi** con le espressioni **ei**, a condizione che **c** sia rispettata.

Le Query possono essere simulate con **transform** che permette di creare, grazie a pattern matching successivi l'espressione **e** e di selezionarla solo nel caso in cui l'espressione booleana **b** sia rispettata. Si può riscrivere la generica Query usando la **transform** come segue:

```

transform e1 with p1 ->
  transform e2 with p2 ->
    ...
    transform en with pn ->
      if b then [e] else []

```

La valutazione dei due comandi produce il medesimo risultato.

Anche se **transform** può sembrare più versatile o espressiva delle Query, quando possibile, è sempre vantaggioso usare queste ultime. I vantaggi sono molteplici:

- la struttura delle Query è più elegante e leggibile, permette di capire con più facilità cosa si sta cercando;
- le Query sono automaticamente ottimizzate con le stesse tecniche di ottimizzazione di SQL, questo è molto vantaggioso soprattutto in ontologie o tesauroi particolarmente popolati;
- CDuce fornisce dei controlli a priori sul risultato della Query permettendo di evidenziare errori che porterebbero ad una Query che viene interpretata correttamente ma produce sempre risultati vuoti.

Oltre che con la struttura del **select from** le Query possono essere espresse anche come proiezioni, come mostrato nel listato 2.3

2.3.1 Esempio

Facendo sempre riferimento all'ontologia con struttura definita nel listato 2.1 scriviamo delle Query per ottenere lo stesso risultato della seconda funzione definita nel listato 2.2

Listato 2.3: basic Query

```

1  (*Query*)
2  let sel = select y
3      from x in [ontology]/Individual,
4           y in [x]/Name
5  in
6  map sel with <rdfs:comment> n -> n;;
7
8  (*projection 1*)
9  map [ontology]/Individual/Name with <rdfs:comment> n -> n;;
10
11 (*projection 2*)
12 map [ontology]/Individual/<rdfs:comment> _ with <rdfs:comment> n -> n;;
13
14 (*error 1*)
15 [ontology]/Name;;
16
17 (*error 2*)
18 let sel = select x
19     from x in [ontology]/Individual
20 in [sel]/Name
21

```

```

22  (*correction error 2*)
23  let sel = select x
24      from x in [ontology]/Individual
25  in sel/Name

```

La query crea una lista di elementi di tipo `Name` a cui si assegna il nome `sel`, questa lista viene poi trasformata in una lista di stringhe grazie alla `map`. La proiezione fa esattamente la stessa cosa in modo ancora più conciso. Sia nella forma del `select from` sia nella forma delle proiezioni si possono usare anche espressioni per fare il match oltre che tipi, come si vede nella `projection 2`.

Gli ultimi due esempi mostrano come `CDuce` possa aiutare nella rilevazione degli errori: se eseguiti restituiscono rispettivamente gli avvertimenti: `Warning: This projection always returns the empty sequence` e `Warning: This branch is not used`, informandoci che il risultato sarà sempre vuoto.

Analizziamo il primo errore. `CDuce` sa che, affinché il risultato possa contenere dei valori, gli elementi della lista chiamata `ontology` (che sono di tipo `Thing`) devono essere sottotipi di `[<_ ..>[Any* Name Any*] *]`; dato che non è così, il risultato sarà sempre vuoto.

Il secondo errore è dovuto al fatto che mettendo le parentesi quadre intorno a `sel` questo viene interpretato come una sequenza (anche se lo è già), ne risulta una sequenza il cui unico elemento è una sequenza (in particolare il tipo attribuito da `CDuce` è `[[Individual*]]`) anche in questo caso controllando il tipo `CDuce` è in grado di capire che il risultato sarà sempre vuoto. Perché la Query restituisca dei valori il tipo deve essere `[Individual*]`, cosa che avviene nella Query successiva che in effetti restituisce il risultato atteso.

Conclusioni

Abbiamo visto come si può descrivere la struttura di un documento XML in `CDuce`, come si può trasformare un elemento di un tipo in elemento di un altro tipo e come creare liste selezionando quali elementi aggiungere. Nei capitoli 3 e 4 applicheremo gli strumenti analizzati ad esempi notevoli che mostrino in che modo e in che contesti `CDuce` possa essere usato per operare, modificandoli, tesauri e ontologie

Capitolo 3

Trasformare un thesauro in un'ontologia

Introduzione

Qui vediamo come sia possibile usare CDuce per creare ontologia a partire da una rappresentazione della conoscenza già formalizzata in altri modi. Per illustrare il processo di creazione, faremo riferimento ad un esempio facilmente replicabile: la trasformazione di un thesauro in un'ontologia (per una trattazione più astratta e formale sulla re-ingegnerizzazione di un thesauro si veda [11]). Commenteremo le scelte effettuate e le strategie adottate. L'esempio permette una discussione lineare e senza eccessivi giri di parole mantenendo comunque una buona generalità qualora si applicassero le stesse scelte e strategie a diversi contesti.

Dopo aver trasformato il thesauro in un'ontologia con CDuce, illustriamo brevemente quali altri strumenti avremmo potuto utilizzare, in particolare cercheremo di ottenere risultati analoghi con Protégé discutendo vantaggi e svantaggi di ciascun approccio.

3.1 Vantaggi del passaggio da thesauro ad ontologia

Potremmo essere interessati a trasformare un thesauro espresso in SKOS in un'ontologia espressa in OWL per varie ragioni: prima fra tutte un'ontologia è una rappresentazione più formale della conoscenza e OWL ha un potere espressivo maggiore di SKOS; potremo quindi effettuare Query più avanzate e utilizzare strumenti di inferenza più potenti.

Per presentare questo argomento prendiamo in considerazione il thesauro: "Europeana Fashion Thesaurus"¹ e tentiamo di trasformarlo in un ontologia. Prima di addentrarci nella parte tecnica forniamo una breve introduzione al thesauro che editeremo.

3.1.1 Europeana Fashion Thesaurus: capturing imagination

L' "Europeana Fashion project"² è un progetto che si è posto come obiettivo quello di organizzare sotto una struttura gerarchica tutto ciò che riguardasse la moda, comprendendo anche sinonimi e contrari. Il thesauro è pubblico può essere scaricato liberamente da chiunque sia interessato. Il thesauro permette di accedere in modo logico, organizzato e strutturato

¹<http://thesaurus.europeanafashion.eu/>

²<http://www.europeanafashion.eu/>

ad una vasta conoscenza. Data la struttura ad albero le relazioni tra oggetti sono chiare e l'aggiunta o la ricerca di informazioni può essere fatta in modo rapido ed efficiente.

3.1.2 Svantaggi del thesauro

Per la sua struttura, il thesauro permette una rappresentazione della conoscenza puramente tassonomica; questo in certi contesti può essere una limitazione, ad esempio nonostante nel thesauro siano presenti sia capi d'abbigliamento che materiali, non c'è nessuna possibilità di mettere in relazione i due concetti in modo formale; inoltre se avessimo degli strumenti per fare inferenza, non saremmo in grado di dedurre che una bandana è un accessorio per il capo, perché quest'ultima non si trova nel ramo degli accessori della testa.

Il passaggio ad un'ontologia permetterebbe di costruire relazioni più complesse per descrivere in modo più ricco gli oggetti del dominio di interesse.

3.2 Struttura del thesauro

La struttura del thesauro che vogliamo re-ingegnerizzare è semplice: in particolare abbiamo quattro concetti fondamentali da cui partono tutti gli altri rami che rappresentano i concetti derivati. I concetti fondamentali sono:

- oggetti di moda;
- eventi di moda;
- tecniche;
- materiali.

Ogni concetto, che sia uno o meno di quelli fondamentali, è rappresentato con un tag **Description**, contenente, a sua volta, una lista di attributi tra cui le **label** che contengono il nome (nelle varie lingue) del concetto, il tag **broader** che nei concetti derivati permette di risalire al nodo padre, le **scopeNote** che contengono una descrizione del concetto, e il tag **exactMatch** che rimanda “The Getty Vocabularies”³.

Analizzata la struttura del thesauro possiamo descriverla formalmente con CDuce in modo da fare il parsing del documento per poter definire funzioni che permettano di trasformare i concetti descritti dal tag **Description** in classi nel linguaggio OWL. Una possibile descrizione del thesauro in CDuce è la seguente:

Listato 3.1: thesaurus.europeana.cd

```
1 type Thesaurus = <rdf:RDF>[Desc *]
2 type Desc = <rdf:Description rdf:about=String>[DescAtt *]
3
4 type DescAtt = AltLabel | InScheme | PrefLabel | ScopeNote | ExactMatch |
   ↪ Broader | Type
5 type AltLabel = <skos:altLabel xml:lang=String> String
6 type InScheme = <skos:inScheme rdf:resource=String> []
7 type PrefLabel = <skos:prefLabel xml:lang=String> String
8 type ScopeNote = <skos:scopeNote xml:lang=String> String
9 type ExactMatch = <skos:exactMatch rdf:resource=String> []
```

³<http://vocab.getty.edu/>

```

10 type Broader = <skos:broader rdf:resource=String> []
11 type Type = <rdf:type rdf:resource=String> [];;

```

3.3 Struttura dell'ontologia

L'ontologia che vogliamo creare a partire da questo tesoro, almeno inizialmente, non potrà contenere più informazioni di quelle già presenti, ci limitiamo quindi a costruire una tassonomia, che potrà poi essere arricchita passando da un albero ad un grafo, con relazioni più ricche tra le classi.

