ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA
CAMPUS DI CESENA
DIPARTIMENTO DI
INGEGNERIA DELL'ENERGIA ELETTRICA E
DELL'INFORMAZIONE
"GUGLIELMO MARCONI"

CORSO DI LAUREA IN INGEGNERIA BIOMEDICA

TITOLO DELL'ELABORATO

Le ontologie: la rappresentazione semantica dell'informazione e la loro applicazione in campo biomedico

Elaborato in

CALCOLATORI ELETTRONICI

Relatore

Prof. Luca Roffia

Presentata da

Elisa Riforgiato

Anno Accademico 2020/2021

Sommario

Abstract

Lo scopo di questa tesi è analizzare il linguaggio semantico delle ontologie e valutare la loro applicazione pratica in ambito biomedico. La sua realizzazione si basa su una selezione di fonti bibliografiche di vario genere (libri, articoli, report, webinar) che tiene conto anche di risultati pubblicati nell'anno in corso. Tra i settori di ricerca biomedici che attualmente muovono interesse verso il modo di rappresentare l'informazione delle ontologie, questa tesi si è focalizzata su due: il *machine learning* e la prioritizzazione genetica. Gli studi analizzati risalgono tutti agli ultimi tre anni (2019-2021) e sebbene mostrino risultati promettenti, suggeriscono anche la necessità di ulteriori sforzi da parte della ricerca per una maggiore validazione dei risultati.

Indice delle Figure

Indice delle Tabelle

1. Introduzione

Fin dalla sua nascita, il Web ha avuto come primo scopo quello di mettere in comunicazione diretta persone e risorse. L'aumentare delle informazioni sul Web fu un incentivo a navigare per le persone, e l'aumento dei browser creati fu lo stimolo a creare più siti Web. Sebbene il Web sia profondamente mutato, questo *loop* è ancora in corso. Negli anni, acquisizione, gestione, rappresentazione e fruibilità dell'informazione sono diventati servizi alla portata di tutti. Tuttavia, la quantità spropositata di dati condivisi, di natura eterogenea, rischia di diventare un limite per l'uomo quanto per le macchine. Fare ricerche, inferenze, ed estrarre informazioni è tanto più difficile quanto più sono diversificati e numerosi i dati da analizzare. È possibile, perciò, che informazioni rilevanti non giungano mai alla luce condizionando così sia la ricerca sia lo sviluppo, ambiti a cui il Web era primariamente rivolto.
Con il Web Semantico, ed in particolare con le Ontologie, si propone come soluzione l'uso di un modello formale che descriva l'informazione secondo specifiche regole sul piano semantico e sintattico. L'Ontologia permette di identificare in modo univoco, esplicito, omogeneo e quanto più possibile automatico concetti, equivalenze, relazioni e proprietà in un dominio di conoscenza. Permette, inoltre, di fare inferenze sulla conoscenza che essa stessa rappresenta. In ambito biomedico e bioinformatico, questo modo di rappresentare l'informazione in forma interoperabile risulta particolarmente utile per gestire conoscenza e dati contenuti in database, articoli, report, brevetti e tanto altro.

Con questo elaborato si propone un'analisi della struttura semantica dell'informazione espressa dalle Ontologie e di come questa possa essere uno strumento in applicazioni di tipo biomedico. Una prima versione di questo progetto aveva lo scopo di identificare le principali ontologie attualmente utilizzate, il loro dominio, struttura e campi di utilizzo. Tuttavia, nella fase

preliminare di ricerca bibliografica e di documentazione sono stati riscontrati problemi di inconsistenza. Molti dei siti Web di riferimento non risultavano più attivi e le poche ontologie di cui di cui è stato possibile analizzare il contenuto dimostravano di essere molto datate nel linguaggio e nel contenuto. La quasi totalità di queste, inoltre, presentava un contenuto esiguo di termini e proprietà, insufficiente per poter effettivamente esprimere la porzione di conoscenza di dominio che asserivano di rappresentare. Tuttavia, da questo studio è emersa l'esistenza di consistenti librerie di ontologie biomediche, aggiornate e di semplice utilizzo. È stata valutata, così, la possibilità di un percorso diverso volto alla scoperta di applicativi che coinvolgessero l'uso di ontologie nel ramo della biomedica. Il progetto è stato sviluppato partendo dalla descrizione di Web Semantico e dei linguaggi che lo compongono. Successivamente, è stato svolto un lavoro di analisi della struttura ontologica in tutte le sue componenti principali, fornendo indicazioni sugli stadi fondamentali del processo di genesi di una ontologia. È stata realizzata una classificazione delle ontologie per tipologia e utilizzo e sono stati evidenziati alcuni limiti nel riuso e nell'integrazione delle strutture ontologiche per generarne di nuove. A seguire sono state presentate alcune delle più note ontologie biomediche, affermate da anni nel panorama scientifico della ricerca e dello sviluppo. Infine, sono stati identificati i settori di ricerca che attualmente sono impegnati nella realizzazione di sistemi e tool in grado di comunicare con specifiche ontologie biomediche.

