ALMA MATER STUDIORUM
UNIVERSITÀ DEGLI STUDI DI BOLOGNA

Learning Path Recommendation

Studio e progettazione di modelli di raccomandazione di learning paths e di risorse didattiche.

Laurea Magistrale in
Ingegneria e Scienze Informatiche

Relatore: Prof.ssa Antonella Carbonaro

Presentata da: Sokol Guri

Anno accademico 2020-2021
Cesena

Indice

Capitolo 1

Introduzione

Il progresso nell'era digitale degli ultimi anni ha influenzato su tutti gli aspetti della vita quotidiana. Il progresso fatto dalle innovazioni tecnologiche sta ridefinendo e ristrutturando le metodologie di apprendimento. Con tale progresso e miglioramento, oggi gli utenti scelgono l'autoapprendimento attraverso l'internet. Al centro di questo cambiamento radicale sta l'E-learning.

Nei giorni d'oggi l'utilizzo di sistemi e-learning che forniscono risorse educative digitalizzate agli utenti, è diventata una normalità. L'e-learning porta tantissimi vantaggi rispetto ai metodi di apprendimento tradizionali con un insegnate che svolge il ruolo principale. Il vantaggio principale è l'aumento dell'accessibilità e della disponibilità delle risorse, riducendo i costi e rispettando la flessibilità degli utenti. L'e-learning è un tipo di apprendimento a distanza che viene svolto tramite internet dove l'utente può accedere alle risorse in qualsiasi momento da qualsiasi posto.

Le metodologi tradizionali di apprendimento incentrate sull'insegnante sono state utilizzate per tantissimi anni come la soluzione più efficacie e fattibile da implementare per ottenere i risultati migliori. Con il progresso fatto a livello tecnologico sono state rilevate tantissime problematiche che i metodi tradizionali portavano. Nella maggior parte degli utenti tali metodi portavano un disorientamento all'utente, fornendo una varietà di risorse disorganizzate e non strutturate. Questa problematica risultava critica se l'utente aveva un esperienza di apprendimento limitata.

Da tantissime ricerche fatte sull'ambito dell'apprendimento è stato affermato che oltre alla digitalizzazione delle risorse, anche l'ordine delle risorse ha un grande impatto sulla qualità dell'apprendimento. Gradualmente alla metodologia di e-learning, sono stati associatati modelli per l'organizzazione

delle risorse didattiche strutturate in una sequenza di materiali che rispettano l'ordine. Tale sequenza rappresenta un percorso di apprendimento personalizzato per un utente che lo guiderà al raggiungimento degli obbiettivi prefissati. L'utilizzo di E-learning con i modelli di raccomandazione potrebbe ridurre significativamente il tempo necessario per raccogliere e organizzare le risorse e in questo modo migliorare l'esperienza dell'apprendimento.

I modelli di raccomandazione forniscono una sequenza di materiali didattici come *learning-path*, però possono essere applicati su diversi domini. Uno di questo domini, sempre collegato con l'insegnamento e l'apprendimento, è academic-program-path. I modelli applicati a tale dominio prevedono un percorso educativo adatto per un utente in base alle informazioni che si forniscono al modello. Rispettando l'organizzazione strutturale delle istituzioni educative come le università, questi modelli prevedono le scuole, i dipartimenti e le facoltà da raccomandare all'utente in base alle sue caratteristiche (passioni, obbiettivi, stile di apprendimento, etc).

Il progetto di tesi propone un modello di raccomandazione del percorso accademico e del percorso di apprendimento per un utente basato sulla modellazione semantica del dominio. *Pathadora* è l'ontologia progettata per rappresentare e modellare i componenti di questo dominio, incorporando ontologie già esistenti sulla accessibilità e l'organizzazione strutturale delle istituzioni educative. Il modello di raccomandazione si basa su regole che estendono e inferiscono nuove relazioni semantiche tra i componenti dell'ontologia. Per ricevere le richieste di interrogazione e manipolazione della *knowledge* dell'ontologia è stato implementato *pathadora-recommender*, che svolge il ruolo di una engine sempre in esecuzione in attesa per computare la risposta alle richieste ricevute.

In questo documento di relazione verranno introdotto una varietà di modelli di raccomandazione che utilizzano diverse metodologie di progettazione, oltre alle regole semantiche. In seguito verrà spiegato in profondità la soluzione ontologica scelta per il modello di raccomandazione e le problematiche incontrate, focalizzandosi sugli vantaggi e svantaggi di tale scelta. Una sezione della relazione sarà dedicata alla progettazione del *pathadora-client* da parte di Andrea, che interagisce con l'engine e fornisce le richieste aspettando la risposta.

La relazione è organizzata come segue:

- Sezione 1- Introduction: introduce senza approfondire il problema e il tema del progetto, paragonando le soluzione tradizionali con quelle più innovative.
- Sezione 2- Background: *to be done*
- Sezione 3- Web Semantico: *to be done*
- Sezione 4- Analisi: *to be done*
- Sezione 5- Progettazione: *to be done*
- Sezione 6- Implementazione: *to be done*
- Sezione 7- Risultati: *to be done*
- Sezione 8- Conclusioni: *to be done*

Capitolo 2

Background

2.1 Ambito

E-learning è uno strumento importantissimo per il miglioramento della qualità dell'istruzione e formazione, basandosi sulle tecnologie d'informazione. La mancanza di adattamento dei contenutati educativi per gli studenti è un problema che e-learning ha portato sin dall'inizio. Mettere a disposizione degli studenti le stesse risorse allo stesso modo, non è la soluzione migliore e più efficiente possibile.

Provando di risolvere il problema della personalizzazione del modello di raccomandazione di risorse didattiche è nato il concetto di *learning path*. Un learning path (percorso di apprendimento) consiste nella progettazione di una sequenza di attività di apprendimento che aiutino lo studente a raggiungere gli obbiettivi prefissati. Durante gli anni la personalizzazione dei learning paths è diventata un problema importante da risolvere, perché tale soluzione si potrebbe applicare in vari domini, oltre all'aspetto didattico. Oltre alla raccomandazione delle risorse didattiche, si potrebbe raccomandare una sequenza di corsi da seguire, di video tutorial da vedere, di libri da leggere, di medicine da prendere, etc. La dinamicità di applicazione lo rende un problema molto complesso e complicato da trovare una soluzione unica e condivisa per tutti.

Dalla fine degli anni 60', i ricercatori hanno tentato di affrontare la personalizzazione, utilizzando diversi parametri, approcci e algoritmi. In seguito si spiegano i sistemi e le tecniche di raccomandazione studiate e implementate seguendo diverse soluzioni e modelli di progettazione.

2.2 Sistemi e tecniche di raccomandazione

Lo sviluppo di modelli di raccomandazione inizia dagli anni 60' utilizzando una sequenza direzionale di risorse didattiche basandosi sui meccanismi di sequenza del curriculum presentato. Questi meccanismi generavano i learning paths, offrendo un unica soluzione per tutti, siccome fornivano le stesse risorse didattiche. Come già accennato prima, le soluzione "*one-fit-all*" causavano tanti problemi, ignorando la diversità degli studenti.

Per risolvere il problema di una soluzione unica, sono stati introdotti *"personalized learning paths"*, focalizzandosi sulla personalizzazione e l'adattamento della soluzione allo specifico studente. In base ad alcuni paramenti iniziali che rappresentano le caratteristiche dello studente, i modelli di raccomandazione generavano diverse soluzioni.

Con l'introduzione di nuove tecniche di raccomandazione comparivano nuove problematiche. Tramite le personalized learning paths, veniva ignorato il progresso che lo studente stava facendo durante tale percorso. Questo problema influenzava negativamente l'efficienza del percorso di apprendimento. Non tenendo in considerazione una tale problematica si rischiava che lo studente non utilizzasse più questo sistema mostrando una mancanza correlazione tra il processo e il metodo di apprendimento.

Secondo Nabizadeh, Jorge, and Leal (2015) and Nabizadeh et al. (2017), i metodi di personalizzazione dei percorsi possono essere classificati in due macro categorie:

- *Path Generation*: questi metodi si focalizzano sulla generazione dell'intero percorso in unica raccomandazione e la valutazione dell'esperienza viene fatta alla fine del percorso.
- *Path Sequence*: questi metodi si focalizzano nella generazione di una sequenza di passi, che presi insieme formano il learning path. La raccomandazione del percorso intero viene fatta man mano che l'utente valuta l'andamento fino a quel punto.

2.2.1 Path Generation

Il processo di generazione del percorso di apprendimento intero consiste in diverse fasi. Durante la prima fase si stabiliscono le caratteristiche e i requisiti che l'utente dovrebbe specificare come parametri iniziali. Durante

la seconda fase, in base a tali parametri allo studente viene generato e consigliato un unico percorso.

Al posto di generare un percorso per ogni studente, alcune soluzioni tendono di raggruppare gli studenti tramite alcuni parametri in comune e poi raccomandare lo stesso percorso per tutti. Kardan et al. ha chiamato tale metodo come ACO-Map e consiste in una soluzione con due fasi principali. All'inizio si applica l'algoritmo K-means per dividere e raggruppare gli studenti e successivamente si ottimizza la soluzione tramite Ant-Colony-optimizzation in modo da generare un percorso per ogni gruppo.

GLPD (*Groupized learning path discovering*) è un altra tecnica di raccomandazione di gruppo, dove inizialmente viene generato un grafo di topics e poi si raccolgono le conoscenze e preferenze dello studente. Il modello dovrà stimare i limiti temporali per tutti i gruppi di studenti sul completamento di un percorso. In base a tali limiti una certa strategia viene scelta per generare il percorso.

Basandosi sulla teoria dei grafi, Belacel propone una tecnica di raccomandazione dove i vertici rappresentano le risorse e gli archi le relazioni e le dipendenze tra le risorse. La fase iniziale consiste nella potatura del grafo, togliendo i vertici irrilevanti e ragruppando quelli simili. L'algoritmo Branch-and-Bound viene utilizzato per individuare il percorso più breve.

Seguendo la logica di Belacel, CourseNavigator è un altro metodo della categoria di Path Generation che genera un percorso di apprendimento tramite i grafi. CourseNavigator si basa sulla logica dei grafi con un algoritmo di ricerca sul un insieme di percorsi generata dalle caratteristiche dichiarate dell'utente. Le caratteristiche servono come vincoli nell'elenco dei percorsi generati.

Seguendo la stessa logica Xu ha progettato un metodo di raccomandazione che ha come obbiettivo finale il completamento del percorso nel minor tempo possibile massimizzando il voto finale.

ECM (*Educational Concept Map*) viene considerato come uno dei metodi più efficienti della categoria Path Generation per la raccomandazione dei percorsi di apprendimento. Presentata da Adorni e Koceva nel 2015, si focalizza sulla conoscenza iniziale che uno studente ha, prima di iniziare il percorso. Lo studente definisce la sua conoscenza selezionando un insieme di argomenti nella mappa dei concetti. ECM utilizza l'algoritmo ENCODE

di Koceva, per linearizzare la mappa in base agli argomenti iniziali e target scelti dallo studente.

RUTICA è un metodo introdotto da Nabizadeh che ha come obbiettivo finale la generazione di un percorso che massimizza il voto finale allo studente con un limite temporale. Utilizzando l'algoritmo DFS (*Depth First Search*) si identificano tutti i possibili percorsi tenendo in considerazione il tempo. Per ogni percorso generato viene stimato e calcolato il voto finale.

Utilizzano i Markov chain, Xia ha progettato un sistema di raccomandazione basandosi su una sequenza di domande che lo studente dovrà rispondere. Le domande sono personalizzate in base allo studente e si focalizzano ad estrarre lo storico dell'esercitazione dello studente e altri concetti comuni che lo studente potrebbe avere con altri studenti.

Christudas, nel 2018, ha proposto una tecnica di raccomandazione basandosi sul CGA (*Compatible Genetic algorithm*). Tale tecnica si basa sullo stile di apprendimento, la conoscenza e l'interattività dello studente.

In uno degli studi più recenti in questo campo, Liu utilizza complex-network-theory definendo i percorsi in tre scenari diversi basandosi sullo storico di apprendimento dello studente.

I metodi di Path Generation sono utilizzati ampiamente anche se portano alcuni svantaggi. Tali metodi ignorano e non prendono in considerazione step intermediari durante il percorso generato. Nel caso dei percorsi di apprendimento, i modelli non considerano l'andamento dello studente e un probabile aggiornamento dei parametri dichiarati. Il percorso è statico ed è impossibile aggiornarlo step-by-step. Un percorso non efficiente porterebbe uno spreco di risorse e un risultato non efficiente per l'utente finale.

2.2.2 Path Sequence

Diversamente dai metodi Path Generation, i metodi Path Sequence raccomandano gli step del percorso man mano che lo studente avanza tenendo traccia dell'andamento. Per risolvere tale problema sono stati utilizzati diversi approcci come: ALN (*Association Link Network*), EAs (*Evolutionary Algorithms*), IRT(*Item Response Theory*), Bayes, etc.