Una volta creata l'ossatura dell'ontologia, potremo andare a definire delle relazioni sugli individui ad esempio per specificare che un dato capo d'abbigliamento è prodotto con un determinato tessuto.

La struttura dell'ontologia, almeno per quello che ci serve per riportare tutte le informazioni contenute nel tesoro, può essere formalizzata in CDuce nel seguente modo:

Listato 3.2: ontology_europeana.cd

```

1 type Ontology = <rdf:RDF xml:base=String> [ Thing * ]
2 type Thing = Ont | Class
3
4 type Ont = <owl:Ontology rdf:about=String> []
5
6 type Class = <owl:Class rdf:about=String> [ ClassAtt * ]
7 type ClassAtt = SubClass | Label | Note | Dictionary
8 type SubClass = <rdfs:subClassOf rdf:resource=String> []
9 type Label = <rdfs:label xml:lang=String> String
10 type Note = <skos:scopeNote xml:lang=String> String
11 type Dictionary = <skos:exactMatch rdf:resource=String> []

```

Notiamo subito come la struttura dell'ontologia si sia semplificata rispetto alla struttura del listato 2.1, questo perché adesso ci interessa solo ricostruire i concetti SKOS con classi OWL.

È interessante come si possano integrare tag SKOS direttamente nel linguaggio OWL: in questo caso li usiamo per aggiungere informazioni umanamente leggibili (le note) e per mantenere il riferimento al dizionario (per una discussione dettagliata sull'interazione tra OWL e SKOS si veda [2]).

Si potrebbero anche eliminare completamente i tag SKOS mappandoli adeguatamente in altri tag (ad esempio trasformando `scopeNote` in `comment`), ma questo non porterebbe nessun vantaggio dal punto di vista della formalità dell'ontologia e renderebbe più complesse le funzioni di trasformazione.

3.4 Da concetto SKOS a classe OWL

Il nostro obiettivo è trasformare tutti i concetti del tesoro in classi di un'ontologia, mantenendo la gerarchia esprimendola come relazione di sottoclasse. Per raggiungere questo scopo definiamo una funzione per mappare gli elementi di tipo `DescAttr` in elementi di tipo `ClassAttr`. Successivamente usiamo queste funzioni per definirne una che ci permetta di

passare da un intero concetto del tesauro ad una classe dell'ontologia; infine, usando la funzione `map` (2.2), possiamo applicare questa funzione a tutti i concetti del tesauro per ottenere le classi che popoleranno l'ontologia.

3.4.1 Trasformare gli attributi

Iniziamo definendo le funzioni per mappare i tag del tesauro in tag dell'ontologia, questo è il primo esempio in cui si possono apprezzare i vantaggi dell'uso di un linguaggio funzionale. Esprimeremo le trasformazioni in modo semplice ed elegante senza perdere in leggibilità, inoltre avremo la garanzia che la trasformazione restituisca esattamente il tipo dichiarato nell'interfaccia della funzione, infine il controllo di tipo e l'inferenza del tipo di un'espressione ci aiuteranno a trovare e correggere eventuali errori. Definiamo una funzione per ciascun tag che desideriamo esportare, in particolare tralasciamo il tag `type` che nel tesauro assume solo 2 valori (`Concept` e `ConceptScheme`). Le funzioni possono essere scritte in `CDuce` come segue:

Listato 3.3: SKOS_to_OWL.cd

```

1  let fun transformNote (ScopeNote -> Note)
2    x -> x
3
4  let fun transformDictionary (ExactMatch -> Dictionary)
5    x -> x
6
7  let fun transformLabel (PrefLabel -> Label ; AltLabel -> Label)
8    | <skos:prefLabel xml:lang=l> lab -> <rdfs:label xml:lang=l> lab
9    | <skos:altLabel xml:lang=l> lab -> <rdfs:label xml:lang=l> lab;;
10
11 let fun transformSubClass (Broader -> SubClass)
12   <skos:broader rdf:resource=res> [] -> <rdfs:subClassOf rdf:resource=res>
    ↪ [];;

```

Le prime due funzioni sono banali: confrontando i listati 3.1 e 3.2 notiamo che gli elementi di tipo `ScopeNote` e `Note` hanno la stessa struttura, così come gli elementi di tipo `ExactMatch` e `Dictionary`. Per questi tag è sufficiente la funzione identità.

Per quanto riguarda le `label`, nel tesauro ci sono due tipi di elementi, nell'ontologia abbiamo deciso di usarne solo uno, vediamo quindi un esempio di overloading della funzione. Per quanto riguarda la trasformazione vera e propria, usiamo il pattern matching per legare la variabile `l` e `lab` rispettivamente alla lingua e al testo della `label`; una volta legate queste variabili verranno usate per costruire il nuovo elemento di tipo `Label`.

La funzione più importante è certamente quella che lavora sull'elemento tipo `Broader`, questa si occupa infatti di trasformarlo in un elemento di tipo `SubClass` in modo da mantenere le relazioni gerarchiche del tesauro. Dal punto di vista della trasformazione però la funzione è molto semplice: leghiamo un'unica variabile `res` alla stringa che identifica il nodo padre e con questa costruiamo il tag `subClassOf`.

3.4.2 Trasformare una singola classe

Per trasformare un singolo concetto espresso in SKOS in una classe OWL definiamo due funzioni: la prima trasforma un attributo del concetto (`DescAttr`) in un vettore di attributi di una classe (`[ClassAttr *]`) attraverso le funzioni definite prima (3.3). La ragione

per cui il risultato deve essere un vettore, è che se l'attributo ci interessa, restituiamo un vettore con un elemento, se l'attributo va scartato, restituiamo un vettore vuoto. La seconda funzione costruisce l'involucro esterno della classe e usa la prima per trasformare tutti gli attributi del concetto in attributi della classe. L'implementazione di queste funzioni potrebbe essere la seguente:

Listato 3.4: concept_to_class.cd

```

1  let fun transformAtt (DescAtt -> [ClassAtt*])
2    | x & PrefLabel   -> [(transformLabel x)]
3    | x & AltLabel    -> [(transformLabel x)]
4    | x & ScopeNote   -> [(transformNote x)]
5    | x & ExactMatch  -> [(transformDictionary x)]
6    | x & Broader     -> [(transformSubClass x)]
7    | Any             -> [];
8
9  let fun transformClass (Desc -> Class)
10    <rdf:Description rdf:about=ab> [ (descAtt :: DescAtt)* ] ->
11      let classAtt = flatten ( map descAtt with x -> transformAtt x) in
12      <owl:Class rdf:about=ab> classAtt;;

```

Le funzioni sono abbastanza semplici. Descriviamo brevemente la seconda: usiamo il pattern matching per legare la variabile `ab` alla stringa che identifica il concetto (useremo questo identificatore anche per la classe); leghiamo poi la variabile `descAtt` al vettore di attributi del concetto SKOS, usando poi la `map` per trasformare questo vettore in attributi di una classe OWL. Siccome la funzione `transformAtt` prende un elemento e restituisce un vettore, al termine della `map` avremo un vettore di vettori, per appiattirne la struttura usiamo la funzione già definita in CDuce `flatten`. A questo punto abbiamo tutti gli elementi per definire la nuova classe che costruiamo assemblando il tag `Class` con la stringa identificativa e il vettore di attributi.

3.5 Costruire la nuova ontologia

Ora che abbiamo una funzione per trasformare ogni concetto SKOS in una classe OWL, possiamo applicarla a tutti i concetti del thesaurus per costruire un'ontologia. Possiamo completare la trasformazione in CDuce in questo modo:

Listato 3.5: thesaurus_to_ontology.cd

```

1  let fashion :? Thesaurus = load_xml "thesaurus.rdf";;
2
3  let newClass = map [fashion]/Desc with x -> transformClass x
4  in
5  let newOnt : Ontology = <rdf:RDF
6    ↪ xml:base="http://www.semanticweb.org/OntEur"> ( [ <owl:Ontology
7    ↪ rdf:about="OntEur"> [ ] ] @ newClass )
8  in
9  dump_to_file_utf8 "Ontologia.rdf" (print_xml_utf8 newOnt);;

```

Alla linea 1, facendo riferimento alla struttura del thesaurus definita in 3.1, carichiamo il thesaurus; alla riga 2 usiamo una proiezione per estrarre un vettore con tutti gli elementi di tipo `Desc` (che sono i concetti del thesaurus), usiamo la `map` per applicare ad ognuno di questi elementi la funzione `transformClass` (3.4), infine diamo un nome al vettore di classi appena creato in modo da poterlo usare nel seguito della trasformazione.