A conclusione dell'elaborato sono state fatte considerazioni sulle attuali applicazioni dell'ontologia in campo biomedico evidenziando non solo i limiti ma anche le possibilità di sviluppi futuri per la ricerca.

2. Web Semantico

"The first goal was to enable people to work together better. [..] the original driving force of the Web was collaboration at home and at work. [..] How much lack of cooperation can be traced to an inability to understand where another party is ''coming from''? The Web was designed as an instrument to prevent misunderstandings. […] In a world of people and information, the people and information should be in some kind of equilibrium. [..] The Web should be a medium for communication between people: communication through shared knowledge. [..] We are forming cells within a global brain, and we are excited that we might start to think collectively. What becomes of us still hangs crucially on how we think individually."

T. Berners Lee- London, 1997

Questo estratto fu parte di un discorso di Berners Lee, padre fondatore del Web, ad un incontro del W3C a Londra nel 1997 e riportato come prefazione del libro *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential* nel 2002. Parte del progetto originale del Web era riuscire a generare una struttura interoperabile che solo protocolli standard globali potevano garantire.
Il Web Semantico, accuratamente definito da standard, rappresenta la realizzazione di un aspetto del Web in grado di potenziarne l'interoperabilità a livello sintattico e semantico [1]. Si afferma, dunque, una visione per la quale il *Web of links*[1] necessita di progredire mediante il *Web of meaning.* Le risorse Web sono descritte in un linguaggio che renda il loro significato sempre più esplicito[2] man mano che si risalgono gli strati della priamide che costituiscono la struttura del Web Semantico (**Figura 1**). Alla base troviamo lo standard XML [2] , gli URI [3], e l'Unicode [4]. Su questi primi strati si basa un altro standard: RDF (*Resource Description Framework*) [5], primo vero strato in grado di descrivere metadati[3] [6]. Infatti, *Hypertext Markup Language* (HTML)[7] e più in generale *Extensible*

[1] L'essenza del Web per come lo conosciamo è uno spazio universale di informazioni alle quali si accede mediante collegamenti ipertestuali.

[2] per grammatica e semantica.

[3] Letteralmente rappresentano i "dati dei dati", includono tutte le proprietà e le informazioni per individuare e descrivere un dato.

Markup Language (XML) forniscono ai documenti struttura, ma non semantica [6].

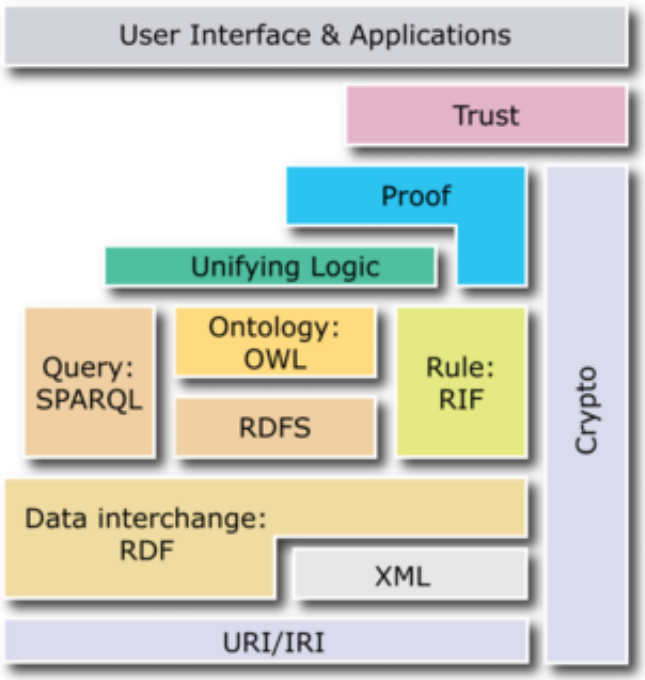


Figura 1 *– Piramide del Web semantico (Fonte:* ***https://commons.wikimedia.org/wiki/File:Semantic_Web_Stack.png*** *).*