Govindarajan, nel 2016, applico un algoritmo EA, in questo caso PSO *(Parallel Particle Swarm Optimizatio)* per restituire un percorso personaliz-

zato e dinamico per gli studenti. Govindarajan raggruppo gli studenti in base alla loro competenza, ovvero mappando una misura numerica alla possibilità di raggiungimento di un obbiettivo target e il cambiamento di tale unità con l'andamento dello studente. La dinamicità del percorso consiste nella raccomandazione dello step successivo basandosi sul cambiamento di tale unità di misura.

Seguendo la logica presentata da Govindarajan, Li sviluppo un metodo per la generazione dei percorsi dinamici basandosi su due algoritmi, MLE *(Maximum Likelihood Estimation)* e GA *(Genetic Algorithm)*. La fase iniziale del metodo consisteva nella creazione di una sequenza di risorse definendo la loro difficoltà in base ai feedback dello studente. Successivamente viene applicato MLE per analizzare le abilità e gli obbiettivi di ogni studente. Alla fine un algoritmo PSO viene utilizzato per generare il percorso tramite i risultati ottenuti negli step precedenti. Per ogni step compiuto, i feedback dello studente servono ad aggiustare il livello di difficoltà delle risorse ed aggiornare le abilità dello studente.

Nel 2013, Yarandi pubblicava un studio sulla modellazione dei learning path utilizzando meccanismi del web semantico. Yarandi proponeva un sistema adattivo di e-learning utilizzando il modello ontologico del dominio focalizzandosi sulle abilità dello studente. Tale sistema riceveva come parametri d'input le abilità dello studente, la conoscenza, lo stile di apprendimento e le preferenze e alla fine generava un percorso basandosi su questi parametri. L'analisi delle risposte veniva fatta tramite IRT e successivamente aggiornava le abilità dell'utente.

Nello stesso anno, anche Salahli aveva proposto un sistema di raccomandazione utilizzando IRT. Il sistema inizialmente identificava i temi, le loro relazioni e la loro difficoltà. L'algoritmo IRT viene applicato per calcolare il grado di comprensione dei temi e il livello di conoscenza necessario per ogni tema. Quando un utente utilizza il sistema, il suo livello di conoscenza e la difficoltà del tema selezionato sono i parametri che servono per stimare il grado di comprensione.

La generazione di percorsi dinamici offre risultati efficienti nella risoluzione del problema di learning path, però porta alcuni svantaggi per quanto riguarda l'istante di tempo quando si devono aggiornare i profili degli studenti in base ai risultati ottenuti. L'utilizzo di un tempo statico e prefissato potrebbe risultare non efficiente in tutti i casi e porterebbe un risultato non corretto. L'aggiornamento continuo dei profili degli studenti nella maggior

parte dei casi porterebbe uno spreco di risorse computazionali.

2.2.3 IRT (Item Response Theory)

La teoria IR (Item Response) viene nota come la teoria della risposta latente, si riferisce a un insieme di modelli matematici con l'obbiettivo di spiegare le relazioni, i risultati, le risposte e le prestazioni tra tratti latenti. Questa teoria stabilisce un collegamento tra le proprietà degli elementi, gli individui che rispondono e il tratto misurato. IRT si è basata sulla relazione tra le prestazioni degli individui su un elemento di prova e i livelli di prestazione dei partecipanti al test su una misura complessiva dell'abilità che l'elemento è stato progettato per misurare. Ogni risposta a un determinato elemento fornisce un grado in inclinazione sul livello o sull'abilità dell'individuo del tratto latente. L'abilità di un individuo è la probabilità di approvare la risposta come corretta e maggiore è l'abilità, maggiore è la probabilità di una risposta corretta.

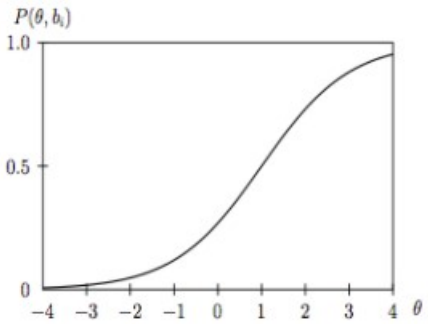

Figura 2.1: Curva caratterista dell'elemento.
La forma del grafico consiste in una S (Sigmoide/Ogiva). La probabilità di ottenere una risposta corretta dipende dalla abilità del intervistato la quale nelle applicazioni rispetta il range `-3` and `+3`.

2.2.3.1 Assunzioni del IRT

- *Monotonicità*: l'assunzione specifica che la probabilità di ottenere una risposta corretta aumenterebbe aumentando il livello del tratto;
- *Unidimensionalità*: si assume che sia un tratto latente dominante da misurare che guida le risposte osservate per ogni elemento della misura;
- *Indipendenza locale*: si assume che le risposte date dagli elementi in un test sono reciprocamente indipendenti dato un livello di abilità;

- *Invarianza*: si possono stimare i paramenti di un elemento da qualsiasi gruppo di soggetto che ha già risposto all'elemento.

Basandosi sulle ipotesi iniziali e considerandole valide, la variazione del tratto latente da parte degli intervistati definisce le differenze nell'osservazione delle risposte corrette. Il modello prevede le risposte degli intervistati su uno strumento basato su elementi tramite la loro posizione nel grafico continuo del tratto latente e alle caratteristiche di tali elementi. Queste caratteristiche vengono considerate paramenti d'input per il modello.

2.2.3.2 Parametri degli oggetti

Le abilità e le capacita delle persone variano e sono molto dinamiche e per questo motivo la loro posizione nel grafico cambia. Il campione degli intervistati e i parametri degli oggetti definiscono la posizione precisa nel grafico.

- *Item Difficulty (bi)*: parametro che definisce il comportamento dell'oggetto lungo la scala delle abilità. Viene calcolato dalla capacità con cui il 50% degli intervistati approvano una risposta come corretta. Gli oggetti che è difficile approvarli sono spostati più a destra, mentre quelli più facili più a sinistra;
- *Item Discrimination (ai)*: parametro che definisce la velocità con cui la probabilità di approvare un oggetto corretto cambia in base ai livelli di abilità. Per ottenere un misura precisa, vengono inclusi elementi con discriminazione alta;
- *Guessing (ci)*: parametro che tiene in considerazione le ipotesi su un determinato elemento limitando la probabilità di approvare la risposta se l'abilità va negativamente all'infinito.

2.2.4 Three parameter logistic model

Nel modello logistico a tre parametri (3PL), la probabilità di una risposta corretta a un oggetto dicotomico i per una domanda a scelta multipla, è:

$$p_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta - b_i)}}$$

dove:

- θ indica che le abilità degli intervistati sono definite in base ad una distribuzione normale per stimare i parametri dell'oggetto. In seguito sono stimate le abilità di ogni intervistato;

- a_i, b_i, c_i sono i parametri dell'oggetto;

Il cambiamento dei parametri viene rappresentato come il cambiamento della forma di una funzione logistica:

$$P(t) = \frac{1}{1 + e^{-t}}$$

dove:

- a - discriminazione, scala e pendenza massimale: $p'(b) = a \cdot (1 - c)/4$; Tale parametro rappresenta il grado della discriminazione dell'oggetto tra regioni diverse del grafico. Per esempio le persone con capacità basse hanno un probabilità molto minore di rispondere correttamente alla domanda.

- b - difficoltà e posizione dell'oggetto: $p(b) = a(1 + c)/2$. $p(b)$ in questo caso è il punto medio tra c_i (min) e 1 (max). In questo punto la pendenza è massimale.

- c - ipotesi: minimo asintotico $p(-\infty) = c$. Il parametro tende a spiegare il caso quando un intervistato con base capacità sceglie la risposta corretta.

Quando $c = 0$, $p(b) = 1/2$ e $p'(b) = a/4$, b è uguale 50% probabilità di avere successo (difficoltà) e a è la pendenza massimale (discriminazione).

2.3 Accessibilità

La potenza del Web sta nella sua universalità.
L'accesso da parte di tutti indipendentemente dalla disabilità è un aspetto essenziale. | Tim Berners-Lee |

Le persone disabili possono utilizzare facilmente le applicazioni e i siti web se sono state progettate e implementate correttamente, considerando i problemi di accessibilità. Molte applicazioni non gestiscono bene tale problema, portando barriere di accessibilità per gli utenti disabili e rendono questi applicazioni difficilmente utilizzabile da loro. Progettare applicazioni con livello alto di accessibilità, porta vantaggi per gli individui, le aziende e la società e per questo motivo sono state definite dei standard internazionali per risolvere tale problematica.

Il Web è stato progettato per essere utilizzato da tutte le persone, indipendentemente dal loro hardware, software, lingua, posizione o capacità. Il Web rimuove le barriere alla comunicazione e all'interazione che molte persone affrontano nel mondo fisico, però quando i siti Web, le applicazioni, le tecnologie o gli strumenti sono progettati male, possono creare barriere che escludono le persone dall'utilizzo del Web.

L'accessibilità al Web significa che i siti Web, le risorse, le informazioni, i tool e le tecnologie sono state progettate e implementate in modo che le persone con disabilità possono comprendere, percepire, navigare, utilizzare e contribuire al web. Le disabilità delle persone comprendono problemi uditivi, cognitivi, neurologici, fisici, di discorso, visivi, etc.

2.3.1 Linee guida per l'accessibilità

WCAG (*Web Content Accessibility Guidelines*) è un documento pubblicato da Accessibility Guidelines Working Group e copre una lista di raccomandazioni per rendere i contenuti Web più accessibili per le persone con disabilità diverse. Le linee guida specificate nel documento riguardano l'accessibilità dei contenuti sui diversi dispositivi elettronici e il rispetto di queste linee guida renderà i contenuti Web più fruibili per gli utenti in generale.

Le WCAG vengono utilizzate da individui che variano ampiamente e per poter soddisfare le diverse esigenze di individui diversi, vengono forniti diversi livelli di guida tra cui:

- *principi generali*: I principi più importanti che forniscono la base dell'accessibilità sono la percepibilità, l'operabilità, la comprensibilità e la robustezza;
- *linee guida generali*: WCAG introduce 13 linee guida che forniscono gli obbiettivi di base verso i quali gli autori dovrebbero focalizzasi per rendere i contenti e le risorse web più accessibili;
- *criteri di successo verificabili*: Per ogni linee guida vengono forniti criteri di successo verificabili per consentire l'utilizzo di WCAG.
- *una ricca raccolta di tecniche*: Per ciascuna delle linee guida e dei criteri di successo nel documento WCAG 2.0 stesso, il gruppo ha documentato un'ampia varietà di tecniche informative che si dividono in due categorie:
  - quelle sufficienti per soddisfare i criteri di successo;
  - quelle consultive che oltre quanto richiesto dai singoli criteri di successo, consentono e richiedono agli autori di affrontare meglio le linee guida;

2.3.2 Accessibilità delle risorse

Una risorsa accessibile è un risorsa creata e studiata per essere facilmente accessibile da un utente indipendentemente, dalla sua disabilità o dal formato della risorsa. Rendere accessibile una risorsa è più semplice quando siamo nelle fasi iniziali della creazione. Quando parliamo di risorse online di diversi tipi (testuali, immagini, video, audio) ci sono alcune considerazioni che dobbiamo verificare in modo da offrire una risorsa accessibile per utenti diversi. Nel caso delle risorse testuali, il livello d'accessibilità di un documento dipende da:

- *la struttura*: una struttura standardizzata di documenti migliora l'accessibilità, perché consente agli utenti di trovare i contenuti in una parte specifica della pagina, senza aver bisogno di leggere tutto il documento;
- *le intestazioni*: i titoli devono essere descrittivi e in un ordine coerente;
- *i floating elements*: questi elementi, comprese anche le immagini, vanno inseriti all'interno della sezione di appartenenza;
- *la risoluzione*: i documenti dovrebbero essere accessibili ai lettori che utilizzano dispositivi con schermi piccoli o ai lettori che utilizzano monitor a bassa risoluzione;

- *il testo*: le dimensioni dei caratteri ridotti o ingranditi, il tipo dei caratteri, l'indentazione e lo spazio sono parametri che dovrebbero essere utilizzati in modo da creare un stile da rispettare per il documento e ciò migliora l'accessibilità della risorsa;

- *i colori*: definiscono un alto livello di accessibilità se definiti con accuratezza, ma non utilizzare testo o sfondo colorato perché lettori non vedenti che accedono tale risorsa tramite una stampa o un dispositivo senza schermo a colori non riceveranno tali informazioni;

- *i blocchi di elementi*: gli elementi dell'elenco non si devono separare lasciando righe vuote o interruzioni di colonna tabulari tra di loro;

Le considerazioni sopraindicate aiutano al miglioramento della accessibilità dei documenti testuali, però ogni tipo di risorsa dovrà rispettare le linee guida specificate. Le immagini che non sono puramente decorative dovrebbero includere un attributo *alt* che sostituisce l'informazione dell'immagine per i lettori non vedenti. Per le immagini è una buona prassi includere una didascalia utilizzando la sintassi dell'immagine incorporata, descrivendo in modo conciso il significato dell'immagine e le informazioni essenziali che trasmette. Il posizionamento delle immagini può essere diverso in base al dispositivo o al formato del documento e per questo motivo si deve evitare il riferimento statico delle immagini come di sinistra o destra, si deve utilizzare invece le didascalie per identificarle e riferirle.