Alla linea 5 creiamo l'ontologia vera e propria aggiungendo tutte le classi appena create, infine facciamo il dump su file generando un documento XML che potrà essere visualizzato con qualsiasi altro strumento incluso Protégè.

3.6 Versione compatta

Per trasformare il thesaurus in un'ontologia, abbiamo definito varie funzioni che ci hanno permesso di trasformare pezzo per pezzo i tag SKOS nei rispettivi OWL. Assemblando progressivamente i pezzi, abbiamo costruito l'ontologia. Proviamo ora a sfruttare tutti i costrutti forniti da CDuce per riscrivere la trasformazione in una forma più compatta

Listato 3.6: thesaurus.to_ontology_compact.cd

```

1 let fun thesaurusToOntology (Thesaurus -> Ontology)
2   <rdf:RDF>[ (concepts :: Desc) * ] ->
3   let newClasses = map concepts with <rdf:Description rdf:about=ab> [
4     ↪ (descAttr :: DescAtt)* ] ->
5     let classAttr = transform descAttr with
6       | x & ScopeNote -> [x]
7       | x & ExactMatch -> [x]
8       | <skos:prefLabel xml:lang=1> lab -> [<rdfs:label xml:lang=1> lab]
9       | <skos:altLabel xml:lang=1> lab -> [<rdfs:label xml:lang=1> lab]
10      | <skos:broader rdf:resource=res> [] -> [<rdfs:subClassOf
11        ↪ rdf:resource=res> []]
12    in
13    <owl:Class rdf:about=ab> classAttr
14  in
15  <rdf:RDF xml:base="http://www.semanticweb.org/OntEur"> ( [ <owl:Ontology
16    ↪ rdf:about="OntEur"> [ ] ] @ newClasses );;
```

Questa funzione permette di passare direttamente da un elemento di tipo `Thesaurus` ad un elemento di tipo `Ontology`. Rispetto alla definizione più estesa abbiamo perso il controllo puntuale sul tipo degli attributi di tipo `ClassAttr`, infatti in questo caso CDuce ci assicura solo che siamo passati correttamente da attributi di un concetto ad attributi di una classe, senza controlli più specifici.

Nonostante questo minore controllo sui tipi questa nuova definizione è molto più compatta rimanendo comunque leggibile: iniziamo con una `map` per trasformare il tag `Description` in un tag `Class`; dentro la `map` innestiamo una `transform` per trasformare gli attributi da SKOS a OWL per ogni elemento estratto con la funzione `map`.

3.7 Aumentare l'espressività

Ora che abbiamo ottenuto un'ontologia, possiamo aggiungere relazioni tra gli elementi, ad esempio possiamo mettere in relazione gli oggetti di moda con i materiali o i colori

e le tecniche con i materiali. Possiamo inoltre definire delle relazioni di sottoclasse più complesse di quella puramente tassonomica. In seguito (4.1.1) presenteremo un'estensione della struttura dell'ontologia definita qui 3.2 in cui la andremo ad arricchire con nuove informazioni ed useremo la nuova versione per sviluppare selezioni e trasformazioni più sofisticate di quelle che avremo potuto definire su un tesauro.

3.8 In Protégé

Provando ad editare il tesauro in Protégé, vediamo che vengono riconosciute come classi quella dei **Concepts** e quella dei **ConceptScheme**, tutti gli elementi del tesauro sono individui. In particolare i 4 concetti fondamentali appartengono alla classe **ConceptScheme**, tutti gli altri alla classe **Concept**. Tra gli strumenti base di refactor di Protégé non ve ne sono per trasformare degli individui in classi, tanto meno mantenendo la struttura gerarchica specificata nel tesauro. Senza voler operare in modo puntuale su ogni individuo per trasformarlo manualmente in una classe (e senza usare i plug-in che permettono di scrivere programmi, ad esempio in java⁴, per operare delle trasformazioni simili a quelle fatte con CDuce), ciò che si può fare è comunque definire le stesse relazioni che abbiamo accennato nel paragrafo precedente (3.7), che avranno in questo caso dominio e range coincidenti (individui di classe **Concepts**). In questo modo possiamo comunque aumentare l'espressività del tesauro, ma con due grandi vincoli:

- non avremo il supporto dei reasoner per verificare la correttezza delle relazioni che valorizzeremo (ad esempio siccome dominio e range della relazione “è fatto di” sono la stessa classe potremo liberamente creare la relazione: “velluto è fatto di alluminio”);
- se volessimo definire delle relazioni di sottogenere più complesse, dovremo farlo tramite la definizione di relazioni, anziché con la più naturale definizione di sottoclasse.

3.9 Conclusioni

Nel contesto della re-ingegnerizzazione di un tesauro CDuce si dimostra molto più versatile di Protégé, questo è dovuto principalmente a due fattori:

- come si legge in [11] una parte importante del processo di trasformazione è la conversione sintattica dei tag, in questo CDuce è uno strumento molto efficace;
- Protégé è uno strumento per editare ontologie espresse in OWL [12] e, nonostante sia molto potente e versatile, non è lo strumento più indicato per operare su un tesauro.

In questo frangente l'uso di un linguaggio di programmazione che ci permettesse di definire esattamente come manipolare i dati è stato essenziale per ottenere un buon risultato di traduzione. Una volta acquisite delle competenze di base abbastanza solide con CDuce e avendo familiarità con gli strumenti che mette a disposizione, la scrittura di un programma che esegue la traduzione da tesauro ad ontologia non risulta particolarmente complessa o lunga in termini di tempo (come testimonia la funzione 3.6).

Dal punto di vista funzionale, abbiamo sfruttato la possibilità di definire funzioni avendo un sofisticato controllo sul tipo delle espressioni che stavamo trattando; questo ha permesso di evitare subito degli errori che, con linguaggi imperativi, non sarebbero emersi fino al momento dell'esecuzione; in particolare dovendo lavorare alternativamente su vettori ed elementi singoli è capitato spesso che il tipo inferito fosse un vettore, mentre il tipo richiesto fosse un elemento o viceversa (in un linguaggio come C con l'uso dei puntatori un errore del genere avrebbe potenzialmente richiesto molto tempo per essere individuato e risolto).

⁴<https://www.java.com/>

Capitolo 4

Merge di ontologie

introduzione

In questo capitolo vediamo come sia possibile usare CDuce per fare il merge di ontologie, cercando di mettere in risalto quelli che possono essere i vantaggi di usare un linguaggio funzionale rispetto ad uno strumento grafico. Alla fine del capitolo paragoneremo l'approccio con CDuce a quello con Protégé. Per mettere in evidenza i vantaggi di CDuce presentiamo un esempio un po' più complesso di quelli visti in precedenza in modo da far risaltare le potenzialità di CDuce e contemporaneamente dare un'idea di funzioni più complesse.

4.1 Obiettivo del merge

Adesso che abbiamo creato un'ontologia a partire dall'“European Fashion Thesaurus” e abbiamo descritto una semplice ontologia per descrivere persone, vorremmo fonderle in una nuova ontologia che possa parlare degli usi e costumi della società, magari riferita ad un ben preciso periodo storico. Vogliamo creare un'ontologia in cui sia possibile rappresentare delle persone, coi rispettivi legami di parentela e nella quale sia possibile associare alle persone uno status sociale, un lavoro e uno specifico abbigliamento.

4.1.1 Ontologie di partenza

Per creare questa ontologia consideriamo 3 ontologie di partenza:

- **society**: è lo scheletro dell'ontologia che vogliamo ottenere: è comodo crearla a priori in modo da non dover creare ex novo tutta la struttura in CDuce, inizialmente rappresenta delle persone sulle quali definiamo le relazioni **born_in**, **is** e **work_as** che associano una persona con la città di nascita, con il suo status sociale e con il proprio lavoro;
- **people**: rappresenta gli individui con le loro parentele: rispetto all'ontologia presentata in 1.1 abbiamo aggiunto la relazione simmetrica **marry**, e la relazione **son_of** come inversa di **parent_of**, aggiungiamo anche l'anno di nascita;
- **fashion**: è l'ontologia che abbiamo creato nel capitolo 3 arricchita con le relazioni **made_of**, **crafted_with**, **color**, che indicano rispettivamente i materiali costituenti, le tecniche realizzative e i colori di un capo d'abbigliamento.

4.1.2 Ontologia di arrivo

Vogliamo modificare **society** in modo da descrivere gli usi e i costumi della società del XII secolo, per fare questo estrapoliamo dall'ontologia **fashion** tutti i vestiti che non siano costituiti da fibre artificiali e scremiamo dall'ontologia **people** tutte le persone che sono vissute in un'epoca che non ci interessa. Vogliamo poi che la classe **People** definita in **society** sia equivalente alla classe **People** definita in **people**, senza perdere tutte le relazioni di parentela, e potendo contemporaneamente sfruttare le nuove relazioni definite in **society**.