Il linguaggio RDF è composta una tripla del tipo *Soggetto – Predicato*[4] *– Oggetto.* Tuttavia, tramite RDF non si ha nessun modo per dichiarare le proprietà espresse tramite predicato, né per definire le relazioni tra queste proprietà e altre risorse [5]. Ad un livello superiore vi sono RDF *Schema* [8] e OWL [9], due linguaggi in grado di connettere asserzioni sulla conoscenza a inferenze logiche su diversi livelli di espressività e complessità computazionale (**Figura 2**). Questi due costrutti definiscono il tessuto semantico e sintattico delle Ontologie. In particolar modo OWL (*Web Ontology Language*) è un linguaggio in grado di esprimere strutture semantiche più complete ed articolate rispetto a proprietà e classi espresse

[4] Proprietà che collega Soggetto ad Oggetto.

tramite RDF *Schema.* Ad oggi OWL e le sue versioni successive come OWL2 sono i linguaggi principalmente utilizzati per sviluppare ontologie [1]. Una comparazione trai i vari livelli di linguaggio semantico è schematizzata in **Tabella 1.** Tutti i linguaggi del Web Semantico sono interrogabili tramite SPARQL (*SPARQL Query Language for RDF*) [10] che consente di estrarre informazioni sulle basi di conoscenza del Web. Ha una struttura simile alle *query in* SQL[5], esempio in **Figura 3**. Successivi allo strato ontologico vi sono, poi, i processi di validazione e verifica dell'ontologia prima che essa venga impegnata in interfacce e applicazioni di vario genere. Il Web Semantico, così descritto, genera un nuovo livello di astrazione dal network sottostante capace di ridurre la confusione e gli errori nell'interpretazione dell'informazione[6], formattare i dati in modo che siano *machine readable,* permettere l'interoperabilità tra sistemi e persone, collegare dataset[7] attribuendo ai metadati sul Web livelli di rappresentazione semantica [12].

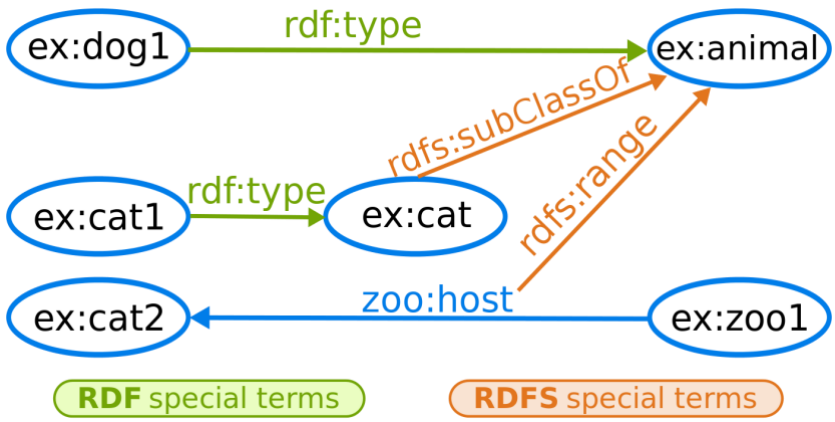


***Figura 2** – Esempio di come costruire un RDF Schema da RDF statement (Fonte: https://it.wikipedia.org/w/index.php?title=RDF_Schema&oldid=118733638).*

[5] *Structured Query Language,* liguaggio standardizzato per database sul modello relazionale [11].
[6] Dovuti anche alla duplicazione delle informazioni.

[7] Come per esempio i Linked Open Data.

RDF	RDFS	OWL	OWL2
*Domain independent. *States fact in triples and establishing the relation between two ends.	*Provide mechanism for defining specific domain. *States class and property relation. *Declares class and subclasses in subsumption, supports property and subproperty, domain and range restriction. *Logical combination beyond its use.	*Compatible with several existing ontology languages e.g. OIL, DAML + OIL. *Extends RDF fact stating ability, and RDFS class and property structure ability. *Declares class and subclasses in subsumption hierarchy. *Classes can be logical combinations (intersection, union, negation) of other classes, or as enumeration of other specific object. *Extends RDFS by declaring properties as transitive, symmetric, functional or inverse. *Expresses disjoint, equivalence, individuality of object, quantification and value restriction. *Relies on XML schema (xsd) for listing datatypes. *Based on SHOIN(D) and NExpTime-complete.	*Compatible with OWL. *Extends and Improves datatype handling. *Additional property and addresses qualified cardinality restriction. *Simplified meta-modelling and extended annotation. *Primary exchange syntax is RDF/XML. Like OWL, OWL2 has other syntaxes e.g. Turtle, XML, RDF graph, Manchester syntax. *Has three profiles OWL2 EL, OWL2 QL, OWL2 RL with trade-offs for expressivity. *Based on STROIQ(D) and 2NExpTime-complete.