Per rendere una risorsa video accessibile, si possono aggiungere i sottotitoli in un formato di testo temporizzato, scaricabile a parte. I sottotitoli serviranno per la trascrizione delle informazioni trasmesse dal video. La stessa cosa deve valere anche per le risorse nel formato audio, dove trascrivendo il discorso tramite i sottotitoli, renderebbe la risorsa facilmente accessibile da un gruppo di utenti più ampio.

Capitolo 3

Web Semantico

Il web semantico è il grafo della conoscenza formato combinando dati collegati di Linked Data con contenuti intelligenti, metadati e altri oggetti informativi su larga scala per facilitare la comprensione e l'elaborazione dei contenuti da parte delle macchine.
https://simplea.com/Articles/what-is-the-semantic-web

3.1 La nascita del web semantico

Con lo sviluppo della tecnologia e del mondo d'internet, è stato possibile navigare diversi testi grazie ai link, che permettono di accedere altre informazioni cliccando tali link. L'internet non consiste solo nella gestione di grandi blocchi di informazioni, ma anche a scoprire e aggregare le associazioni tra tali blocchi, collegate tramite i link. Le relazioni tra i blocchi sono state rilevate da una serie di algoritmi che hanno rivoluzionati il Web da un contenitore di contenuti, ad un ambiente dove facilmente si possono trovare i contenuti richiesti tramite le associazioni tra informazioni e dati. Queste associazioni permettono di rintracciare tutto quanto concesso ad un concetto, un dato o una parola.

Per tanti anni si parlava di web di documenti, considerando i documenti come l'elemento più importante della navigazione nei sistemi web based. Nei giorni d'oggi si menziona sempre di più, web di dati, come nei database. L'obbiettivo del web di dati è di abilitare i compilatori ad effettuare un lavoro più efficiente e di sviluppare sistemi che possano supportare iterazioni sicure nella rete.

Il termine *web semantico* rappresenta una grande quantità di dati interconnessi tramite relazioni e facilmente estendibili e accessibili dagli utenti. L' articolo Scientific American considera il web semantico come *"un'estensione del web attuale dove alle informazioni viene dato un significato ben definito, consentendo ai computer e alle persone di lavorare in cooperazione."*

I dati, arricchiti con semantica, struttura e collegamenti significativi e interpretabili dalla macchina, consentono ai computer di trovare e manipolare le informazioni con maggiore precisione.

3.2 Lo sviluppo di web semantico

Per molti ricercatori web semantico è stato evoluto durante due fasi importantissime, la prima fino al 2006 e la seconda fino nei giorni d'oggi. Durante la prima fase di sviluppo si è lavorato sulla progettazione di un metodo sintetico oppure chiamato diversamente delle ontologie. Durante la seconda fase si è stato utilizzato il metodo analitico, o dei linked data. Il processo di sviluppo si è diviso in due fasi, perché il metodo sintetico era insufficiente a ottenere risultati concreti e intelligenti richiesti dagli utenti. L'obbiettivo finale del mondo del web semantico è quello di permettere ai calcolatori di comprendere al meglio la semantica delle ricerche dell'utente.

Qualsiasi algoritmo di ricerca è caratterizzato da tre elementi fondamentali:

- *organizzazione della conoscenza*: consiste nell'identificazione degli elementi che serviranno per la ricerca;
- *rappresentazione della conoscenza*: consiste nella descrizione dei dati;
- *strumentazione per accedere alla conoscenza*: tramite strumenti software sarà effettuata la ricerca di tutte le rappresentazioni di un dato che si trovano in rete.

Secondo Berners-Lee, i due componenti principali per sviluppare una repository d'informazione sono i metadati e il risultato di un collegamento ipertestuale. L'introduzione dei metadati permetteva ottenere informazioni sulle informazione trovate sul web che potevano essere interpretate dai calcolatori. L'utilizzo di valori ipertestuali è stato utile ad indirizzare le ricerche sulla rete. Il web semantico potrebbe essere definito come uno sviluppo del Web 3.0, che si focalizza sul miglioramenti dell'infrastruttura dei dati, in particolare sull'etichettamento dei dati in modo sa supportare le ricerche in

linguaggio naturale.

Aggiungere informazioni sulle informazioni, per molti ricercatori è stato considerato come una operazioni pesante e intensiva per il sistema, ma tale idea è stata implementata da Yahoo nel 2008.

3.3 Stack del web semantico

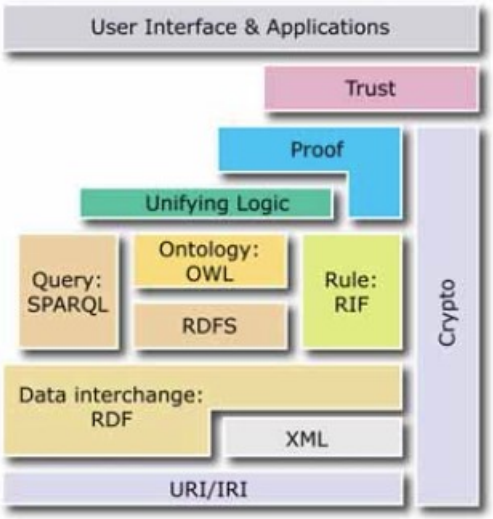

Figura 3.1: Stack del web semantico

La struttura del web semantico si rappresenta da tre livelli importanti, dove ogni strato viene suddiviso in sotto livelli:

- *livello base*: comprende le basi dei protocolli di comunicazione via Web (IRI, URI, Unicode);
- *livello corse*: comprende una serie di linguaggi con lo scopo di descrivere semanticamente le informazioni (RDF, OWL, SPARQL, etc);
- *livello finale*: comprende meccanismi che sono ancora in fase di sviluppo, con tendono a sincronizzare la progettazione semantica con intelligenza artificiale.

3.3.1 IRI e URI

Il livello base racchiude le tecnologie ipertestuali che forniscono le basi per il web semantico. IRI (*Internationalized Resource Identifier*), permette l'i-

dentificazione univoca delle risorse condivise nella rete. Le IRI aiutano nella generazione dei URI (*Uniform Resource Identifier*). Un URI è una stringa di caratteri che viene utilizzata per identificare univocamente una risorsa generica su Internet. Questo meccanismo di identificazione permette l'iterazione in rete con le risorse usando protocolli specifici.

Il mondo del web semantico necessita questo meccanismo di identificazione di informazioni per consentire la ricerca e la manipolazione delle risorse dai livelli superiori.

3.3.2 Unicode

La manipolazione e la rappresentazione delle informazione in diversi linguaggi viene effettuato tramite lo strumento di unicode. Tale strumento è uno standard informatico che permettere ai compilatori di rappresentare in maniera consistente e di manipolare i testi espressi in diverse lingue del modo. Non è altro che un sistema di codifica che assegna dei bit ad ogni carattere indipendentemente dal programma, dalla piattaforma o dalla lingua. Nello standard sono stati codificati i caratteri di tutte le lingue, i simboli matematici e chimici, i segni cartografici, gli ideogrammi, i simboli musicali, etc. In totale sono stati codificati più di 107.000 caratteri.

3.3.3 XML

Il livello core, racchiude le tecnologie più importanti che modellano semanticamente le risorse nella rete. XML (*eXtensible Markup Language*) è un metalinguaggio modella la struttura dei documenti e delle informazioni tramite un insieme di regole semantiche.

3.3.4 Namespace

Un namespace consiste in un insieme di attributi identificati univocamente da un identificatore e vengono utilizzati per dichiarare diversi sorgenti d'informazione per lo stesso concetto. L'associazione delle informazioni tra di loro necessità il riferimenti di più fonti.

In modo da utilizzare una lista di termini, si dovrebbe specificare una indicazione precisa dei vocabolari che verranno utilizzati. In questo modo si

ottiene un significato non ambiguo dagli identificatori e il documento ha una leggibilità migliore.

3.3.5 RDF

RDF (*Resource Description Framework*) permette la definizione di informazioni descrittive sulle risorse documentali tramite un insieme di regole. RDF viene utilizzato per la modellazione delle informazioni utilizzando le notazioni sintattiche e formati di serializzazione dei dati.

Il modello RDF assomiglia con i classici approcci di modellazione concettuale. Il modello si basa sulle asserzioni che vengono rappresentate tramite triple (soggetto, predicato, oggetto) che effettuano una relazione binaria tra gli elementi. Il predicato denota caratteristiche della risorsa ed esprime la relazione tra il soggetto e l'oggetto.

3.3.5.1 Classi e proprietà di RDF

RDF Schema fornisce strumenti per combinare tra loro le asserzione e le descrizioni in un singolo vocabolario. Gli elementi principali che permettono la definizione di un vocabolario sono:

- le classi:
  - `rdf:XLMLiteral` : la classe dei valori XLM;
  - `rdf:Property` : la classe delle properità;
  - `rdf:Statement` : la classe delle dichiarazioni;
  - `rdf:Alt, rdf:Bag, rdf:Seq` : le classi che rappresentano i contenitori di diversi tipi (di ordinati, non ordinati);
  - `rdf:List` : la classe che modella le liste;
  - `rdf:nil` : istanza di `rdf:List` che rappresenta una lista vuota;
  - `rdfs:Resource` : la classe delle risorse;
  - `rdfs:Literal` : la classe dei valori;
  - `rdfs:Class` : la classe delle classi;
  - `rdfs:Datatype` : la classe dei tipi di dato RDF;
  - `rdfs:Container` : la classe dei contenitori RDF;

- le proprietà:
  - `rdfs:type`: istanza di `rdf:Property` utilizzata per affermare che tale risorsa è istanza di una certa classe;
  - `rdfs:first`: il primo elemento di una lista RDF;
  - `rdfs:rest`: il resto della lista, ignorando il primo elemento;
  - `rdfs:value`: proprietà utilizzata per i valori strutturati;
  - `rdfs:subject`: il soggetto si una dichiarazione;
  - `rdfs:predicate`: il predicato di una dichiarazione;
  - `rdfs:object`: l'oggetto di una dichiarazione;
  - `rdfs:subClassOf`: il soggetto è una sottoclasse di una classe;
  - `rdfs:subPropertyOf`: il sogetto è una sotto proprietà di una proprietà;
  - `rdfs:domain`: dominio del soggetto di una proprietà;
  - `rdfs:range`: l'intervallo del soggetto di un proprietà;
  - `rdfs:label`: il nome del soggetto;
  - `rdfs:comment`: la descrizione del soggetto della risorsa;
  - `rdfs:isDefinedBy`: la definizione del soggetto di un a risorsa;

3.3.5.2 Serializzazione dei documenti RDF

Diversi formati di serializzazione sono stati progettatati per la modellazione semantica tramite RDF delle risorse nel web.

Turtle (*Terse RDF Triple Language*) è un formato di sintassi e di documento per rappresentare le informazioni tramite il modello di dati RDF. Questo formato prevede un raggruppamento di URI dei componenti della tripla (soggetto, predicato, oggetto) abbreviando l'informazione, per esempio aggregando le parti in comune delle URI.

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

[ foaf:name "Alice" ] foaf:knows [
    foaf:name "Bob" ;
    foaf:knows [
        foaf:name "Eve" ] ;
    foaf:mbox <bob@example.com> ] .
```

N-Triples è un altro formato per effettuare la memorizzazione e il trasferimento dei dati. Progettato per offrire una serializzazione più semplice del formato Turtle, più facile da parsare dai componenti software e rappresentare meglio le risorse innestate.

```
_:alice <http://xmlns.com/foaf/0.1/knows> _:bob .
_:bob <http://xmlns.com/foaf/0.1/knows> _:alice .
```

JSON-LD (*JavaScript Object Notation for Linked Data*) è un altro metodo per rappresentare la sintassi dei file che contengono informazioni semanticamente modellate. Il vantaggio di tale formato rimane la facilita di trasformazione dei file JSON in un JSON-LD, offrendo ai sviluppatori la possibilità di manipolare questi file utilizzando i metodi tradizionali.

```
{
  "@context": {
    "name": "http://xmlns.com/foaf/0.1/name",
    "Person": "http://xmlns.com/foaf/0.1/Person"
  },

  "@id": "https://me.example.com",
  "@type": "Person",
  "name": "John Smith",
}
```

RDF/XML è una dei formati di serializzazione maggiormente utilizzato e condiviso nel mondo del web semantico. Tale formato rappresenta un grafo RDF come un documento XML.