4.2 Struttura generica di un'ontologia

Per fare il merge dobbiamo poter caricare le 3 ontologie in CDuce, quindi descriviamo una struttura generale per fare il parsing di un'ontologia generica:

Listato 4.1: general structure

```
1 type Ontology = <rdf:RDF xml:base=String> [ Thing * ]
2 type Thing = Ont | AnnProperty | ObjProperty | DataProperty | Class |
   ↳ Individual
3
4 type Ont = <owl:Ontology rdf:about=String> []
5
6 type AnnProperty = <owl:AnnotationProperty rdf:about=String> []
7
8 type DataProperty = <owl:DatatypeProperty rdf:about=String> [ PropAttr* ]
9 type ObjProperty = <owl:ObjectProperty rdf:about=String> [ PropAttr * ]
10
11 type PropAttr = Inverse | Domain | Range | PropType
12 type Inverse = <owl:inverseOf rdf:resource=String> []
13 type Domain = <rdfs:domain rdf:resource=String> []
14 type Range = <rdfs:range rdf:resource=String> []
15 type PropType = <rdf:type rdf:resource=String> []
16
17 type Class = <owl:Class rdf:about=String> [ ClassAtt * ]
18 type ClassAtt = SubClass | EqClass | Label | Note | Dictionary
19 type SubClass = <rdfs:subClassOf rdf:resource=String> []
20 type EqClass = <owl:equivalentClass> [ EqAttr ] | <owl:equivalentClass
   ↳ rdf:resource=String> []
21 type EqAttr = <owl:Restriction> [ AnyXml* ]
22 type Label = <rdfs:label xml:lang=String> String
23 type Note = <skos:scopeNote xml:lang=String> String
24 type Dictionary = <skos:exactMatch rdf:resource=String> []
25
26 type Individual = <owl:NamedIndividual rdf:about=String> [ IndAttr * ]
27 type IndAttr = IndClass | IndProp | DataProp
28 type IndClass = <rdf:type rdf:resource=String> []
29 type IndProp = <_ rdf:resource=String> []
30 type DataProp = <_ rdf:datatype=String> String
```

Rispetto alle strutture definite in precedenza notiamo varie aggiunte:

- si specifica meglio l'elemento `EqClass`: questo infatti può essere una restrizione (un genitore è una persona con dei figli) oppure un'equivalenza senza condizioni (ci serve per rendere uguali i concetti di persona definiti nelle 2 ontologie);
- definiamo il tipo `DataProperty`: questo rappresenta una proprietà degli individui, nel nostro caso la usiamo per specificare l'anno di nascita di una persona. Gli individui nella loro lista di attributi potranno ora averne uno di tipo `DataProp`;
- il tipo `AnnProperty` serve perché avendo editato l'ontologia `fashion` in Protégé (per aggiungere le relazioni non presenti nel tesauro) i tag SKOS sono stati correttamente riconosciuti [2] e classificati come tag `owl:AnnotationProperty`;

4.3 Merge

4.3.1 Funzioni utili

Prima di passare alla costruzione della nuova ontologia vediamo alcune funzioni utili che si possono applicare a qualsiasi ontologia e che sono servite per costruire le funzioni specifiche adatte a manipolare la particolare ontologia di interesse.

Listato 4.2: usefull function

```
1 let fun loadOntology (Latin1 -> Ontology)
2   x -> load_xml x :? Ontology;;
3
4 let fun head ([ Any* ] -> Any)
5   | ([ x ] @ _) -> x
6   | [] -> "error empty list";;
7
8 let fun subClasses (Class -> Ontology -> [ Class* ])
9   <owl:Class rdf:about=ab> [ _* ] -> fun (Ontology -> [ Class* ])
10  ont ->
11    select x from
12      x in [ont]/Class,
13      y in [x]/SubClass
14    where (y = <rdfs:subClassOf rdf:resource=ab> []);;
15
16 let fun subClassesRec ( Class -> Ontology -> [ Class* ])
17   cl -> fun (Ontology -> [ Class* ])
18   ont ->
19     let subCl = subClasses cl ont in
20     subCl @ flatten (map subCl with y -> subClassesRec y ont)
21
22 let fun andList ( [ Bool* ] -> Bool )
23   | ([ x ] @ y) -> (x && (andList y))
24   | [] -> `true;;
25
26 let fun orList ( [ Bool* ] -> Bool )
27   | ([ x ] @ y) -> (x || (orList y))
```

```

28 | [] -> `false;;
29
30 let fun classOf (Individual -> Ontology -> [Class*])
31   <owl:NamedIndividual ..> [ (tp::IndClass | _) *] -> fun (Ontology ->
32     ↳ [Class*])
33   ont ->
34     transform tp with <rdf:type rdf:resource=str> [] ->
35     transform [ont]/Class with x ->
36       match x with <owl:Class rdf:about=a>[ _* ] -> if a = str then [x]
37       ↳ else [];;
38
39 let fun contains (Any -> [Any*] -> Bool)
40   obj -> fun ([Any*] -> Bool)
41   lst ->
42     let intersect = select x from x in lst
43     where (x = obj) in
44     match intersect with
45     | [] -> `false
46     | [ Any* ] -> `true;;
47
48 let fun isInClasses (Individual -> [Class*] -> Ontology -> Bool)
49   ind -> fun ([Class*] -> Ontology -> Bool)
50   classes -> fun (Ontology -> Bool)
51   ont ->
52     let c1 = classOf ind ont in
53     let res = map c1 with x -> contains x classes in
54     andList res;;

```

Per la prima volta abbiamo fatto ricorso ad applicazioni parziali, questo è utile per poter parametrizzare anche l'ontologia di riferimento per l'operazione; dato che in questo caso ne manipoliamo contemporaneamente tre, è importante poter specificare volta per volta a quale ci riferiamo.

Vediamo per la prima volta degli esempi di funzioni ricorsive:

- **andList** e **orList** lavorano su liste di booleani e restituiscono il risultato della congiunzione o disgiunzione logica tra tutti gli elementi di una lista. Per ottenere questo risultato si fa il pattern matching della lista che può essere un elemento in testa alla lista seguito da una lista che chiamiamo coda; la funzione restituisce l'operazione logica tra la testa e la chiamata ricorsiva alla funzione, passando come parametro la coda della lista. Quando la lista è vuota (seconda clausola del matching) la funzione restituisce l'elemento neutro dell'operazione logica;
- **subClassRec** richiama ricorsivamente se stessa per costruire l'intero albero di sottoclassi a partire da una classe data (e dall'ontologia di riferimento). Ovviamente perché questa funzione possa terminare, la struttura delle classi deve essere un grafo aciclico (questa non è una grossa limitazione, infatti se la classe *c2* è contemporaneamente sopraclasse e sottoclasse di *c1*, allora *c1* e *c2* sono equivalenti e possono essere accorpate per eliminare i cicli, discorso analogo vale per cicli più lunghi).

Per verificare se un individuo si trova in un certo albero di classi, si potrebbe operare al contrario rispetto a **subClassRec**, risalendo l'albero fino a quando non si trova il padre

desiderato (restituendo `true`) oppure la radice delle classi (restituendo `false`). Questo approccio, però, presenta delle problematiche:

- un individuo può appartenere a più classi, quindi bisognerebbe risalire n alberi dove n è il numero di classi a cui appartiene l'oggetto;
- ogni classe può essere sottoclasse di più classi diramando così ulteriormente la ricerca.

Facendo alcuni test, si nota che un approccio di questo genere, oltre ad essere impegnativo dal punto di vista implementativo è anche poco efficiente dal punto di vista prestazionale. Come vedremo in 4.3 per ogni capo d'abbigliamento dovremo chiederci se è costituito da materiali artificiali e questo rallenta il processo di merge. Per evitare questo problema costruiamo solo una volta la lista di materiali artificiali e, quando dobbiamo stabilire se un materiale è naturale o meno, verifichiamo se appartiene alla lista dei materiali artificiali. Questo approccio è vantaggioso in quanto la lista di materiali artificiali andrebbe creata in ogni caso per andare ad aggiungere alla nuova ontologia tutti i materiali che non lo sono (conviene creare la lista dei materiali artificiali piuttosto che quella dei materiali naturali perché la prima contiene molti meno elementi, di conseguenza è più veloce da creare).

Per implementare questa ricerca usiamo `isInClasses` che prende un individuo e una lista di classi, valuta a che classi appartiene l'individuo (con `classOf`) e usa la funzione `contains` per creare una lista di valori booleani uno per ogni classe dell'individuo, verificando se è contenuto nella lista di classi fornita; infine si verifica se la lista di booleani contiene solo `true` con la funzione `andList`

4.3.2 Selezione degli abiti

Mostriamo come si possano importare solo le fibre naturali e i vestiti costituiti esclusivamente da questi materiali, questo ha lo scopo di presentare come si possa fare della semplice inferenza sulla base di conoscenza usando solamente CDuce e senza ricorrere a reasoner esterni più sofisticati e complessi. Non mostriamo le funzioni per selezionare solo le persone vissute nel periodo storico considerato, questo per non inserire troppo codice e perché sarebbe poco istruttivo (non introdurremmo nuovi concetti rispetto a quelli che presentiamo in seguito).