Tabella 1 *- Tabella di comparazione tra RDF, RDFS, OWL, OWL2.*
(Fonte: https://www.researchgate.net/figure/Comparison-of-RDF-RDFS-and-OWLlanguages_fig1_344393345).

```
PREFIX foaf:   <http://xmlns.com/foaf/0.1/>

SELECT ?name

WHERE { ?x foaf:name ?name }

ORDER BY ?name
```

Figura 3 *– Esempio di Query su SPARQL (Fonte: https://www.w3.org/TR/rdf-sparql-query/).*

3. Ontologia

In filosofia, l'ontologia è lo studio e la rappresentazione dell'essere mediante caratteristiche capaci di esprimerne il significato.

Il termine "ontologia" deriva dall'unione di due parole greche, "*onto"* participio presente del verbo essere [13] e "*logos"* ovvero parola [14], più propriamente il significato esprimibile mediante la parola o il pensiero.

La definizione più rappresentativa dell'ontologia appartiene a Gruber: *"Un'ontologia è una specificazione formale ed esplicita di una concettualizzazione condivisa"* [15].

In particolare:

- la "*specificazione"* è possibile mediante la descrizione dei metadati con una sintassi e una struttura specifica [16];
- la "*concettualizzazione*" consiste nel definire un modello astratto del mondo e della realtà (oggetti, concetti, e relazioni) [17].
- l'essere "*esplicito*" significa che tutto ciò che viene utilizzato per esprimere la concettualizzazione è descritto in modo chiaro e comprensibile;
- l'essere "*formale*" implica che l'ontologia sia costruita seguendo specifiche regole ed assiomi, tali per cui essa possa essere comprensibile anche dalle macchine, secondo vari livelli di formalismo [18];
- l'essere "*condivisa*" riflette il fatto che la conoscenza espressa mediante ontologia sia riconosciuta da più comunità.

Alcuni modelli di riferimento possono essere il modello entità-relazione e modello orientato ad oggetti, in quanto forniscono costrutti per rappresentare la conoscenza della realtà mediante definizione di entità e relazioni in un dominio [1]. L'ontologia può essere vista come una quintupla [19], formata da:

1. Classi e sottoclassi;
2. Proprietà riferite ad oggetti;
3. Attributi;
4. Assiomi;

5. Istanze.

Inoltre, per essere considerata tale un'ontologia deve avere [20] :

- un vocabolario finito e controllato;
- rigide relazioni gerarchiche di sottoclassi tra le classi;
- interpretazione non ambigua delle classi e delle relazioni tra le varie entità.

Altre caratteristiche desiderabili ma non obbligatorie sono proprietà e valori specifici definiti in base alla classe, inclusione di istanze e specificazione di classi disgiunte. Un vocabolario controllato, eventualmente estensibile, fornisce una rappresentazione esplicita del concettualismo che si vuole esprimere [1]. La rappresentazione della conoscenza mediante ontologie non è importante solo nel suo costituire un modo formale, sintatticamente e semanticamente coerente della realtà, ma anche nella possibilità di fare inferenze sulla base di conoscenza espressa [1]. Il tipo e la specificità delle inferenze che si possono desumere dalle ontologie sono strettamente legati alla qualità descrittiva dell'ontologia che dipende soprattutto proprietà e assiomi.

3.1 Struttura dell'Ontologia

3.1.1 Classi

Classi e sottoclassi rientrano in una categoria nota anche come "concetti" od "entità". Questi esprimono concettualmente la porzione di conoscenza che si vuole rappresentare.

Un concetto può essere una qualsiasi entità concreta o astratta (oggetto, nozione, idea, processo) [21].

Quindi, ne fanno parte esempi come:

- "malattia";
- "processo di vascolarizzazione";
- "diagnosi".

Le entità sono raggruppate tramite proprietà comuni. Le proprietà (attributi) relative alle entità di tipo superclasse[8] sono ereditate da tutte le sottoclassi che da esse discendono (**Figura 4**).