```
<?xml version="1.0"?>

<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:si="https://www.w3schools.com/rdf/">
    <rdf:Description rdf:about="https://www.w3schools.com">
        <si:title>W3Schools</si:title>
        <si:author>Jan Egil Refsnes</si:author>
    </rdf:Description>
</rdf:RDF>
```

Il meccanismo dei RDF per descrivere le risorse, è un componente importantissimo del web semantico che ha permesso la possibilità di archiviare, scambiare, manipolare e utilizzare le informazioni del web in un formato machine-readable. In questo modo è stato consentito agli utenti di gestire le risorse con maggiore efficienza, certezza e facilità. La collezione di dichiarazioni nel formato RDF, rappresenta un multi-grafo diretti e etichettati, rendendo il modello di dati più adatto certi tipi di rappresentazione della conoscenza.

3.3.6 OWL

OWL (*Ontology Web Language*) è un linguaggio di markup utilizzato per rappresentare il significato delle informazioni utilizzando le relazioni tra le informazioni e i vocabolari. OWL è stato progettato per rappresentare la conoscenza arricchita e complessa sugli elementi, gruppi di elementi e le relazioni tra loro. Secondo W3C, OWL è un linguaggio computazionale basato sulla logica, in modo che la conoscenza potrebbe essere utilizzata dai compilatori e dai programmi. I documenti OWL, vengono conosciuti come ontologie e una volta pubblicate nel WWW, potranno riferire o essere riferite da altre ontologie. La versione più recenti di OWL è OWL 2, sviluppata e pubblicata nel 2012.

Lo modellazione ontologica assomiglia con l'approccio del paradigma ad oggetti, organizzando le informazioni il classi e gerarchie di classi. Tramite le classi, le ontologie modellano informazioni e risorse che cambiano costantemente e velocemente nel web. Le ontologie riescono a rappresentare i dati provenienti da diversi fonti e sorgenti in modo molto flessibile.

Un dominio rappresentato da una ontologia tramite OWL, viene interpretato come una lista di *"individual"* e una lista di asserzioni di proprietà. Le asserzioni stabiliscono connessione tra diversi individui. Un'ontologia è costituita da un insieme di assiomi che definiscono vincoli su insiemi di individui e sui tipi di relazioni consentite tra di essi. Questi assiomi forniscono la semantica consentendo ai sistemi di dedurre ed inferire informazioni aggiuntive basate sui dati forniti esplicitamente.

3.3.6.1 Sottolinguaggi di OWL

Il linguaggio OWL prevede una varietà di sottolinguaggi con un maggiore livello di espressione, tra i quali:

- *OWL Lite*: linguaggio che supporta gli utenti a cui serve una gerarchia di classificazione e semplici vincoli di funzionalità.
- *OWL DL*: linguaggio che supporta gli utenti che vorrebbero massimizzare l'espressività senza perdere la completezza e la correttezza computazionale e la decidibilità del sistema di ragionamento. OWL DL, prende il nome dalla Description Logics, che si focalizza sulla decidibilità basandosi sulla logica del primo ordine;
- *OWL Full*: linguaggio che maggiore espressività e un sintassi svincolata con nessuna garanzia a livello computazionale.

3.3.6.2 Contenuto dell'ontologia

Un documento owl che rappresenta una ontologia inizia con la dichiarazione del namespace con gli identificatori e le URI dei vocabolari. Una volta dichiarato il namespace, viene specificata un insieme di asserzioni sotto `owl:Ontology`, aggiungendo informazioni sull'ontologia. Queste informazioni sono metadati che servono per controllare la versione del documento, le ontologie importate, eventuali commenti etc.

Importando una ontologia, si potrebbe accedere è utilizzare tutti i componenti di tale ontologia, definendo relazioni tra elementi di diverse ontologie. Le ontologie importate saranno sincronizzate con il namespace dichiarato. A volte l'importo di una ontologia potrebbe non avere successo siccome il sistema dovrebbe cercare le risorse condivise nella rete, che non è sempre possibile.

In seguito verranno spiegate gli elementi e i tag più importanti che OWL offre per la modellazione di un ontologia.

3.3.6.3 OWL classes

Le classi in OWL, sono uno strumento di astrazione importantissimo per raggruppare le risorse con caratteristiche simili. Come nelle classi RDF, ogni classe OWL è associata ad un insieme di individui, che vendono considerati come istanze della classe o estensione della classe. Il significato di una classe non è correlato con il significato della sua estensione, perché due classi possono avere la stessa estensione ma sono comunque diverse.

Le classi nell'ontologia vengono descritte tramite un insieme di assiomi utilizzando il nome della classe oppure specificando l'estensione della classe

come individuo di una classe anonima. OWL offre diversi modi per descrivere le classi tramite:

- un riferimento URI della classe: la descrizione viene fatta tramite il nome della classe.
- una enumerazione esaustiva degli individui che raggruppati insieme formano un istanza della classe;
- una restrizione di proprietà: l'insieme gli individui soddisfano la restrizione;
- l'intersezione, l'unione o il complemento di diverse classi.

3.3.6.4 OWL object properties

Le proprietà permettono di specificare caratteristiche associate ai membri di una certa classe e agli individui. Le proprietà degli oggetti definiscono relazioni tra istanze di due classi. Un assioma di proprietà definisce le caratteristiche della proprietà. La dichiarazione di una relazione tra le istanze si potrebbe vincolare definendo delle assiomi:

- sui costrutti del schema RDF: il domino e l'intervallo dei possibili valori utilizzando rispettivamente `rdf:domain` e `rdf:range` ;
- sulle relazioni con altre proprietà utilizzando `owl:equivalentProperty` oppure `owl:inverseOf` ;
- sulle cardinalità globali in questo caso utilizzando `owl:FunctionalProperty` e `owl:InverseFunctionalProperty` ;
- sulle caratteristiche logiche della proprietà tramite `owl:SymmetricProperty` e `owl:TransitiveProperty` .

Le assiomi sopraelencato sono quelle più importanti, però OWL offre un insieme molto ampio di assiomi per poter aggiungere caratteristiche e informazioni sulle proprietà.

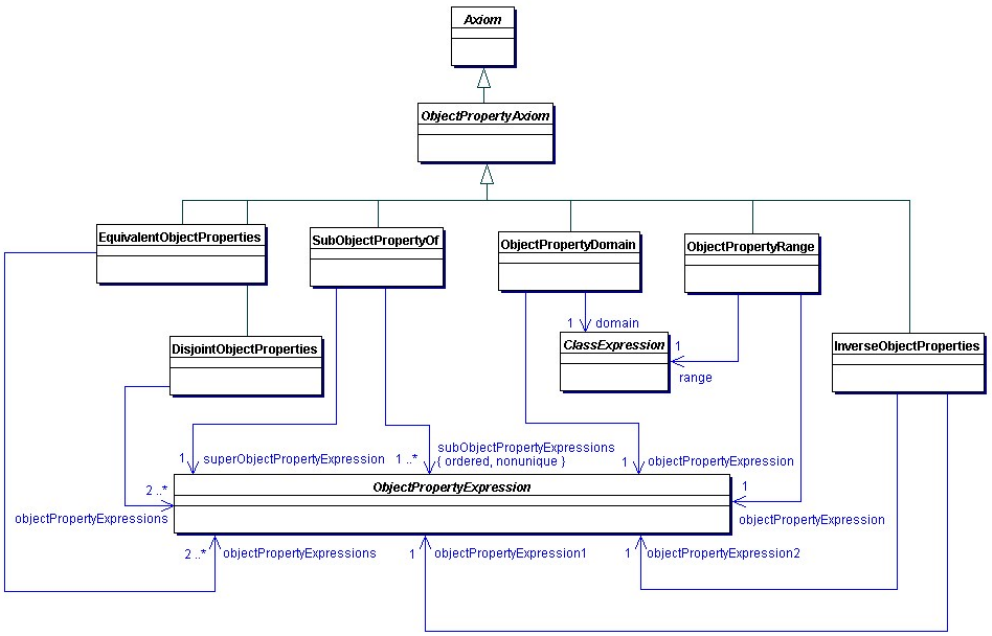

Figura 3.2: Le assiomi più importanti delle proprietà di oggetti

3.3.6.5 OWL data type properties

Le proprietà nelle ontologie vengo divise in base agli elementi che si riferiscono. Quando le proprietà si riferiscono a tipi di dati, parliamo di data type properties e si basano sui tipi di dati definiti tramite RDF e XML Schema. I tipi di dati più utilizzati con OWL sono: `xsd:string`, `xsd:decimal`, `xsd:integer`, `xsd:long`, `xsd:boolean`, `xsd:byte`, etc.

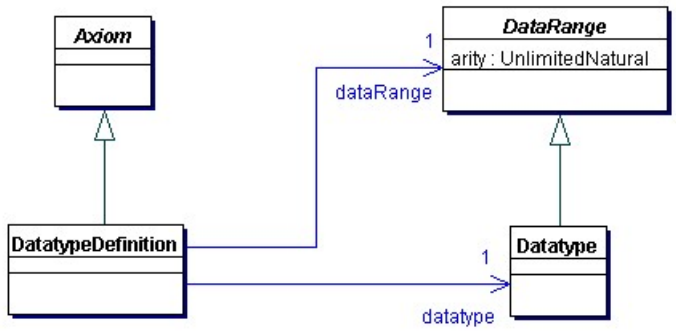

Figura 3.3: Definizione dei tipi di dato

3.3.6.6 OWL annotation properties

OWL offre le annotazioni come strumento per associare nuove informazioni all'ontologia, alle entità, alle assiomi, etc. La sintassi di OWL offre un meccanismo per incorporare i commenti nei documenti dell'ontologia.

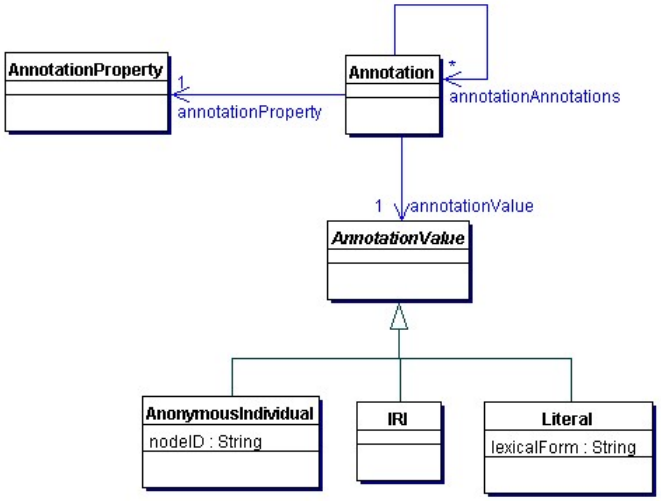

Figura 3.4: Annotazione delle ontologie in OWL

3.3.6.7 OWL individuals

Gli individui in OWL 2, rappresentano oggetti del dominio e possono essere di due tipi:

- *named individuals*: agli individui è stato dato un nome esplicito;
- *anonymous individuals*: gli individui non hanno un nome globale e sono locali all'ontologia alla quale fanno parte.

Gli individui nominati vengono identificato tramite il riferimento IRI perché vengono considerati come entità. Nel caso degli individui anonimi che non debbano essere utilizzati al di fuori dell'ontologia, loro vengono identificati tramite un nodo locale.

3.3.7 SPARQL

Lo SPARQL (*SPARQL Protocol and RDF Query Language*) è un linguaggio semantico per le interrogazioni dei database offrendo la possibilità di ricerca e di manipolazione dei dati memorizzati nei documenti rappresentati tramite il formato RDF.

Questo linguaggio potrebbe essere utilizzato per esprimere query tra diverse sorgenti di dati indipendentemente dal formato dei documenti. Tutto questo è possibile tramite i middleware che sono applicazioni che si interpongono in modo da offrire l'accessibilità sui diversi formati di documenti. Lo SPARQL offre la possibilità di fare query su diversi modelli di grafo messi insieme applicando su di loro delle operazioni. Il risultato delle interrogazioni potrebbe essere un insieme di dati oppure grafi RDF.

La sintassi e la semantica del linguaggio per interrogare i documenti si basa strettamente su due specifiche:

- `SPARQL Protocol for RDF`: definisce il protocollo per fare ed eseguire le query e ottenere il risultato da remoto;
- `SPARQL Query Result XML Format`: definisce il formato per rappresentare i risultati delle query eseguite.

Uno dei vantaggi più importati del linguaggio SPARQL sta nel fato che permette all'utente di applicare query non ambigue siccome si basa su identificatori URI non ambigui.

3.3.7.1 Sintassi delle interrogazioni

La maggior parte delle query rispettano il formato a triple, chiamato anche "*basic graph pattern*". Tale formato è molto simile alle triple RDF, ma con la differenza che il soggetto e l'oggetto possono essere variabili. L'accoppiamento tra basic graph pattern e un sottografo di informazioni RDF succedere quando gli elementi del sottografo possono essere sostituti al posto delle variabili del pattern e il risultato sarebbe un grafo RDF equivalente al sottografo.