Per ragionare sul singolo abito ci domandiamo di che materiali è fatto usando la relazione `made_of` che associa uno o più materiali ad un abito. Una volta che abbiamo i materiali, possiamo considerare la struttura ad albero dei materiali dell'ontologia `fashion` (la struttura è quella importata dal tesoro "European Fashion Thesaurus") per distinguere materiali naturali da artificiali (le due categorie devono essere disgiunte). Un vestito verrà considerato valido per la nuova ontologia solo se costituito unicamente da fibre naturali¹.

Una possibile implementazione delle funzioni che ci permettono di stabilire se un abito è costituito esclusivamente da fibre naturali è la seguente:

Listato 4.3: test artificial

```
1 let fun madeOf (Individual -> Ontology -> [Individual*])
2   <owl:NamedIndividual ..> [ (ip::IndProp | _) *] -> fun (Ontology ->
3     ↳ [Individual*])
   ont ->
```

¹possiamo fare inferenza solo coi dati in nostro possesso, in particolare la classe dei materiali si divide in tre sottoclassi che rappresentano materiali naturali, artificiali e materiali di decorazione; su questi ultimi non possiamo fare alcun ragionamento sull'origine, si è operata una scelta permissiva considerandoli tutti naturali


```

4     transform ip with <ns1:made_of rdf:resource=str> [] ->
5     transform [ont]/Individual with x ->
6     match x with <owl:NamedIndividual rdf:about=a>[ _* ] -> if a = str
    ↪ then [x] else [];;
7
8
9 let fun isArtificial (Individual -> [Class*] -> [Class*] -> Ontology ->
    ↪ Bool)
10 ind -> fun ([Class*] -> [Class*] -> Ontology -> Bool)
11 materials -> fun ([Class*] -> Ontology -> Bool)
12 artificialMats -> fun (Ontology -> Bool)
13 ont -> if (isInClasses ind materials ont)
14     then (isInClasses ind artificialMats ont)
15     else
16     let matMade = madeOf ind ont in
17     orList (map matMade with x -> isInClasses x artificialMats ont);;

```

La funzione `madeOf` restituisce una lista di individui che sono tutti i materiali di cui è costituito un abito, per fare questo `madeOf` prende come parametri l'abito e l'ontologia di riferimento.

La funzione `isArtificial` ha 2 comportamenti differenti a seconda dell'individuo che le viene passato come parametro:

- materiale: se l'individuo appartiene all'albero dei materiali, il controllo consiste nello stabilire se questo materiale appartiene alla lista dei materiali artificiali: se è così si restituisce `true`
- indumento: se l'individuo appartiene ad una sottoclasse dei vestiti, prima si crea la lista di materiali di cui è costituito, poi si valuta: se almeno uno di questi è artificiale si restituisce `true`

Per poter fare queste operazioni la funzione riceve come parametri: l'individuo da analizzare, la lista di tutti i materiali, la lista dei materiali artificiali e l'ontologia di riferimento.

4.3.3 Costruzione dell'ontologia

Assembliamo tutti i pezzi visti finora per manipolare l'ontologia `society` aggiungendo tutte le classi e gli individui che ci interessano. Presentiamo subito l'implementazione descrivendola poi passo passo.

Listato 4.4: assemble ontology

```

1 let fashion = loadOntology "europeana_formatted.rdf";;
2 let people = loadOntology "people.rdf";;
3 let society = loadOntology "society.rdf";;
4
5 let materials = subClassesRec <owl:Class
    ↪ rdf:about="http://thesaurus.europeanafashion.eu/thesaurus/10346"> []
    ↪ fashion;;
6 let artificialMats = subClassesRec <owl:Class
    ↪ rdf:about="http://thesaurus.europeanafashion.eu/thesaurus/10358"> []
    ↪ fashion;;

```

```

7  let fashionObjects = subClassesRec <owl:Class
   ↪  rdf:about="http://thesaurus.europeanafashion.eu/thesaurus/10000"> []
   ↪  fashion;;
8
9  let newMaterials : [Class*] = select x from x in materials
10     where (not contains x artificialMats);;
11
12 let newFashionIndividual : [Individual*] = select x from x in
   ↪  [fashion]/Individual
13     where (not isArtificial x materials
   ↪     artificialMats fashion);;
14
15 let socPeople = head ([society]/<owl:Class
   ↪  rdf:about="http://www.semanticweb.org/society#people"> _) :? Class;;
16
17 let newSociety : [Thing*] = select x from x in [society]/(Thing \ Ont)
18     where (not x = socPeople);;
19
20 let socPeople :? Class = match socPeople with <owl:Class rdf:about=str> [
   ↪  (attr::ClassAtt)* ] -> <owl:Class rdf:about=str> (attr @ [
   ↪  <owl:equivalentClass rdf:resource="http://www.people#People"> [] ]) ;;
21
22 let newPeople : [Thing*] = [people]/(Thing \ Ont);;
23
24 let newOnt :? Ont = <owl:Ontology
   ↪  rdf:about="http://www.semanticweb.org/society_merged"> [];;
25
26 let newThing : [Thing*] = [ newOnt ] @ newSociety @ [ socPeople ] @
   ↪  newFashionIndividual @ newMaterials @ newPeople;;
27
28 let newOntology :? Ontology = <rdf:RDF
   ↪  xml:base="http://www.semanticweb.org/society_merged"> newThing ;;
29
30 dump_to_file_utf8 "society_merged.rdf" (print_xml_utf8 newOntology);;

```

Da riga 1 a riga 3 carichiamo le ontologie che useremo, poi andiamo a creare le liste di classi di materiali che ci servono per utilizzare le funzioni definite in 4.2 e 4.3: tutti i materiali e materiali artificiali. Creiamo anche la lista di tutte le classi di vestiti.

Alla riga 9 costruiamo la lista di classi dei materiali ammissibili per la nuova ontologia selezionando dalla lista totale tutti quelli non artificiali. In riga 12 estraiamo tutti i vestiti (individui) costituiti solo da materiali naturali (nella nuova ontologia inseriamo tutte le classi di vestiti, queste possono potenzialmente rappresentare individui fatti di fibre naturali).

Alla riga 17 prendiamo tutti gli elementi dell'ontologia `society` escluso l'elemento di tipo `Ont` e la classe `people`, questo perché l'elemento `Ont` verrà ricreato da zero alla riga 24 e la classe `people` andrà modificata per renderla equivalente alla classe `people` dell'ontologia `people` (riga 20)

In questo esempio prendiamo tutti gli elementi dell'ontologia `people`; come detto prima, presentare il codice per selezionare solo alcune persone sarebbe poco istruttivo. Una volta create tutte le nuove liste di elementi le assembliamo alla riga 26 e, alla riga 28 costruiamo la nuova ontologia; in riga 30 la salviamo su file.

4.3.4 Risultato finale

L'ontologia che abbiamo costruito permette di descrivere la società nelle modalità che ci eravamo prefissati all'inizio del capitolo (4.1), inoltre abbiamo importato già tutte le persone che avevamo modellato nell'ontologia **people** mantenendo le loro relazioni di parentela; possiamo definire nuove relazioni di parentela sulle persone già modellate in **society** (oppure attribuire loro una data di nascita) e usare le relazioni definite in **society** per arricchire la descrizione di un individuo presente in **people**. Abbiamo anche importato tutte le classi di vestiti e tutti i materiali naturali dall'ontologia **fashion**, possiamo ora creare nuove relazioni tra persone e vestiario per aggiungere informazioni sugli usi e costumi della società che intendiamo descrivere.

4.4 In Protégé

In protégé esiste un comando per fare il merge di ontologie, una volta aperte tutte le ontologie che siamo interessati a fondere possiamo farne il merge e Protégé si occupa di creare una nuova ontologia e inserire tutte le classi, gli individui, le relazioni e le proprietà di tutte le ontologie di partenza.

Questo approccio è molto comodo se siamo interessati a prendere tutti i concetti dalle ontologie di partenza, in questo caso l'unico lavoro che ci rimane da fare è quello di mettere in relazione corretta le classi (in questo caso rendere equivalenti le due classi che descrivono le persone).

Nel caso in cui volessimo fare una selezione più fine su quali concetti importare dovremmo modificare le ontologie di partenza costruendo e scegliendo a prescindere (mediante delle query) le classi e gli individui che vogliamo importare oppure, una volta fatto il merge, rimuovendo ciò che non ci interessa.

4.5 Conclusioni

Il merge con CDuce richiede di scrivere funzioni che permettano di selezionare cosa importare, questa precisione nella selezione comporta, però, più responsabilità da parte del programmatore che deve fare in modo che i dati continuino ad essere consistenti e che le informazioni non vengano alterate durante il merge. CDuce viene incontro al programmatore con un potente controllo sul tipo restituito dalle funzioni, in questo caso possiamo notare che man mano che creiamo la nuova ontologia in 4.4 verifichiamo che ogni elemento abbia effettivamente il tipo che ci aspettiamo. Controllando il tipo in fase di costruzione siamo sicuri che il documento che salviamo alla fine sia effettivamente un'ontologia e aprendola con un altro strumento (Protégé ad esempio) verrà riconosciuta correttamente e non ci saranno inconsistenze.