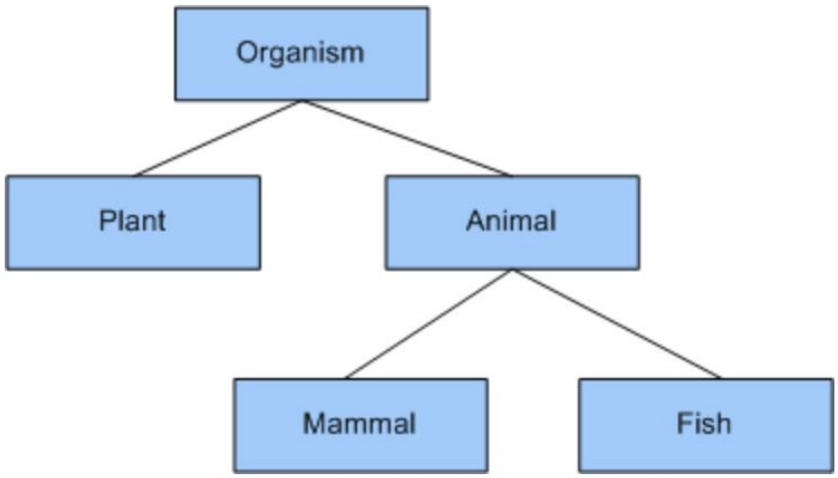


Figura 4 – *Rappresentazione gerarchica di classi e sottoclassi (Fonte: https://jena.apache.org/documentation/ontology/).*

[8] Classe ad un livello superiore, più generico.

3.1.2 Proprietà

Le proprietà [16] sono di due tipi:

- attributo o funzione;
- relazioni o proprietà riferite agli oggetti.

Nel primo caso ci si riferisce a una serie di caratteristiche che definiscono i concetti, ovvero gli attributi.

Nella struttura gerarchica precedente in **Figura 4,** si può per esempio dire che attributi di “Organismo” sono:

- “nome”;
- “specie”;
- “si trova in”.

Gli esempi appena fatti dipendono molto dal progetto dell’ontologia. Se l’intenzione fosse stata rappresentare il mondo animale in base alla sua geolocalizzazione molto probabilmente l’attributo “si trova in” sarebbe piuttosto una relazione tra classe “Animale(x)” e “Posto_del_Mondo (y)”.

Il secondo tipo di proprietà, infatti, più che definire un gruppo di sottoclassi, mette in relazione due o più entità dell’ontologia.

Questo è il caso, per esempio, della più comune relazione “is-a” presente nelle ontologie, utile a definire sottoclassi di classi: “Animale è (*is-a*) un Organismo.”

3.1.3 Assiomi

Gli assiomi sono per definizione affermazioni sempre vere. Funzionano come regole e vincoli, necessari per fare inferenze sull’ontologia [1]. Tramite gli assiomi è possibile:

- vincolare l’informazione contenuta nell’ontologia;
- dedurre nuova conoscenza;
- verificare la correttezza.

Il contenuto assiomatico è ciò che differenzia maggiormente i livelli di linguaggio del Web Semantico e che ne aumenta la capacità espressiva. Esistono diversi tipi di assiomi diversificati per le classi, proprietà, e istanze. Una classificazione delle regole assiomatiche è proposta in dettaglio nella tabella in **Tabella 3**.

No.	Logic	OWL	%
1	$A \sqsubseteq B$	subClassOf(A B)	51
2	$A \sqsubseteq \exists P.B$	subClassOf(A someValuesFrom(P B))	33
3	$[a, b] \in P$	propertyAssertion(P a b)	8
4	$a \in A$	classAssertion(A a)	4

Tabella 2 *- Quattro più comuni schemi assiomatici e la relativa percentuale media di utilizzo nel linguaggio OWL (Fonte:* ***https://aclanthology.org/W10-4222.pdf*****).**

Type	Axioms	Meaning
Classes	**EquivalentTo**	A class expression that is equivalent to the current selected class.
	SubClassOf	A class expression that the current selected class is a subclass of. In other words, each row is a superclass of the current selected class.
	DisjointWith	A list of class expressions that this class is disjoint with. A DisjointClasses axiom can contain 2 or more classes (the current selected class is removed from the list for clarity).
	DisjointUnion Of	Specifies that this class is the main class in a DisjointUnion class axiom.
Properties	**Equivalent To**	The selected property is equivalent to each of the properties listed in this section.