```
<http://example.org/book/book1>
<http://purl.org/dc/elements/1.1/title> "SPARQL Tutorial" .
```

Listing 3.1: Esempio con i dati nel formato RDF

```
SELECT ?title
WHERE
{
    <http://example.org/book/book1>
    <http://purl.org/dc/elements/1.1/title> ?title .
}
```

Listing 3.2: Esempio di una semplice query SPARQL

?title
"SPARQL Tutorial"

Tabella 3.1: Il risultato prodotto dall'applicazione della query SPARQL sull'insieme di dati sopraelencati

Il risultato di una query viene aggregato in una lista di soluzioni basandosi sulla corrispondenza del graph pattern con il sottografo dei dati in RDF. SPARQL offre la possibilità di utilizzare anche i literal RDF, i tipi numerici, i tipi di dati arbitrari nella struttura della query.

La struttura delle query non è uniforme perché SPARQL permette alle interrogazioni di assumere diverse forme. Le interrogazioni `SELECT` restituiscono un legame che i dati dichiarati nei documenti hanno con le variabili utilizzate nelle query. Le interrogazioni `CONSTRUCT` restituiscono un grafo RDF costruito basandosi su un modello che è stato utilizzato per generare triple RDF in base al matching del graph pattern con il dati.

SPARQL offre altri costrutti come `FILTER` per restringere il risultato della query solo a quelle tuple che soddisfano la condizione del filtro, `ORDER BY` per ordinare il risultato in base ad una regola, `DISTINCT` per ottenere un risultato con elementi con duplicati, etc.

3.3.8 SKOS

Lo SKOS (*Simple Knowledge Organization System*) è un insieme di linguaggi formali che fornisce un modello per rappresentare la struttura e il contenuto di schemi concettuali come glossari, tassonomie, classificazioni e altri tipi di vocabolari strutturati. Per molti ricercatori SKOS, viene considerato un vocabolario RDF per rappresentare la conoscenza semi formale. Il fato che SKOS si basa su RDF lo rende machine-readable.

Lo SKOS fornisce un linguaggio leggero, semplice ed intuitivo per i modelli concettuali permettendo di sviluppare e distribuire in rete nuovi KOS. Lo scopo di tale meccanismo e di fornire un percorso a basso costo per portare gli sistemi verso il web semantico. Lo SKOS svolge un ruolo da intermediario tra i formalismi ben organizzati dei linguaggi ontologici e quelli informali, scarsamente strutturati.

Il modello di dati SKOS è stato definito come una ontologia di tipo OWL Full e i dati SKOS sono stati rappresentanti come triple RDF da essere serializzate utilizzando qualsiasi sintassi. I concetti SKOS possono essere collegati ad altri concetti tramite le relazioni semantiche. Tali concetti possono essere raggruppati insieme in collezioni, a loro volta manipolate tramite l'ordinamento, etichettamento, etc.

Come nel caso di RDF, anche per il modello di dati SKOS, gli elementi più importanti della rappresentazione delle informazioni sono le classi e le proprietà. Le caratteristiche logiche delle classi e proprietà e loro dipendenze definiscono la struttura e l'integrità del modello. Tuttavia, lo SKOS non è un linguaggio formale per la rappresentazione della conoscenza, la quale viene espressa come un insieme di assiomi e fatti.

3.3.9 RIF

Il RIF (*Rule Interchange Format*) è uno standard ancora in fase di sviluppo, però un componente fondamentale dello stack del web semantico. Lo scopo principale di questo progetto è lo scambio delle regole tra diversi linguaggio di regole che già esistono. Nel 2005, fu istituito il *RIF Working Group* che aveva come obbiettivo di attirare coloro che creavano regole nei mercati commerciali sviluppando un formato di scambio tra i sistemi delle regole già esistenti.

La regola è un costrutto di tipo `IF-THEN` ed è probabilmente la nozione più semplice d'informatica. Questo costrutto descrive lo step da processare in caso una condizione viene verificata. I sistemi basati sulle regole utilizzano una nozione di predicato per rappresentare una certa informazione. Gli oggetti che sono argomenti della regola, vengono mantenuti connessi in base a tale predicato.

La progettazione e l'implementazione di motori di ricerca che inferiscono e processino le informazioni in base alle regole, è più facile siccome la conoscenza viene dichiarata tramite i predicati. Un sistema di regole è basato su un insieme di regole semantiche includendo quantificatori esistenziali, funzioni logiche disgiunzione, unione, negazione, etc. Nei giorni d'oggi tali sistemi vengono chiamati *sistemi esperti*, però il loro sviluppo data dagli anni '70.

RIF Group ha specificato tre dialetti standard del RIF:

- `Core`: dialetto che include un sottoinsieme comune di molte rule engines;
- `BLD`: estende il dialetto Core, aggiungendo caratteristiche che non sono direttamente disponibili;
- `PRD`: aggiunge la nozione della regola "forward-chaining".

3.3.10 SWRL

SWRL (*Semantic Web Rule Language*) è un linguaggio basato sulle regole per il mondo del web semantico ed è una combinazione dei sottolinguaggi dell'OWL con quelli del Rule Markup. Il linguaggio rimane una proposta da parte di W3C e non è stata approvata dal Consorzio.

La struttura delle regole si basa su due componenti principali: testa (consequent) e corpo (antecedent). La regola consiste in alcune condizioni espresse nell'antecedent e se le condizioni sono vere, allora dovrebbero essere vere anche le condizioni nella parte consequent. L'antecedent e il consequent consistono in zero o più atomi. Nel caso di un antecedent vuoto verrà considerato come sempre vero, mentre un consequent vuoto viene considerato come sempre falso e non sarà soddisfatto da nessuna interpretazione.

3.3.10.1 La sintassi delle regole

La sintassi del SWRL viene estratta da qualsiasi sintassi di documenti OWL per consentire un utilizzo più facile ed efficiente. Come discusso prima, qualsiasi ontologia OWL nella sua sintassi astratta contiene una sequenza di assiomi e di fatti. Tale ontologia potrebbe essere estesa definendo assiomi di regole come $axiom ::= rule$.

```
rule ::= 'Implies(' [URIreference ]{annotation }
antecedent consequent ')'
antecedent ::= 'Antecedent(' {atom }')'
consequent ::= 'Consequent(' {atom }')'
```

Listing 3.3: La struttura di un assioma di regole

```
atom ::= description '(' i-object ')'
    | dataRange '(' d-object ')'
    | individualvaluedPropertyID '(' i-object i-object ')'
    | datavaluedPropertyID '(' i-object d-object ')'
    | sameAs '(' i-object i-object ')'
    | differentFrom '(' i-object i-object ')'
    | builtIn '(' builtinID {d-object }')'
builtinID ::= URIreference
```

Listing 3.4: Composizione dell'atomo di una regola

Gli atomi in queste regole possono essere della forma:

- $C(x)$: valido se x è una istanza della classe descritta C;
- $P(x, y)$: valido se x è associato a y tramite la proprietà P;
- $sameAs(x, y)$: valido se x e y sono interpretatati come lo stesso individuo;
- $differentFrom(x, y)$: valido se x e y sono interpretatati come individui differenti;
- $builtin(r, x, \dots)$: valido se la relazione *builtin* di r è valida sulle interpretazione degli argomenti;

I valori di x e y potrebbero essere variabili, individui OWL o tipi di dati OWL.

3.3.11 Proof, Trust, Digital Firms

Il livello finale e la cima dello stack del web semantico consiste in un insieme di tecnologie che anche nei giorni d'oggi sono in fase si sperimentazione e di sviluppo senza avere una pubblicazione concretata e condivisa dalla maggior parte della comunità scientifica. Lo scopo delle tecnologie di questo livello è la realizzazione di sistemi in grado di formulare principi logici e permettere alle macchine di ragionare utilizzando tale principi.

Lo sviluppo dei sistemi logici permetterebbe di sfruttargli per provare e dimostrare la verità delle informazioni e le relazioni tra gli elementi. Il risultato delle dimostrazioni logiche deve restituire informazioni valide e vere. Qualsiasi utente che avrà la possibilità di accedere in rete, potrebbe scrivere istruzioni logiche e le machine potrebbero utilizzare tali riferimenti semantici per costruire le dimostrazioni. La costruzione delle dimostrazioni è molto difficile a livello operativo, ma potrebbero esser facilmente controllabili. Una volta ottenuto questo passaggio, il web potrà essere considerato come un meccanismo in grado di processare informazioni.

In una visione futura potremo vedere processori euristici che utilizzando le regole e gli istruzioni forniti dagli utenti, riusciranno ad aggregare autonomamente delle conclusioni. Dalla comunità scientifica viene chiamato *procedimento euristico* un approccio alla soluzione dei problemi che si basa sulla intuizione e allo stato temporaneo delle istanze al fine di generare nuova

conoscenza senza seguire un percorso chiaro e strutturato.

La firma digitale nel livello finale serve per garantire autenticità e sicurezza durante l'esecuzione dei processi. L'autenticità delle asserzioni viene garantita tramite un sistema di crittografia, dove la responsabilità del materiale pubblico cade sull'ente, sull'individuo o sull'organizzazione che lo ha condiviso. La firma digitale sarà univocamente associata ai documenti che si trovano nella rete, assicurando l'utente sulla provenienza di tale risorsa.

Il termine *"Web of Trust"* viene introdotto per rappresentare un modo della rete concentrato sulla fiducia, riservatezza e sicurezza. Condividendo la fiducia che due utenti avranno tra di loro, permetterà lo sviluppo di tale mondo associando a tutte queste relazioni un grado di fiducia. Alla fine sarà il livello di *User interface* che permetterà a tutti gli utenti di usufruire le applicazioni del web semantico.

3.4 Knowledge graph

Dopo avere parlato dei componenti principali dello stack del web semantico, adesso verranno discussi gli argomenti più importanti che riguardano i knowledge graph.

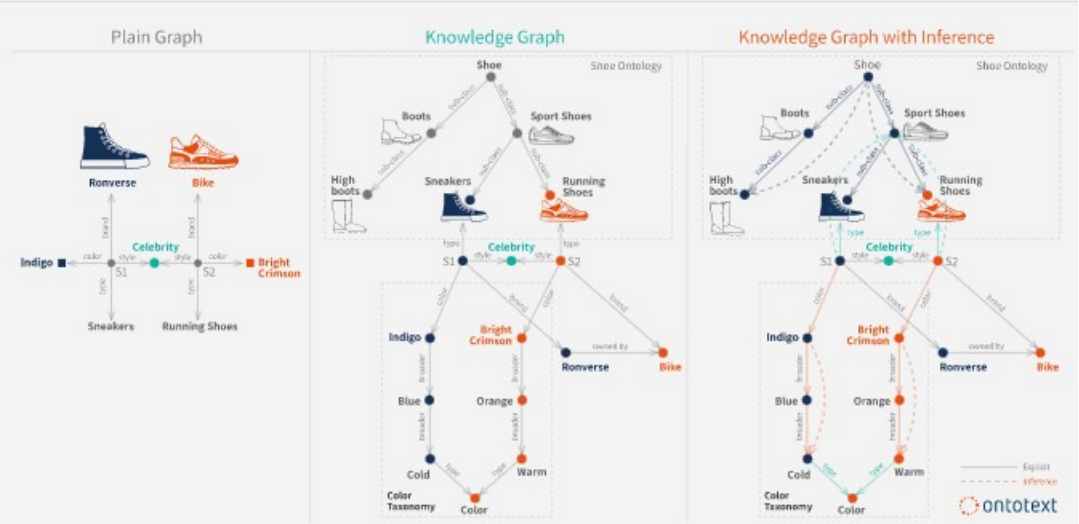

Figura 3.5: Esempio di una grafo di conoscenza

3.4.1 Cos'è un grafo di conoscenza?

Un knowledge graph oppure un grafo di conoscenza rappresenta un insieme enorme di entità che sono connesse tra di loro. Tutte queste relazioni formano una rete di entità intraconnesse e interconnesse. L'informazione viene memorizzata in database specifici per la gestione dei grafi di conoscenza e per la visualizzazione di tale informazione.

Come in tutti i casi, quando parliamo di un grafo, parliamo di una struttura composta da nodi e archi. Le etichette vengono utilizzate per facilitare la gestione e la leggibilità del grafo, aggiungendo informazioni e strumenti di ricerca per i nodi e gli archi. Gli archi permettono di definire relazioni tra diversi nodi che compongono il grafo.

Nella comunità scientifica c'è molta discussione sulle differenze tra le ontologie e i grafi di conoscenza, dove alcuni gli considerano la stessa cosa. La maggior parte condivide il fato che le ontologie servono per creare un formato di rappresentazione delle entità del grafo. Secondo *OntoText*, un grafo di conoscenza rappresenta una collezione di descrizioni di entità interconnesse, degli oggetti e eventi del mondo reale, e degli concetti astratti dove:

- le descrizioni hanno una semantica formale che permette agli utenti e ai processori di processarle in un modo efficiente e non ambiguo;
- le descrizioni delle entità contribuiscono a formare una rete di entità, dove ogni entità rappresenta una parte della descrizione associata a tale entità e fornisce contesto per l'interpretazione.

I knowledge graph combinano un insieme di caratteristiche di diversi paradigmi per la gestione dei dati tra i quali:

- *base di dati*: perché i dati possono ricercati, aggiornati e manipolati tramite interrogazioni strutturate;
- *grafi*: perché i dati possono essere analizzati e visualizzati come qualsiasi formato di dati in rete;
- *base di conoscenza*: perché portano una semantica formale, utile per interpretare i dati e dedurre nuova conoscenza.

Il knowledge graph rappresentato tramite RDF, offre il miglior framework per l'integrazione dei dati, la loro unificazione e riutilizzo. I grafi di conoscenza trovano un ampio dominio applicativo siccome riescono a combinare:

- *l'espressività*: Il core dei knowledge graph sono gli standard del web semantico, RDF e OWL, che vengono utilizzati per una rappresentazione fluida e semplice dei diversi tipi di dati e contesti. Gli schemi di dati, le tassonomie, i vocabolari e i metadati servono a modellare le informazioni in modo efficacie;

- *la prestazione*: Utilizzando il knowledge graph è stato provato nella pratica che consente una gestione efficiente di grafi su una grande quantità di fatti e proprietà;

- *l'interoperabilità*: Avendo una varietà di specifiche per la serializzazione dei dati, l'accesso (protocollo SPARQL per gli endpoint), la gestione (SPARQL Graph Store) e la federazione è possibile l'integrazione e la sincronizzazione con diversi sorgenti d'informazione. L'uso di identificatori univoci globali facilita l'integrazione e la pubblicazione dei dati;

- *la standardizzazione*: Tutte le caratteristiche sopra elencate sono state standardizzate dalla comunità scientifica, in questo caso W3C, per garantire che i requisiti siano sempre soddisfati.

Non tutti i grafi RDF e knowledge base sono grafi di conoscenza. Un elemento importantissimo dei knowledge graph è che le descrizioni delle entità devono essere interconnesse con una l'altra. La definizione di una entità include una altra entità e tale connessione forma il grafo di conoscenza. Le basi di conoscenza che non rispettano una struttura semantica formale, non possono essere considerate come grafi di conoscenza.

Nella rete sono miglia di enormi knowledge graph pubblici e condivisi, tra i quali DBPedia, Geonames, Wordnet, etc.

3.4.1.1 DBPedia

Questo progetto sfrutta la struttura inerente agli infobox di Wikipedia per creare un enorme dataset e un'ontologia che ha una copertura enciclopedica di entità come persone, luoghi, film, libri, organizzazioni, specie, malattie, ecc. Questo set di dati è al centro del movimento Open Linked Data. Questo progetto è stato importantissimo per le organizzazioni in modo da progettare i propri grafi di conoscenza interna con milioni di entità. DBpedia consente agli utenti di interrogare semanticamente le relazioni e le proprietà delle risorse di Wikipedia, inclusi i collegamenti ad altri set di dati correlati.

DBpedia estrae informazioni dalle pagine di Wikipedia, consentendo agli utenti di trovare risposte a domande in cui le informazioni sono distribuite su più articoli di Wikipedia. L'utente sarebbe in grado di accedere alle informazioni utilizzando un linguaggio di query simile a SQL per RDF, SPARQL.