Nel caso specifico il controllo di tipo è stato importante, infatti, la modifica della classe **socPeople** (riga 20 in 4.4) non era andata a buon fine e aveva generato un elemento XML che non faceva più il match con l'elemento di tipo **Class** definito in 4.1. Avremmo comunque potuto salvare il risultato, ma se avessimo tentato di trattarlo come un file che descrive un'ontologia aprendolo con un altro software avremmo ottenuto un messaggio d'errore (come se il file fosse corrotto) perché il programma non sarebbe stato in grado di interpretare il file come un file **.rdf** ben formato che descrive un'ontologia.

Il merge con Protégé per quanto lineare richiede comunque di formulare delle query più o meno raffinate oppure di eliminare tutto ciò che non serve agli obiettivi del merge, il processo quindi non è molto più guidato o automatico rispetto alla controparte in CDuce. Come si può vedere dal codice in 4.2 e 4.3 la maggior parte del lavoro consiste nel creare

funzioni che permettano di selezionare cosa importare, una volta fatto questo il codice in [4.4](#) che assembla l'ontologia risulta quasi banale. Selezionare elementi con funzioni in CDuce oppure con query in protégé richiede, in entrambi i casi, un certo impegno a livello di programmazione; considerando inoltre che in Protégé una volta selezionati gli elementi con delle query è comunque necessario eliminare tutti gli altri, CDuce si rivela più efficace potendo importare esattamente ciò che si vuole (a patto di poterlo selezionare).

In conclusione Protégé è uno strumento vantaggioso nei casi in cui si voglia fare il merge tra ontologie mantenendo tutti i concetti rappresentati dalle ontologie di partenza, inoltre manipolando le informazioni con delle query in Protégé siamo sicuri di ottenere delle ontologie consistenti. Quando è importante selezionare finemente i dati da importare, CDuce si rivela lo strumento migliore avendo capacità espressiva e manipolativa superiore al set di strumenti base di Protégé (anche in questo caso non prendiamo in considerazione i plug-in che permettono di sviluppare programmi java per manipolare l'ontologia in Protégé). Con il controllo di tipi, CDuce fornisce comunque un certo supporto per verificare che le manipolazioni sulle ontologie portino a strutture consistenti.

Capitolo 5

Conclusioni

5.1 Maturità di CDuce

Prima di utilizzare un software in applicazioni importanti è bene domandarsi se lo stesso sia affidabile e venga mantenuto; questo al fine di evitare situazioni in cui tutto l'applicativo costruito al di sopra di uno strumento crolli al crollare delle fondamenta. Vediamo quindi qual è lo stato di sviluppo di CDuce e quali sono stati le difficoltà tecniche dell'utilizzarlo.

5.1.1 Il progetto CDuce

Come si può leggere nella pagina GitLab del progetto¹ il progetto nasce e si sviluppa come una ricerca condivisa tra il gruppo di ricerca sui linguaggi all'ENS² di Parigi e il gruppo di ricerca sui Database all'LRI³ ad Orsay. Attualmente il progetto è mantenuto dai PPS laboratory⁴ e dal Toccata Group⁵.

Lo scopo del progetto è quello di costruire un linguaggio per manipolare documenti XML che usi seriamente i tipi XML; questo porta i vantaggi descritti in 1.2.3. L'implementazione attuale ha lo scopo di mostrare proprio queste feature innovative e di validare le scelte di progettazione effettuate.

5.1.2 Stato di sviluppo attuale

Sulla pagina del progetto⁶ le ultime informazioni risalgono al 2021 e spiegano che gli sviluppatori si stanno impegnando in una riscrittura completa del compilatore per separare la libreria che gestisce i tipi dal resto del compilatore. La riscrittura del compilatore potrebbe provocare inconsistenza con le ultime versioni di OCaml pertanto si descrive un modo per aggirare il problema oppure si consiglia di utilizzare la vecchia versione. Nonostante le ultime informazioni sul sito siano del 2021 controllando la pagina GitLab si vede che, anche se il branch “stable” non riceve aggiornamenti da un anno, altri branch sono tutt'ora attivi⁷.

¹<https://gitlab.math.univ-paris-diderot.fr/cduce/cduce>

²<https://www.ens.psl.eu/>

³<https://www.lri.fr/>

⁴<https://www.irif.fr/>

⁵<https://toccata.gitlabpages.inria.fr/toccata/>

⁶<https://www.cduce.org/>

⁷alla data in cui si scrive (10/10/2023) l'ultima attività risulta del 09/10/2023

5.1.3 Difficoltà incontrate

Ho testato CDuce su una distribuzione Ubuntu⁸ e su Arch⁹; su Ubuntu non ho avuto successo (seguono dettagli), quindi tutte le prove descritte sono state fatte su Arch.

Ottenere CDuce

Sulla pagina del progetto si legge che CDuce è pacchettizzato per le più importanti distribuzioni Linux, ma questi pacchetti, almeno per Ubuntu risalgono a versioni molto vecchie (Ubuntu 12.10 e 13.04) e non funzionano nelle distribuzioni attuali.

Seguendo le istruzioni per compilare la nuova versione (sia polimorfa che monomorfa) di CDuce il processo fallisce, probabilmente questo è dovuto al fatto che da quando è stata scritta la guida ad oggi, i pacchetti di OCaml sono stati ulteriormente aggiornati e il workaround descritto non è più sufficiente. Ho provato, questa volta con successo, ad installare la vecchia versione di CDuce (version 0.6.1-rc1), ho scelto la versione monomorfa perché la versione polimorfa veniva descritta come “buggy and experimental”

Il processo di installazione descritto sul sito consiste a grandi linee nel:

- creare un ambiente virtuale per OCaml;
- installare in questo ambiente le vecchie versioni dei pacchetti necessari a compilare CDuce;
- compilare i sorgenti;
- ottenere l'eseguibile di CDuce.

Reperire informazioni

Sul sito si possono trovare molte risorse utilissime per imparare ad usare CDuce, sono presenti in particolare 2 documenti:

- una guida utente¹⁰ che spiega in modo molto rigoroso l'approccio del linguaggio ai vari tipi, operazioni e funzioni. La spiegazione è molto tecnica, presenta ogni argomento nel modo più generale possibile facendo estensivo uso dei pattern, per leggerla è necessaria qualche base sui linguaggi funzionali e su come il pattern matching operi (soprattutto se si intendono consultare solo alcune sezioni);
- un tutorial¹¹ che parte direttamente da degli esempi per spiegare il modo in cui CDuce opera. Gli esempi presentati sono molto ben curati e commentati, e permettono di cominciare subito a sperimentare con lo strumento.

Entrambe le guide hanno una controparte in PDF molto comoda per poter cercare all'interno dell'intero manuale un certo termine.

Le difficoltà sorgono nel momento in cui siamo interessati a certe sezioni del tutorial, alcune parti mancano per intero, ed essendo uno strumento relativamente di nicchia non c'è un forum o una community estesa alla quale si possa ricorrere per domande o dubbi. Questo rende particolarmente difficile capire certe parti della guida utente che senza esempi risultano particolarmente oscure. In particolare durante gli esperimenti eseguiti è stato importante fare pattern sulle sequenze, questa sezione (come le precedenti di quel capitolo) è assente nel

⁸<https://ubuntu.com/desktop>

⁹i use Arch btw: <https://archlinux.org/>

¹⁰<https://www.cduce.org/manual.html>

¹¹<https://www.cduce.org/tutorial.html>

tutorial. Avendo delle basi di Haskell¹² e cercando di interpretare la guida che è piuttosto criptica in quel capitolo si può provare a immaginare quale sia la struttura corretta per fare il match (è servito nelle funzioni `andList` e `orList` in 4.2).

Output

Nei linguaggi funzionali le operazioni di input e output sono sempre delicate, in CDuce per poter esportare un documento XML dobbiamo fare 2 passi: prima trasformiamo l'elemento di tipo `AnyXml` (che contiene il file XML che vogliamo salvare) in una stringa poi facciamo il dump della stringa su file. Per la conversione da XML a stringa abbiamo 2 possibilità, se il documento contiene solo caratteri previsti dalla norma ISO-8859-1¹³ si può usare `print_xml` altrimenti per preservare tutti i caratteri si usa `print_xml_utf8`. Ottenuta la stringa si può fare il dump su file con `dump_to_file` oppure `dump_to_file_utf8`, questo sarà un file XML che nel nostro caso rappresenta un'ontologia.

Provando ad aprire il risultato di un dump con un editor di testo ci si tende conto che tutto il codice XML si trova su una sola riga, manca completamente qualsiasi formattazione e il file risulta dunque illeggibile. Aprendo il file con uno strumento come Protégé non ci sono problemi, il file è ben formato e viene interpretato correttamente, ma se vogliamo vedere il risultato della manipolazione con CDuce senza ricorrere ad altri software appositi ci troviamo in difficoltà¹⁴. Se siamo intenzionati a vedere il risultato del dump di CDuce dobbiamo riformattare il documento per renderlo fruibile. Per fare questo ci sono molte opzioni, in questo caso è stato sviluppato un piccolo programma in java che riformatta il file XML rendendo possibile un'ispezione dello stesso mediante un editor di testo.