	SubProperty Of	The selected property is a subproperty of each of the properties listed in this section.
	Inverse Of	The selected property is the inverse of each of the properties listed in this section.
	Disjoint With	The selected property is disjoint with each property that is listed in this section.
	Symmetrics With **Asymmetric With** **Reflexive With** **Transitive WIth**	The selected property is respectively symmetric, asymmetric, reflexive, transitive with each property that is listed in this section.
	Domain	The selected property has each class expression listed in this section in its domain. If a given property has a given class in its domain this means that any individual that has a value for the property (i.e. is the subject of a relation along the property), will be inferred to be an instance of that domain class.
	Range	The selected property has each class expression listed in this section in its range. If a given property has a given class in its range this means that any individual that is the value for the property (i.e. is the object of a relation along the property), will be inferred to be an instance of that range class.
	SubPropertyChainOf	The selected property is a super property (i.e. implied by) each chain of properties listed in this section.

Instances	**Same Individual As**	Displays a list of individuals that the selected individual is asserted or inferred to be the same as.
	Different Individuals	Displays a list of individual that the selected individual is *asserted* to be different from.
	Types	Displays a list of class expressions that the selected individual is a direct instance of.

Tabella 3 *- Classificazione dei principali assiomi esistenti per i linguaggi OWL e OWL2 rispettivamente per Classi, Proprietà e Istanze.*
*(Fonte: **http://protegeproject.github.io/protege/views**).*

3.1.4 Istanze

Le istanze rappresentano specifici oggetti (concreti o meno) del mondo realtà. Possono essere anche fatti o semplicemente affermazioni [16].
Definite a partire da classi, ne ereditano attributi e proprietà.
Per esempio, possiamo dire che uno specifico paziente "Paz", riconosciuto con uno specifico numero (da 1 a n), sia una istanza della classe "Pazienti" e che può essere affetto ("affetto_da", proprietà) una specifica sottoclasse di "Patologie" (**Figura 5**). A loro volta le sottoclassi di "Patologie" potrebbero essere o sottoclassi, appunto, oppure istanze della classe "Patologie".
Per discriminare una istanza semplice[9] da una istanza relativa a fatti e affermazioni si propone il seguente esempio:

- "Uomo" è istanza di "Mammifero" → istanza *semplice*;
- "L'uomo vive perlopiù sulla terra" → istanza di tipo *fatto*;
- "L' Antropologo dice che l'uomo vive perlopiù sulla terra" → istanza di tipo *affermazione*.

[9] Descrive un oggetto della realtà.

È importante fare questa distinzione perché le risorse del Web Semantico sono in grado di fare affermazioni.

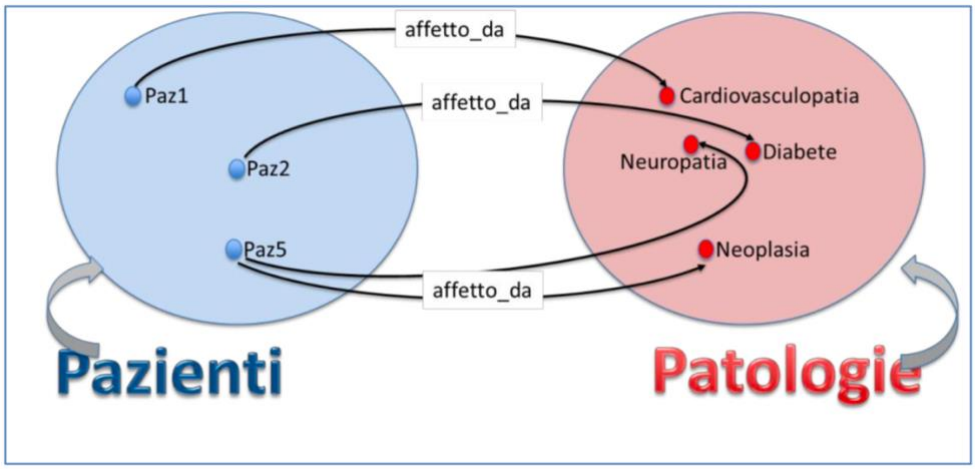


Figura 5 *– Esempi di istanze con dominio "Pazienti" affetti da "Patologie".*
(Fonte: https://core.ac.uk/download/pdf/37830677.pdf).

3.2 Genesi e Sintesi di una Ontologia

Le ragioni per sviluppare un'ontologia sono molteplici:

- condivisione di strutture di informazione comuni tra persone o macchine;
- creare una base di conoscenza riutilizzabile ed integrabile;
- rappresentazione di un dominio in modo esplicito;
- possibilità di fare ricerche e *query* su necessità.