```
PREFIX dbprop: <http://dbpedia.org/ontology/>
PREFIX db: <http://dbpedia.org/resource/>
SELECT ?who, ?WORK, ?genre
WHERE {
    db:Tokyo_Mew_Mew dbprop:author ?who .
    ?WORK dbprop:author ?who .
    OPTIONAL { ?WORK dbprop:genre ?genre } .
}
```

Listing 3.5: Esempio di una query SPARQL interrogando il dataset sulla serie di manga giapponese *Tokyo Mew Mew* per trovare i generi di altre opere scritte dalla sua illustratrice Mia Ikumi.

DBpedia ha una vasta gamma di entità che coprono diverse aree della conoscenza. Questo lo rende uno strumento naturale e importantissimo per la connessione di fonti di dati, dove i fonti di dati esterni potrebbero collegarsi ai suoi concetti. Il set di dati DBpedia è interconnesso a livello RDF con vari altri set di dati Open Data sul Web, consentendo alle applicazioni di arricchire le informazioni di DBpedia.

3.4.1.2 GeoNames

Fondato nel 2005, il progetto GeoNames è un database geografico modificabile e accessibile dagli utenti tramite vari servizi web. Sotto Creative Commons, gli utenti del set di dati Geonames hanno accesso a 25 milioni di entità geografiche. Tutte le informazioni sono classificate e raggruppate in nove classi e ulteriormente sotto-categorizzati in base a 645 codici.

Ogni entità di GeoNames è rappresentata come una risorsa web identificata da un URI in modo da fornire accesso attraverso la navigazione del contenuto, sia alla pagina wiki HTML, sia a una descrizione RDF della caratteristica, utilizzando elementi dell'ontologia GeoNames. Questa ontologia descrive le proprietà delle caratteristiche di GeoNames utilizzando il linguaggio dell'ontologia Web, le classi di caratteristiche e i codici descritti nel linguaggio SKOS. Attraverso l'URL degli articoli di Wikipedia collegati nelle descrizioni RDF, i dati di GeoNames sono collegati ai dati di DBpedia e ad altri dati collegati a RDF.

3.4.1.3 WordNet

Il progetto Wordnet è uno dei database lessicali più conosciuti per le lingue, che fornisce definizioni e sinonimi, spesso utilizzato per migliorare le prestazioni della PNL e delle applicazioni di ricerca. WordNet è un database lessicale di relazioni semantiche tra parole in più di 200 lingue, collegando le parole in relazioni semantiche inclusi sinonimi e iponimi. I sinonimi sono raggruppati in synset con brevi definizioni ed esempi di utilizzo. WordNet può quindi essere visto come una combinazione ed estensione di un dizionario e di un thesaurus.

Sebbene sia accessibile agli utenti umani tramite un browser web, il suo uso principale è nell'analisi automatica del testo e nelle applicazioni di intelligenza artificiale. WordNet è stato creato per la prima volta in lingua inglese e gli strumenti software sono stati rilasciati con una licenza BSD.

3.4.2 Come funziona un grafo di conoscenza?

Il knowledge graph è composto da un insieme di dataset provenienti da diversi fonti o sorgetti che utilizzano diversi formati e strutture per rappresentare le informazioni memorizzate. Gli schemi dei dati, gli identificatori e il contesto vengono mischiati insieme per fornire una struttura comune ai dati diversi.

I knowledge graph che vengono alimentati tramite machine learning effettuano un processo di arricchimento semantico per fornire una visone completa dei nodi, archi ed etichette. Quando una sorgente di dati fornisce informazioni, il processo consente l'identificazione dei singoli oggetti e la comprensione delle loro relazioni. La conoscenza generata e inferita tramite questo processo viene confrontata con altri dataset. Gli sistemi che attendono risposte delle interrogazioni fate, riceveranno il risultato appena il grafo della conoscenza è completo e consente le query. L'integrazione dei dati e dei risultati potrebbe supportare la creazione di nuova conoscenza definendo relazioni tra dati che prima sarebbe stato impossibile definire.

3.4.3 Definizione di un grafo di conoscenza

Come già menzionato prima, un grafo della conoscenza è un grafo diretto etichettato costituito da nodi, archi ed etichette. Qualsiasi concetto potrebbe essere rappresentato come un nodo, ad esempio persone, aziende, eventi,

luoghi, etc. Un arco effettua il collegamento tra una coppia di nodi e cattura la relazione d'interesse tra di loro, mentre le etichette catturano il significato della relazione.

Secondo la definizione formale, dato in un insieme di nodi N e un insieme di etichette L, un knowledge graph è un sottoinsieme del prodotto $N \times L \times N$. Ciascun elemento di questo insieme di dati viene rappresentato come una tripla.

Figura 3.6: Visualizzazione di due nodi connessi del grafo

Un grafo come quello sopraindicato viene chiamato un grafo di dati. Nel caso i nodi rappresentano classe di oggetti e gli archi catturano le relazioni dei sottoclassi, tale grafo viene chiamata tassonomia.

La navigazione sul grafo porta lo vantaggio di ridurre e risparmiare molti calcoli, che inizialmente sembrerebbero pesanti computazionalmente. In un knowledge graph della conoscenza che modella le amicizie, se vogliamo calcolare gli amici di un amico di una persona A, basterebbe navigare il grafo da A a tutti i nodi B collegati ad esso tramite una relazione etichettata *amico* e continuare la navigazione ricorsiva su tutti i nodi C, collegati a B tramite la stessa relazione.

Un *path* (percorso) in un grafo di conoscenza G è un insieme di nodi $(v_1, v_2, ..., v_n)$ dove per ogni $i \in N$ con $1 \leq i \leq n$, esiste un arco da v_1 a $v_i + 1$. Un percorso con nodi non ripetuti e distinti viene chiamato percorso semplice, mentre se il primo e l'ultimo nodo corrispondo viene chiamato percorso ciclico. Durante la navigare e l'attraversamento dei nodi è possibile definire numerose proprietà aggiuntive (ad esempio componenti connessi, fortemente connessi, etc).

3.4.4 Use case di knowledge graph

I grafi della conoscenza sono un meccanismo avanzato ed innovativo e negli ultimi anni ha trovato numerose applicazione nel mondo della ricerca e in

diverse industrie. In seguito verranno spiegate le applicazione dei knowledge graph che hanno portato un aumento della sua popolarità recentemente.

3.4.4.1 Organizzare la conoscenza nella rete

L'utilizzo principale dei gradi della conoscenza è l'organizzazione delle informazioni in rete in modo da essere facilmente accessibili, manipolabili ed interrogabili. Wikidata è un esempio di come potrebbe essere strutturata la conoscenza in rete. Tale progetto offre un archivio centrale per i dati strutturati che si trovano nel Wikipedia.

Un esempio dell'utilità del Wikidata viene mostrato da una semplice ricerca su Wikipedia della città di Winterthur in Svizzera. Nella pagina Wiki di questa città vengono elencate alcune città gemelle che si trovano in Europa. Anche se non era specificato nella pagina di Winterthur, nella pagina Wiki di Ontario in California viene specificato la città Winterthur come una città sorella. Per gli utenti le relazioni sorella e gemella vengono riconosciute come uguali, ma per il sistema no. Al contrasto di Wikipedia, la rappresentazione di Wikidata mappa queste due relazioni come simili e siccome l'inferenza della informazioni è automatica, Ontario e Winterthur sono città gemelle e specificate nelle loro istanze in Wikidata.

Il grafo della conoscenza di Wikidata è il più grande grafo della conoscenza disponibile oggi. Molti dati in Wikidata possono provenire da informazioni estratte automaticamente, ma devono essere facilmente compresi e verificati secondo le politiche editoriali di Wikidata. Il progetto di Wikidata si focalizza esplicitamente sulla definizione semantica di diversi nomi di relazioni utilizzando una varietà di vocabolari. In questo modo Wikidata migliora l'esperienza di navigazione e di ricerca per gli utenti sul web.

3.4.4.2 Integrazione dei dati nelle industrie

L'integrazione dei dati è un processo di combinazione di dati provenienti da vari fonti, fornendo all'utente una visone strutturata, unificata e chiara dei dati. In diverse industrie, i dati vengono gestiti e memorizzati utilizzando diversi meccanismi di persistenza, aumentando il rischio e i costi. La progettazione e l'implementazione di uno schema globale condiviso da tutti per la modellazione dei dati è un processo difficile. Gli ricercatori si sono focalizzati sulla risoluzione del problema seguendo un approccio diverso, conversione dei basi di dati in uno schema generico di triple. L'accumulazione di tutte le

triple permetterebbe la creazione di un grafo di conoscenza.

3.4.4.3 Intelligenza artificiale

Sin dall'inizio, i knowledge graph sono stati utilizzati come rappresentazione dell'intelligenza artificiale. Negli ultimi anni i grafi della conoscenza sono stati rappresentati tramite i grafi concettuali, le logiche descrittive e linguaggi di regole.