Il problema è probabilmente dovuto al fatto che le funzioni di `pretty-printing` di CDuce usano per le operazioni di output funzioni di OCaml, avendo forzato l'allineamento tra i pacchetti di OCaml e CDuce è possibile che non sia tutto esattamente compatibile, questo fa sì che le componenti non funzionali di CDuce e in generale l'interazione tra CDuce e OCaml non funzioni correttamente.

È in ogni caso un peccato che attualmente l'unico modo per poter leggere il risultato dell'esecuzione di un programma scritto con un linguaggio funzionale, usato senza ricorrere a nessuno stratagemma imperativo richieda però un programma in java per essere fruito.

5.1.4 Sviluppi futuri

Considerando che il progetto sembra essere ancora di interesse per i gruppi di ricerca che vi hanno lavorato e che il repository risulta attivo ed utilizzato è auspicabile che la riscrittura del compilatore iniziata nel 2021 termini e la nuova versione di CDuce sia polimorfa e sfrutti l'ultima versione di OCaml, questo ridurrebbe drasticamente i problemi descritti sopra e potrebbe spingere più persone ad interessarsi al progetto e a collaborare per rendere CDuce uno strumento professionale a tutti gli effetti; attualmente CDuce viene considerato dagli sviluppatori stessi come un prototipo di ricerca e non adatto ad applicazioni stabili¹⁵.

5.2 Punti di forza di CDuce

Nonostante le difficoltà incontrate CDuce si è rivelato uno strumento molto potente e versatile per la manipolazione di ontologie, in particolare il fatto di essere un linguaggio

¹²<https://www.haskell.org/>

¹³<https://www.iso.org/standard/28245.html>

¹⁴senza contare che Protégé quando si salva il file sul quale si ha lavorato si lo riformatta correttamente, ma riposiziona gli elementi e aggiunge parti non presenti in origine (anche semplicemente i commenti) rendendo impossibile conoscere il documento originale

¹⁵<https://gitlab.math.univ-paris-diderot.fr/cduce/cduce/-/wikis/home>

funzionale lo rende particolarmente sintetico e leggibile una volta presa dimestichezza coi concetti base.

5.2.1 Sistema di tipi

I tipi descrivono un insieme di valori costruiti in un certo modo, è estremamente facile definire un tipo e la sua struttura, d'altra parte avere dei tipi definiti correttamente ci permette di scrivere funzioni che li manipolano come ci aspettiamo e che restituiscono un elemento esattamente del tipo che vogliamo.

Tutta la verifica dei tipi in CDuce avviene staticamente, saremo quindi sicuri che una funzione trasformi un concetto di un tesoro in una classe di un'ontologia ancora prima di eseguire questa funzione perché CDuce verifica che tutte le trasformazioni che applichiamo permettano di passare da un elemento del primo tipo ad un elemento del secondo.

Il controllo di tipo avviene per tutte le funzioni che definiamo (a patto di specificarne correttamente l'interfaccia) e permette di scrivere del codice in cui la maggior parte del debug può essere fatta in fase compilazione e non a run-time.

5.2.2 Funzioni di ordine superiore

Come tutti i linguaggi funzionali anche CDuce permette di definire funzioni di ordine superiore, queste hanno numerosi vantaggi:

- possiamo scrivere funzioni semplici di cui è facile verificare la correttezza e assemblarle in funzioni più articolate che diventano più trattabili, gestendo in questo modo la complessità;
- le funzioni di ordine superiore permettono di generalizzare il codice in modo estremamente efficace, consideriamo il caso di una funzione che rimuove alcuni elementi di una lista secondo un certo criterio, nel momento in cui cambia il criterio dovremmo riscrivere almeno parte della funzione, se invece implementiamo una funzione per la rimozione selettiva che fra i vari parametri accetta una funzione che specifica se un elemento della lista va eliminato possiamo scrivere una funzione assolutamente generale che possa lavorare con un qualsiasi criterio che andremo volta per volta a specificare. Su questa possibilità torneremo dopo in [5.4.2](#)

5.2.3 Pattern matching

Grazie al pattern matching è possibile effettuare complesse operazioni di estrazione dei dati e manipolazione degli stessi, è un aspetto centrale di CDuce e ne incrementa moltissimo la potenza espressiva. Purtroppo non è lo strumento più semplice da padroneggiare e ad una prima vista può sembrare ambiguo il modo in cui descrivere un certo pattern. Lievi differenze nella definizione del pattern producono effetti differenti, e bisogna prestare particolare attenzione a descrizioni che apparentemente possono sembrare equivalenti.

5.3 Spunti per paragoni futuri

In questo lavoro abbiamo provato a fare dei paragoni tra CDuce e Protégé, sono due strumenti molto differenti e le scale di paragone per forza, a seconda dei parametri considerati, fanno risaltare uno strumento come molto efficace e l'altro come inappropriato. Potrebbe essere quindi interessante confrontare CDuce con altri strumenti più simili.

5.3.1 Confronto con un linguaggio imperativo

Esistono vari linguaggi che incorporano librerie per la modellazione di documenti XML, uno fra tutti è sicuramente java.

JAXP

JAXP[9] è l'API che java mette a disposizione per processare dati in XML. I packages contenuti in questa API consentono alle applicazioni di fare il parsing, trasformare, validare e interrogare con delle query documenti XML. JASP è dotata di un layer che permetta l'indipendenza tra il codice dell'applicazione e il particolare processor XML implementato. JAXP, rispetto a CDuce che è attualmente in stato di sviluppo ed è considerato un prototipo, garantisce la possibilità di sviluppare applicazioni per il processing di file XML che siano perfettamente funzionanti, integrabili e stabili.

Alla luce dei vantaggi di affidabilità offerti da progetti molto più grandi di CDuce potrebbe essere interessante valutare come l'uso di un linguaggio funzionale possa apportare benefici a livello di programmazione, correttezza del codice e leggibilità rispetto ad un linguaggio essenzialmente imperativo come java.

5.3.2 Confronto con un linguaggio funzionale

CDuce nasce con lo scopo di processare documenti XML, esistono altri linguaggi più general purpose che implementano la capacità di elaborare XML tramite librerie. Uno dei linguaggi che presenta ampia scelta di librerie è Haskell. Questo linguaggio è sicuramente più noto di CDuce e sono disponibili, pertanto, numerose risorse online per chiarire domande e dubbi che possono sorgere. Fra le librerie per l'elaborazione di documenti XML citiamo HaXml¹⁶ e Haskell XML Toolbox (HXT)¹⁷, che si basa sulle idee di HaXml, entrambi i progetti risultano attivi e documentati.

Il vantaggio di usare una di queste librerie è la possibilità di interfacciarsi direttamente con Haskell potendo integrare le funzionalità della libreria scelta con le numerosissime funzioni offerte da Haskell (sia prese dai pacchetti base sia implementate da librerie esterne). Anche CDuce dovrebbe fornire integrazione con OCaml, sia permettendo di richiamare OCaml in CDuce sia viceversa; queste funzionalità attualmente, almeno sulla base delle prove effettuate, non funzionano.

5.3.3 Confronti possibili

Abbiamo presentato altri strumenti alternativi a CDuce, i confronti possibili potrebbero basarsi sulle seguenti caratteristiche

- velocità di sviluppo del codice: ci aspettiamo che un linguaggio funzionale possa presentare dei vantaggi soprattutto nell'esprimere delle trasformazioni degli elementi rispetto ad un linguaggio imperativo; sfruttando poi la funzione `map` (sia in CDuce che Haskell) risulta anche immediato applicare la trasformazione ad intere liste di elementi;
- correttezza del software: anche in questo caso un linguaggio funzionale che esegua un controllo statico dei tipi in fase di compilazione può aiutarci ad evidenziare errori prima che si presentino a run-time;

¹⁶<https://archives.haskell.org/projects.haskell.org/HaXml/>

¹⁷<https://wiki.haskell.org/HXT>

- scalabilità e integrazione: probabilmente da questo punto di vista un linguaggio come java offre dei vantaggi essendo pensato per anche per lo sviluppo di applicativi di grandi dimensioni più che Haskell o CDuce (in ogni caso nulla vieta di integrare CDuce in applicazioni scritte in altri linguaggi, questa possibilità è approfondita in [5.4.2](#));
- velocità di esecuzione: durante gli esperimenti si è notato un drastico rallentamento dell'esecuzione quando si usavano funzioni ricorsive per creare gli insiemi di sottoclassi, può essere utile provare almeno in java ad implementare una versione iterativa dello stesso algoritmo per verificare se le prestazioni cambiano significativamente. Per quanto riguarda le query può essere interessante un confronto dato che queste ultime in CDuce sono ottimizzate con le stesse tecniche di ottimizzazione della logica classica SQL.