Affinchè queste azioni siano possibili, è necessario che l'ontologia abbia una struttura consistente e coerente. Non esiste un modo assolutamente "corretto" per costruire un'ontologia: dipende molto dal tipo di uso applicativo che se ne vuole fare. Tuttavia, esistono problematiche comuni quando ci sia approccia ad un progetto di questo tipo. Lo sviluppo di una ontologia è un processo iterativo in cui si possono identificare *step* principali da raggiungere per evitare l'insorgere di problemi strutturali o logici [22]. La genesi di un'ontologia prevede la pianificazione del lavoro di sintesi in stati progressivi e la preliminare padronanza della conoscenza nel dominio scelto (**Figura 6** , **Figura 7**).

Una volta creata l'ontologia, il lavoro continua nei processi di mantenimento. Questo *step* è particolarmente importante sotto molti aspetti.

Mantenere un'ontologia significa:

- permettere che il suo contenuto (OWL, RDFS, Turtle[10]) sia visibile e scaricabile (anche mediante licenza);
- aumentare la sua visibilità nel Web inserendola all'interno di librerie note e creando siti Web di riferimento mantenuti attivi;
- aggiornarla periodicamente;
- adattarla a nuove necessità.

Parte del lavoro di questo elaborato si è dovuta confrontare con problemi propri del mantenimento. Fornire ontologie citate con collegamento ipertestuale a un sito Web inesistente è inutile almeno quanto non poter scaricare il file relativo per analizzarne il contenuto o riscontrare che una ontologia è datata.

[10] Terse RDF Triple Language.

Dare visibilità alle ontologie significa mantenerle vive, utilizzabili, e significative [1]: se questi aspetti non vengono curati le ontologie si perdono nel Web esattamente come succederebbe ad un qualsiasi altro dato.

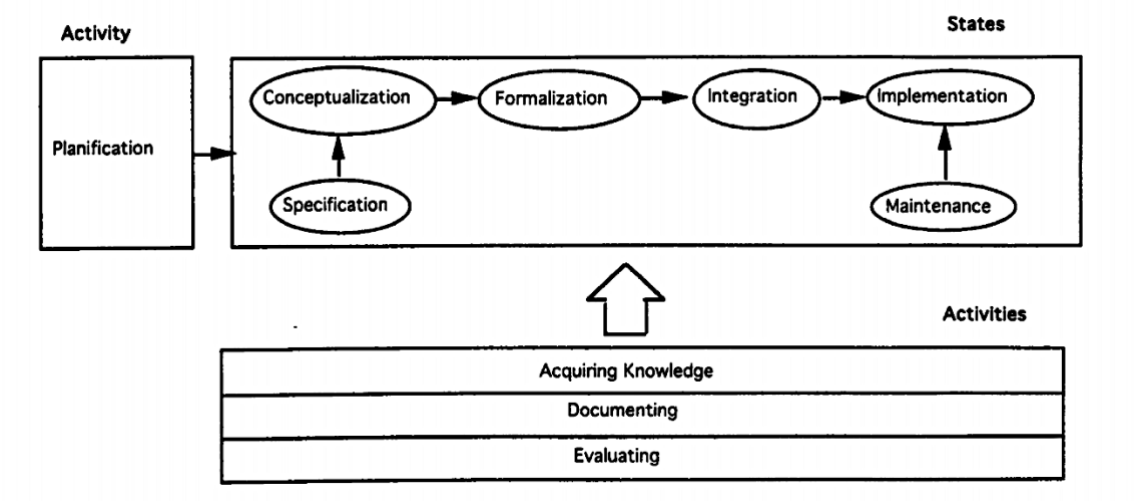


Figura 6 *– Processo di sintesi di una Ontologia: principali macro-step. (Fonte: http://oa.upm.es/5484/1/METHONTOLOGY_.pdf)*

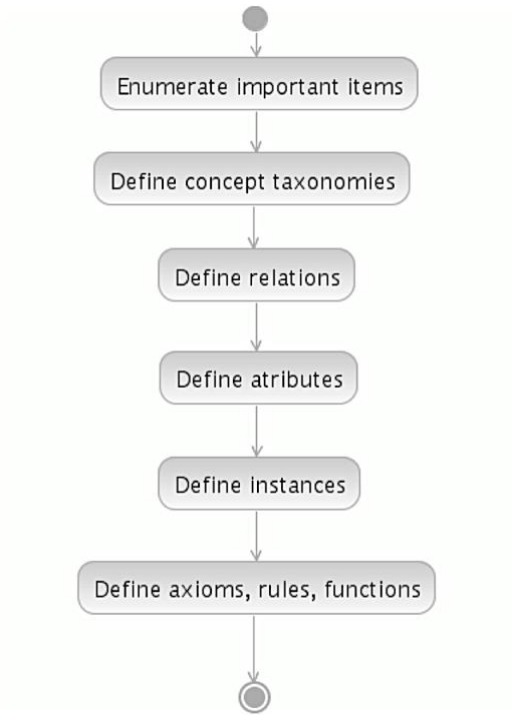


Figura 7 *– Schema metodologico del processo di sintesi: principali micro-step (Fonte: http://www.aslab.org/documents/controlled/ASLAB-R-2007-004.pdf)*

3.2.1 Passo 1: definire lo scopo

In generale, per strutturare un progetto è necessario avere la consapevolezza del fine ultimo da raggiungere.