Le due sfide principali dell'intelligenza artificiale sono la rappresentazione della conoscenza e l'acquisizione. sua catturazione tramite la rappresentazione scelta. Il processo dell'acquisizione consiste nella catturazione tramite la rappresentazione scelta in modo semplice e scalabile. Le tecniche utilizzate nei giorni d'oggi si basano sull'apprendimento induttivo e la generazione automatica dell'apprendimento.

I knowledge graph vengono utilizzati per la memorizzare e la visualizzazione delle conoscenze automaticamente apprese. La visualizzazione delle informazioni servirebbe per migliorare il processo dell'apprendimento automatico.

3.4.4.4 Input e output di machine learning

3.4.4.4.1 Input

I modelli di machine learning si basano maggiormente su un input numerico e richiede che qualsiasi struttura di dato potrebbe essere convertita in una rappresentazione numerica. La rappresentazione numerica applicata sull'input testuale, chiamata word embedding, migliora notevolmente le prestazioni dei task di *natual language processing* includendo l'estrazione delle entità e delle classi, il parsing, etc.

Word embedding viene utilizzato nel mondo dell'web semantico per autocompletare le interrogazioni di ricerca, prevedendo le parole che probabilmente seguiranno la query parziale che l'utente avrebbe già digitato. Calcolando il numero delle occorrenze di una parola nel testo, è possibile rappresentare il testo come un grafo di conoscenza dove ogni parola è un nodo e tra due parole consecutive si trova un arco etichettato. L'obbiettivo è di rappresentare ogni nodo del grafo tramite un vettore, in modo che le similarità tra i nodi potranno essere calcolate come le differenze dei loro vettori. Questo processo

viene chiamato come *graph embeddings*.

In modo da calcolare i graph embeddings per ogni nodo del grafo, viene definito un metodo di encoding che consiste in una funzione che calcola le similarità tra i nodi e poi applica un algoritmo di ottimizzazione. Il processo di encoding di un nodo viene chiamato come *node embedding*.

3.4.4.4.2 Output

I knowledge graph vengono utilizzati come rappresentazione dell'output target per l'elaborazione del linguaggio naturale e gli algoritmi di visione artificiale. L'estrazione di entità e l'estrazione di relazioni dal testo sono due obbiettivi fondamentali nell'elaborazione del linguaggio naturale. I grafi della conoscenza forniscono un mezzo naturale per estrarre informazioni correlate da più parti del testo.

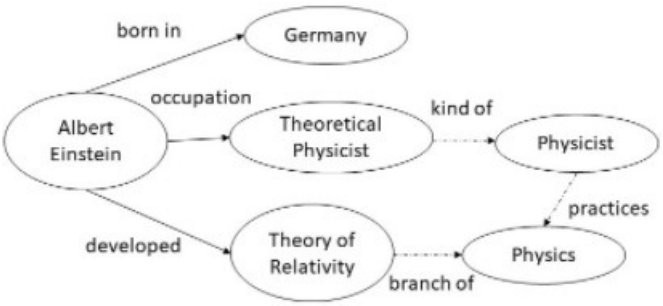

Figura 3.7: *Albert Einstein was a German-born theoretical physicist who developed the theory of relativity.* Passando come input l'informazione precedente, possiamo estrarre i concetti della frase e rappresentarle semanticamente costruendo un grafo.

Lo scopo della visone artificiale è la comprensione di un'immagine creando un modello in grado di rilevare e descrivere gli oggetti presenti specificando le relazioni.

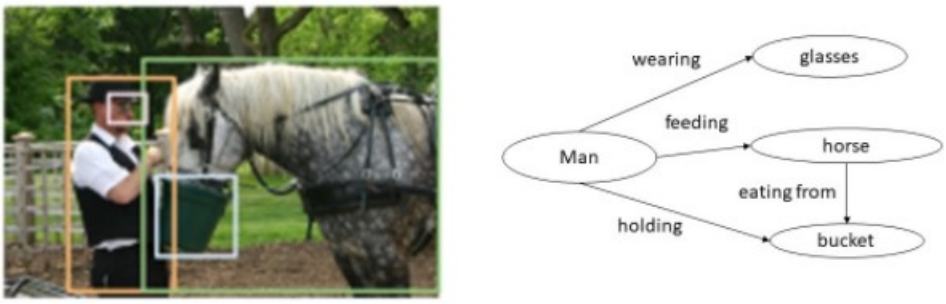

Figura 3.8: Esempio di un sistema di comprensione delle immagini che produce un grafo della conoscenza, dove i nodi sono gli output di un rilevatore di oggetti e gli archi le relazioni tra gli oggetti.

3.5 Linked Data

Il progetto Linked Data consiste in un insieme di principi di progettazione per la condivisione di dati machine-readable sul Web. La fusione di tale progetto con Open Data si chiama LOD (*Linked Open Data*). LOD è in grado di gestire enormi set di dati provenienti da fonti diversi e collegarli agli Open Data, il che aumenta la scoperta della conoscenza e fornisce strumenti per un analisi efficiente basata sui dati.

Il Web semantico ha lo scopo di offrire meccanismi per la creazione di collegamenti tra set di dati comprensibili non solo per gli esseri umani, ma anche per le macchine. I Linked Data forniscono le migliori pratiche per rendere possibili questi collegamenti. In altre parole, i Linked Data sono un insieme di principi di progettazione per la condivisione di dati interconnessi leggibili dalla macchina sul Web.

Quando parliamo di Web of Data, è fondamentale avere a disposizione un enorme quantità di dati in un formato standard, raggiungibile e gestibile dagli strumenti del Web semantico. Le relazioni tra i dati dovrebbero essere rese disponibili per creare un Web of Data, diversamente da una semplice raccolta di set di dati. Questa raccolta di set di dati interconnessi sul Web può anche essere indicata come Linked Data.

Per ottenere e creare Linked Data, dovrebbero essere disponibili tecnologie per un formato comune come RDF, in modo da effettuare la conversione e l'accesso al volo a database esistenti (relazionali, XML, HTML, ecc.). La configurazione degli endpoint delle interrogazioni è importantissima per poter accedere a tali dati velocemente e in modo più conveniente. W3C fornisce

una gamma di tecnologie (RDF, GRDDL, POWDER, RDFa, il prossimo R2RML, RIF, SPARQL) per accedere ai dati.

Capitolo 4

Analisi

4.1 Scopo del progetto

Pathadora ha come obbiettivo finale la generazione dei learning paths dinamici ed efficienti basandosi sugli parametri iniziali specificati dall'utente e dalle relazioni generate tra le istanze del sistema.

4.1.1 Modellazione ontologica

Pathadora deve modellare tramite gli strumenti del web semantico diversi domini tra i quali:

- **l'istituzione accademica**: l'organizzazione dell'università in scuole, dipartimenti, facoltà e corsi. La modellazione dell'istituzione tramite una gerarchia permetterebbe di generare una sequenza si dipartimenti, facoltà e corsi tendono in considerazione le associazioni tra tali componenti. Le informazioni dichiarati dallo studente, permettono al sistema di iniziare la ricerca in cima della gerarchia, ovvero le scuole per poi scendere step-by-step fino ai corsi.
- **le informazioni dello studente**: tale informazioni possono essere:
  - generali: informazioni che specificano i dati anagrafici, le passioni, le lingue che parla, etc;
  - personali: informazioni sulle disabilità, sugli obbiettivi, sul metodo di apprendimento, etc;
  - accademici: informazioni che enfatizzano gli obbiettivi raggiunti durante il suo percorso accademico.

- le risorse didattiche: modellazione dei materiali didattici focalizzando sulle informazioni che riguardano l'accessibilità della risorsa e il topic del corso associato.

- l'accessibilità delle risorse; rappresenta l'accessibilità come un macro concetto e si focalizza sulla modellazione di informazioni sulla accessibilità che possono essere utilizzati per usufruire diversi tipi di risorse con diverse caratteristiche. L'ontologia mappa metodi di accessibilità di una risorsa e metodi di conversione di una risorsa in modo da offrire un altra soluzione di accessibilità.

4.1.2 Rule and query base model

La logica del Pathadora si dovrebbe basare sulle regole in modo da creare un sistema *rule-based*. Tale sistema sarebbe in grado di memorizzare e manipolare la conoscenza e interpretare le informazioni in diversi modi. Pathadora utilizzerà questo sistema per fare deduzioni o scelte durante la generazione dei learning path basandosi sulle informazioni inferite dall'applicazione delle regole.

4.1.3 Pathadora engine

Pathadora-engine sarà una engine sempre in esecuzione che servirà le richieste ricevute da parte dell'utente. Tale engine offrirà i meccanismi necessari per interrogare la knowledge dell'ontologia.

4.2 Analisi dei requisiti

In questa sezione verranno spiegati i requisiti funzionali e non funzionali del progetto.

4.2.1 Requisiti funzionali

Il sistema finale del progetto dovrà soddisfare i seguenti requisiti funzionali:

- le ontologie:
  - pathadora-ontology: L'ontologia che modellerà tutto il dominio del sistema, racchiudendo i concetti dell'organizzazione dell'università, dei studenti, delle risorse e la loro accessibilità;

  - accesible: Sincronizzare l'ontologia dell'accessibilità con l'ontologia del sistema, pathadora-ontology;
  - lom: Sincronizzare l'ontologia che modella i metadati per le risorse didattiche con l'ontologia del sistema.

- gestire le richieste: L'utente potrà interagire con il sistema mandando delle richieste dal interfaccia web e il sistema dovrà essere in grado di accettare e servire queste richieste;

- manipolazione delle ontologie: Il sistema dovrà offrire meccanismi di manipolazione delle ontologie con l'inserimento di nuove istanze oppure i loro aggiornamento;

- manipolazione delle regole: L'utente deve essere in grado di personalizzare e parametrizzare le regole da applicare sulle ontologie presenti;

- manipolare le interrogazioni: L'utente deve essere in grado di manipolare le interrogazioni che verranno fatte al sistema di persistenza;

- Pathadora Recommender: Progettazione dell'engine che racchiude tutti i componenti del sistema.

4.2.2 Requisiti non funzionali

Il sistema finale del progetto dovrà soddisfare i seguenti requisiti non funzionali:

- la prestazione: Il sistema deve essere in grado di gestire la richiesta e produrre una risposta nel minor tempo possibile, sfruttando al massimo le risorse computazionali messe a disposizione;