5.4 Sviluppo di nuovi strumenti

5.4.1 Criticità di CDuce

CDuce si è rivelato essere un valido e potente strumento per la manipolazione di file XML; a prescindere dai problemi riscontrati durante gli esperimenti, che auspicabilmente verranno tutti risolti con il rilascio della nova versione stabile, la più grande difficoltà che un utente incontra quando si appropria all'uso di questo strumento è proprio la difficoltà di esprimere degli algoritmi con un linguaggio funzionale.

Se chi vuole usare CDuce per manipolare delle ontologie non ha grandi competenze in campo di programmazione ed ha usato fino a quel momento solo strumenti grafici per la creazione o manipolazione di ontologie si scontra con una notevole difficoltà di sviluppo degli strumenti, che solo in parte sono guidate dal tutorial presente sul sito, in numerosi casi bisogna essere in grado di interpretare la guida che ad un utente non esperto può sembrare particolarmente oscura.

Discorso analogo vale per un utente già competente nell'uso di linguaggi imperativi che si trova a dover fare i conti con un nuovo paradigma di programmazione senza che la guida o il tutorial aiutino particolarmente. Probabilmente un utente di questo tipo si troverà ancora più confuso di un utente novizio alla programmazione.

CDuce presuppone che i suoi utenti posseggano delle competenze già abbastanza avanzate nell'uso di linguaggi funzionali, che si trovino a proprio agio nello sviluppare algoritmi ricorsivi e ad usare intensivamente il pattern matching. Date le competenze richieste e la curva d'apprendimento particolarmente ripida CDuce potrebbe rivelarsi uno strumento ostico per la maggior parte degli utenti che vogliono concentrarsi sulla creazione di un'ontologia che sia corretta e ben formata e non sullo sviluppo di codice per la creazione stessa.

5.4.2 Interfaccia grafica

Fatte le precedenti considerazioni ne risulta che nonostante CDuce sia uno strumento potente ed espressivo sia anche particolarmente poco fruibile. Per arginare la difficoltà d'uso si potrebbe pensare ad un interfaccia che permetta di rendere grafiche la maggior parte delle operazioni, relegando la parte di programmazione a task specifici o molto puntuali sulla particolare ontologia che si sta manipolando.

Esistono numerose interfacce da cui trarre ispirazione (l'interfaccia di Protégé per citarne una) e CDuce si presta particolarmente bene a creare codice generale:

- parsing di una generica ontologia: come si può notare in [4.1](#) la struttura definita permette di fare correttamente il parsing di tre ontologie completamente differenti, la

struttura non è esaustiva e si potrebbero dover aggiungere altre definizioni di tipi per renderla completamente generale senza rinunciare ai controlli di correttezza¹⁸;

- funzioni generali per trattare con ontologie: guardando il codice in 4.2 ci si rende conto che queste funzioni non hanno alcuna attinenza con le particolari ontologie che stiamo trattando, un parco sufficientemente ampio di queste funzioni sarebbe in grado di estrarre la maggior parte delle informazioni utili da un'ontologia, di selezionare e di assemblare le parti che l'utente vorrebbe poter mantenere;
- funzioni di ordine superiore: laddove le funzioni generali non bastano si potrebbero pensare funzioni di ordine superiore complesse a piacere che servano per trattare in modo molto specifico le ontologie che si stanno trattando, queste funzioni vengono messe a disposizione dell'utente che deve solo integrare piccole e semplici funzioni che prese come argomento dalla funzione di ordine superiore la rendano specifica per la particolare ontologia d'interesse. Questo limita notevolmente la necessità di programmare dell'utente anche quando si voglia costruire degli strumenti molto specifici per la propria ontologia;
- creazione guidata di query: esistono già numerosi strumenti che permettono la creazione di query, anche particolarmente complesse, in modo grafico e guidato; data la profonda somiglianza tra le query espresse in linguaggio SQL e la sintassi in CDuce sarebbe possibile adattare uno degli strumenti già esistenti allo scopo. In questo modo anche le query che sono uno strumento potente ed efficiente messo a disposizione da CDuce diventerebbero facilmente accessibili.

5.5 Conclusioni

CDuce nonostante sia classificato come prototipo è uno strumento per la manipolazione di documenti XML ed in particolare di ontologie completo ed espressivo¹⁹. Offre una documentazione online ben curata e accessibile ad un utente con solide basi di programmazione funzionale, la documentazione è inoltre accompagnata da una serie di esempi che guidano, almeno fino ad un certo punto, un utente più novizio a muovere i primi passi con lo strumento potendo già realizzare programmi utili.

Paragonandolo ad altri strumenti si rivela acerbo²⁰ e il sito riflette questa condizione essendo finito solo per metà (numerosi link e componenti non sono attualmente funzionanti). Nonostante questo, rispetto a software molto più grandi e più specificatamente orientati alla manipolazione di ontologie CDuce permette di esprimere una grande quantità di lavoro in poche righe di codice, in modo elegante e leggibile.

L'elaborazione di molti dati, la selezione secondo certi criteri e la trasformazione di elementi seguendo certi pattern risulta molto naturale in questo linguaggio; se a questi si unisce il fatto che il controllo statico sui tipi che garantisce la costruzione di ontologie ben formate CDuce, nonostante non sia stato pensato specificatamente per questo scopo, si dimostra uno strumento molto buono, migliore, in questi casi d'uso, di Protégé.

¹⁸se si usasse il tipo `AnyXml` la struttura sarebbe generalissima ma non offrirebbe più controlli puntuali sulla correttezza dei tipi ritornati dalle funzioni

¹⁹si fa riferimento alla versione del software 0.6.1-rc1

²⁰gli sviluppi sono sicuramente rallentati dal fatto di essere un progetto portato avanti da un piccolo team

Elenco listati

1.1	persone.rdf	2
2.1	persone.cd	8
2.2	basic functions	10
2.3	basic Query	11
3.1	thesaurus_europeana.cd	14
3.2	ontology_europeana.cd	15
3.3	SKOS_to_OWL.cd	16
3.4	concept_to_class.cd	17
3.5	thesaurus_to_ontology.cd	17
3.6	thesaurus_to_ontology_compact.cd	18
4.1	general structure	21
4.2	usefull function	22
4.3	test artificial	24
4.4	assemble ontology	25

Bibliografia

- [1] Grigoris Antoniou e Frank van Harmelen. “Web ontology language: Owl”. In: *Handbook on ontologies* (2009), pp. 91–110.
- [2] Sean Bechhofer e Alistair Miles. “Using OWL and SKOS”. In: *W3C. Web page* (2008).
- [3] Véronique Benzaken, Giuseppe Castagna e Alain Frisch. “CDuce: an XML-centric general-purpose language”. In: *ACM SIGPLAN Notices* 38.9 (2003), pp. 51–63.
- [4] Tim Berners-Lee, James Hendler e Ora Lassila. “A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities”. In: *Scientific american* 284.5 (2001), pp. 34–43.
- [5] Dan Brickley, Ramanathan V Guha e Andrew Layman. *Resource description framework (RDF) schema specification*. Rapp. tecn. Technical report, W3C, 1999. W3C Proposed Recommendation. <http://www.w3.org/2001/sw/2001/05/rdf-schema/>, 1998.
- [6] Thomas R Gruber. “A translation approach to portable ontology specifications”. In: *Knowledge acquisition* 5.2 (1993), pp. 199–220.
- [7] Pascal Hitzler. “A review of the semantic web field”. In: *Communications of the ACM* 64.2 (2021), pp. 76–83.
- [8] Haruo Hosoya e Benjamin C Pierce. “XDuce: A statically typed XML processing language”. In: *ACM Transactions on Internet Technology (TOIT)* 3.2 (2003), pp. 117–148.
- [9] *Java API for XML Processing*. URL: <https://docs.oracle.com/javase/8/docs/technotes/guides/xml/jaxp/index.html>.
- [10] Adam Kilgariff e Colin Yallop. “What’s in a Thesaurus?”. In: *LREC*. 2000, pp. 1371–1379.
- [11] Daniel Kless, Ludger Jansen, Jutta Lindenthal e Jens Wiebensohn. “A method for re-engineering a thesaurus into an ontology.” In: *FOIS*. 2012, pp. 133–146.
- [12] *Protégé 5 Documentation*. URL: <https://protegeproject.github.io/protege/getting-started/>.
- [13] Stuart J Russell e Peter Norvig. *Artificial Intelligence A Modern Approach*. London, 2010. Cap. 7, p. 235.
- [14] R Sivakumar e PV Arivoli. “Ontology visualization PROTÉGÉ tools a review”. In: *International Journal of Advanced Information Technology (IJAIT) Vol 1* (2011).
- [15] Padmini Srinivasan. “Thesaurus construction”. In: *Information Retrieval: data structures and algorithms* (1992), pp. 161–218.
- [16] Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez e Mariano Fernandez-Lopez. “The NeOn Methodology framework: A scenario-based methodology for ontology development”. In: *Applied ontology* 10.2 (2015), pp. 107–145.