- la scalabilità: A livello di codice il sistema deve essere facilmente estendibile se sarebbe necessario l'introduzione di nuovi componenti;

- la disponibilità: Pathadora deve essere sempre in ascolto e servire le richieste, definendo tecniche di fault-tolerance;

- la consistenza: Il sistema deve essere in grado di mantenere la consistenza nel caso di inserimento o aggiornamento della knowledge, verificando lo stato degli strumenti utilizzati per la persistenza.

Maybe Glossario dei termini here

Capitolo 5

Progettazione

Pathadora è un sistema di raccomandazione di facoltà e corsi da seguire da uno studente in base alle caratteristiche specificate. Il sistema oltre alle facoltà e corsi, genera una sequenza di risorse didattiche da consigliare per un determinato corso scelto. Il learning path delle risorse dipenderà principalmente dal livello di accessibilità delle risorse e dalle eventuali disabilità dichiarate dello studente.

Pathadora gestisce le caratteristiche che riguardano gli studenti, le facoltà, i corsi e le risorse tramite la modellazione ontologica utilizzando strumenti e meccanismi del web semantico. Sfruttando tali strumenti, Pathadora è in grado di produrre un risultato dinamico e aggiornabile con nuove entità o caratteristiche che vengono aggiunte nel sistema. La generazione di relazioni tra le entità del sistema offre la possibilità di ottenere percorsi non uniformi e uguali per tutti, aumentando l'efficacia del sistema.

Gli utenti possono interagire con il sistema, tramite l'interfaccia web che Pathadora offre in modo da catturare e servire le richieste. *Pathadora-engine* si occupa della gestione degli strumenti ontologici del sistema e la produzione dei learning paths e la verifica di tali risultati.

5.1 Architettura

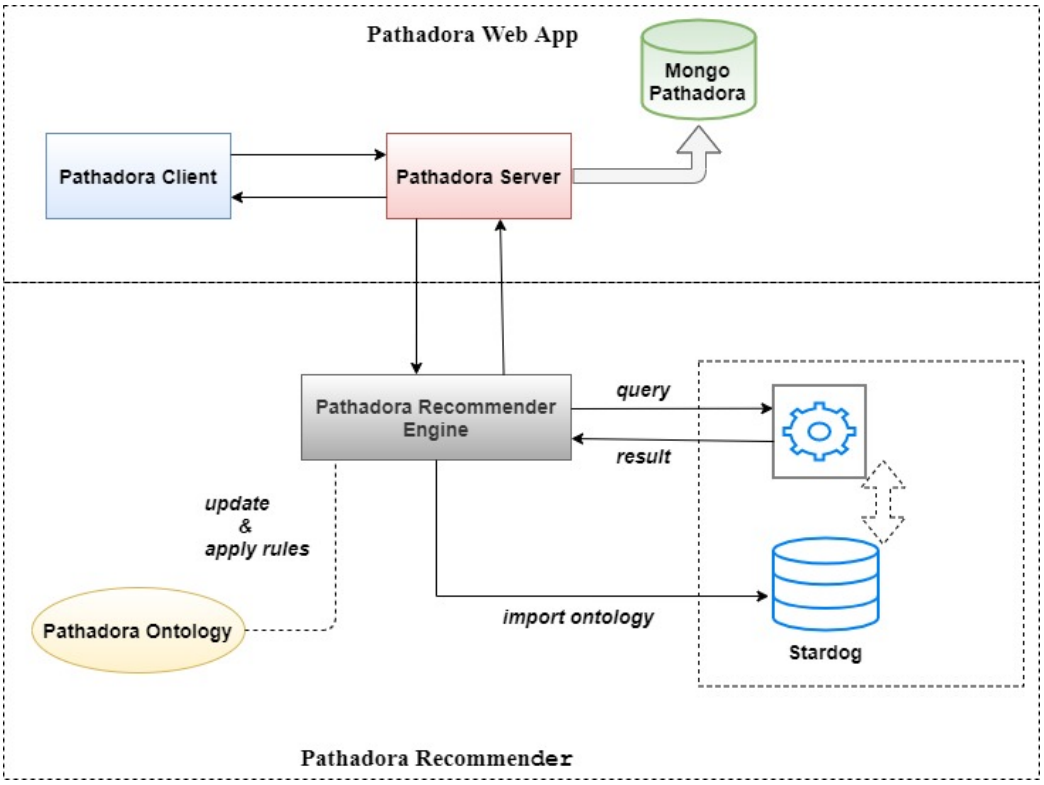

Figura 5.1: Architettura di Pathadora

Il sistema Pathadora consiste in due macro concetti principali, Pathadora Web App e Pathadora Recommender. Pathadora Web App offre una interfaccia web permettendo all'utente di interagire con il sistema. L'applicazione web è composta da tre blocchi importanti: il client, il server e il database.

Pathadora Ontology è un componente del sistema che ha come obbiettivo di gestire e manipolare l'ontologia utilizzata per rappresentare il dominio del progetto. Tale componente interagisce con l'engine in caso diversi aggiornamenti o interrogazioni vengono richieste.

Pathadora Rule Model ha il compito di parametrizzare le regole ed effettuare la loro applicazione, notificando l'engine sull'andamento di tale operazione. Questa attività viene eseguita con l'inizializzazione del Pathadora Recommender e in quel momento il sistema è pronto ad ricevere richieste di interrogazione.

Per connettere le informazioni a livello di elaborazione e non di archiviazione, viene utilizzata una piattaforma esterna, Stardog. Tale componente ha l'obbligo di ospitare il knowledge graph dell'ontologia e preparare le risposte delle richieste fatte.

Pathadora Recommender racchiude tutta la modellazione e la logica dietro la generazione dei learning path. Questa engine rimane sempre in esecuzione ad intercettare le richieste e inoltrale al componente che ha il compito di gestirle.

5.2 Progettazione del knowledge graph

Per poter organizzare ed esprimere in modo logico le relazioni semantiche tra gli elementi del dominio, è stato progettato un knowledge graph. Tramite il knowledge graph vengono aggregate e dedotte le relazioni tra i concetti inseriti nella grafo. L'informazione verrà memorizzato in un database a grafo e visualizzato come una struttura a grafi.

Il grafo è composto da nodi e archi, dove i nodi modellano un elemento e l'arco modella la relazione di tale elemento con un altro. Qualsiasi componente del sistema potrebbe essere un nodo. La modellazione dell'ontologia Pathadora definisce un formato di rappresentatine delle istanze del knowledge graph.

La conoscenza viene rappresentata da un knowledge framework multidimensionale. Ogni concetto del dominio è stato modellato tramite le classi e le sottoclassi per le informazioni più dettagliate. La modellazione multidimensionale permette di aggregare le informazioni a diversi livelli di dettaglio.

5.2.1 L'università

L'università come istituzione è stata modellata tramite una gerarchia a tre livelli. In cima di questa gerarchia sono le scuole, per poi scendere fino ai corsi. Utilizzando la gerarchica e una organizzazione a livelli, è più facile mappare le preferenze di uno studente con determinato learning path. In questo modo, se si ottene una scuola che soddisfa alcune caratteristiche dello studente, allora si assicura che scendendo sono questo ramo, i risultati ottenuti continuano a soddisfare le sue preferenze.

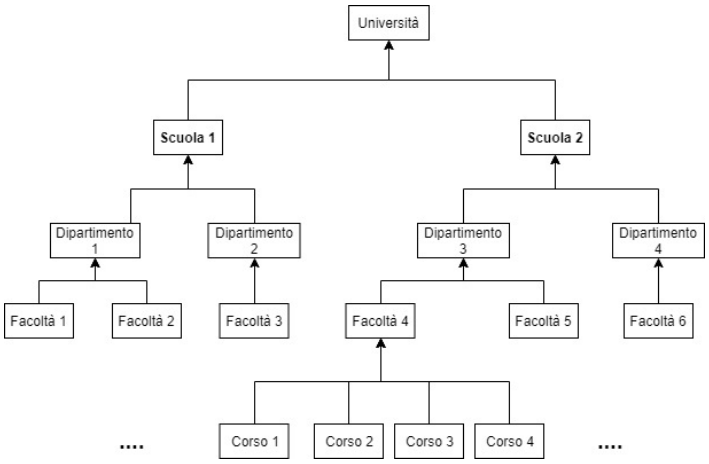

Figura 5.2: Organizzazione dell'università

5.2.2 Lo studente

La modellazione delle informazioni dello studente si basa su una knowledge graph. Per rappresentare le relazioni tra i diversi concetti in base al livello di dettaglio, le informazioni dell'utente sono state divise in tre classi: General, Personal e Accademic.

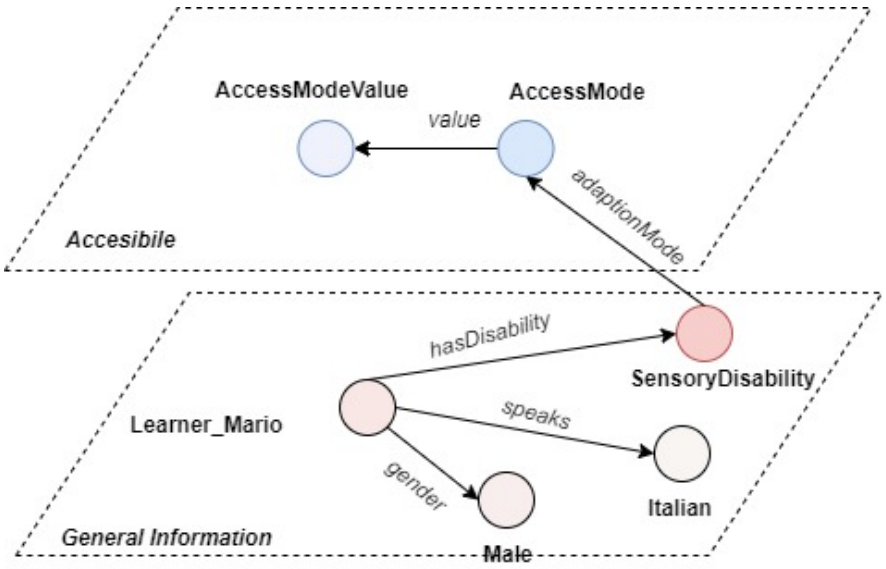

Figura 5.3: Knowledge graph sulle disabilità dello studente

5.3 Gestione delle iterazioni e richieste

Il sistema dovrà gestire le richieste che arrivano da parte del cliente e istanziare il componente adeguato a computare la risposta.

5.3.1 Inserimento

La prima tipologia di richiesta da gestire è l'inserimento di nuove istanze nell'ontologia. In questa richiesta dovranno essere specificate tutte le informazioni che serviranno all'engine per aggiornare la knowledge graph. La richiesta potrebbe dichiarare relazioni tra elementi che non sono presenti e in tal caso, l'engine prima creerà questi elementi per poi definire la relazione tra loro. Nel caso della richiesta di inserimento tutti gli elementi verranno aggiunti nell'ontologia, indipendentemente se esistono, se rispettano il formato giuso oppure se le relazioni sono semanticamente correte.

5.3.2 Generazione di facoltà

La secondo tipologia di richiesta consiste nella generazione del percorso delle facoltà. Lo studente nella richiesta di generazione dovrà specificare il tipo di diploma che vuole conferire come parametro per selezionare le facoltà da consigliare. L'aggregazione delle facoltà verrà fatta applicando una regola, parametrizzata con il tipo di diploma da conferire. Oltre al tipo di diploma, il reasoner utilizzerà tutte le informazioni dichiarate dallo studente durante la registrazione per aggregare una lista di facoltà.

5.3.3 Generazione di corsi

La generazione dei corsi, viene fatta sempre dopo la generazione delle facoltà. L'utente dovrà interagire con il sistema per fare delle scelte che serviranno per generare il risultato del prossimo step. La generazione del percorso di studio, step-by-step, necessita che lo studente scegliesse la facoltà preferita dall'elenco delle facoltà raccomandate. Fatta tale scelta, lo studente dovrà specificare l'anno dei corsi che vorrebbe ricevere come raccomandazione. Il sistema Pathadora, una volta ricevuto tale scelta, inizializzerà il knowledge graph database e lo interrogherà parametrizzando la query con le informazioni dello studente.

5.3.4 Generazione di risorse

L'ultima richiesta da gestire consiste nella raccomandazione delle risorse, focalizzandosi sulla accessibilità delle risorse e sulle disabilità dello studente. Per ogni risorsa verranno definite informazioni che rappresentano il livello di accessibilità utilizzando coefficienti che specificano tale livello. Lo studente durante la registrazione definirà le disabilità, specificando un livello a scala numerica. Pathadora-engine dovrebbe mappare la correlazione tra questi coefficienti di disabilità e di accessibilità per aggregare una lista di risorse da raccomandare allo studente. Oltre alla accessibilità, la raccomandazione delle risorse si focalizzerà sulle caratteristiche dello studente come stile di apprendimento, obbiettivi, etc. Anche in questo caso, la richiesta della generazione delle risorse, dovrà essere fatta dopo la generazione dei corsi, in modo che lo studente scegliesse il corso per il quale vorrebbe ricevere una lista di risorse.

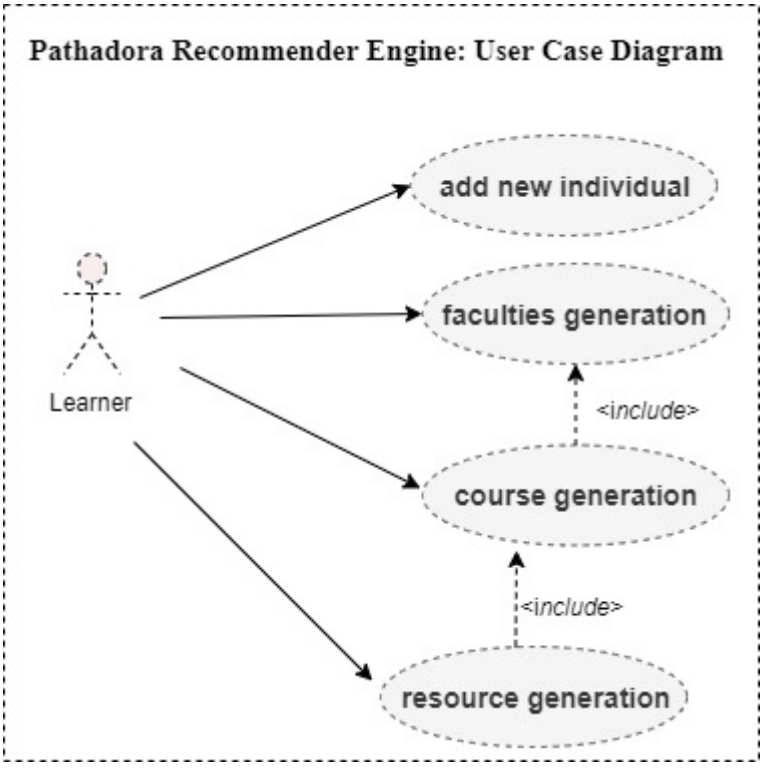

Figura 5.4: Pathadora Recommender Engine: User case diagram

Capitolo 6

Implementazione

6.1 Fase di sviluppo

6.2 Tecnologie

6.2.1 RDF

6.2.2 OWL

6.2.3 SWRL

6.2.4 Web Scrapping

6.3 Tools

6.3.1 SWRL & OWL API

6.3.2 Protege

6.3.3 Stardog

6.3.3.1 Docker

Capitolo 7

Risultati

Capitolo 8

Conclusioni

Conclusioni to be completed

References

[1] A learning path recommendation model based on a multidimensional knowledge graph framework for e-learning, https://www.sciencedirect.com/science/article/pii/S095070512030085X?via%3Dihub

[2] <Name of the reference here>, <urlhere>

Elenco delle figure