Quindi, in questo caso, per costruire un'ontologia è utile chiarire [23]:

- dominio di appartenenza;
- utilizzo dell'ontologia;
- tipo di domande a cui l'ontologia deve saper rispondere;
- a chi o che cosa è rivolta;

Nel corso del suo sviluppo, le risposte ai tre punti precedenti potrebbero cambiare, ma l'importante è averne ben chiara la risposta.

Per il mantenimento di questo primo *step* si possono elaborare delle *domande di competenza* [24] ovvero domande specifiche a cui la base di conoscenza dell'ontologia deve saper rispondere. Queste domande possono essere usate come test di verifica a fine progetto per valutare se l'ontologia ha abbastanza informazioni per poter rispondere o se invece alcune domande necessitano di un livello di dettaglio più specifico in determinate aree dell'ontologia.

Definito lo scopo principale per l'ontologia che si vuole sviluppare, è consigliabile fornirsi di una buona documentazione e richiedere il supporto di esperti in quel dominio di conoscenza [21].

Una buona strategia è riutilizzare altre ontologie: se esistono ontologie nello stesso dominio di conoscenza modificabili, si possono rielaborare affinchè rispondano allo scopo prefissato. Questa, tuttavia, è un'operazione delicata che va affrontata avendo cura di non generare ambiguità, soprattutto nel caso di *merging*[11] di più ontologie (il riuso di ontologie è argomento del capitolo 3.4 Integrazione e Riuso di Ontologie: alcuni limiti).

[11] Fusione di più ontologie. Gli errori più frequenti nel *merging* sono ambiguità nei termini: stesso termine ma con significati semantici differenti in due ontologie, due termini diversi (da due ontologie) con stesso significato, proprietà contrastanti definite per uno stesso termine.

3.2.2 Passo 2: creare una lista di termini

Per iniziare *ex novo* una ontologia potrebbe essere utile stilare una lista di tutti i termini (o frasi complete) [25] che l'ontologia deve contenere affinchè possa descrivere in modo completo la realtà che vuole rappresentare.
Questa lista comprende concetti, proprietà e relazioni senza nessuna distinzione.
A seguito di questa fase inizia la vera e propria "Concettualizzazione", discriminando prima classi e sottoclassi, poi relazioni e proprietà, e con loro gli assiomi.

3.2.3 Passo 3: definire le classi e gerarchia tra le classi

Come anticipato, in questo passaggio si devono distinguere nella lista fatta le classi dalle proprietà e dalle relazioni. Le entità devono essere raggruppate distinguendo classi e relative sottoclassi.
Questo procedimento può essere realizzato in tre modi diversi [26]:

1. *top-down*: da classi più generali si prosegue verso una specificità maggiore;
2. *bottom-up*: dai concetti più specifici si risale verso quelli più generici, fino alla radice ("*Thing*");
3. *middle-out*: una combinazione tra le prime due. Nel definire una sottoclasse di una superclasse si può decidere (anche in un secondo momento) di inserire una classe intermedia.

Anche in questo caso non esiste il metodo migliore, la scelta può dipendere da una preferenza dello sviluppatore e dal tipo ontologia che si vuol realizzare.
Qualunque sia la scelta che si decide di fare, rimane vincolante il modo con cui si differenziano classi (superclassi) e sottoclassi.
"Se una classe A è una superclasse di B, allora ogni istanza di B è anche istanza di A" [23].
Errori comuni in questo *step* sono:

- cicli nella gerarchia: quando una classe A ha una sottoclasse B e allo stesso tempo B è superclasse di A;

- parentele con gradi di generalità diverse: ogni parente deve essere allo stesso livello di generalità rispetto alla radice. Solo la radice in quanto tale è esente da questa regola [23];
- sottoclassi di una classe con proprietà in più, relazioni e restrizioni diverse rispetto alla superclasse: implica la necessità di aggiungere una nuova classe.

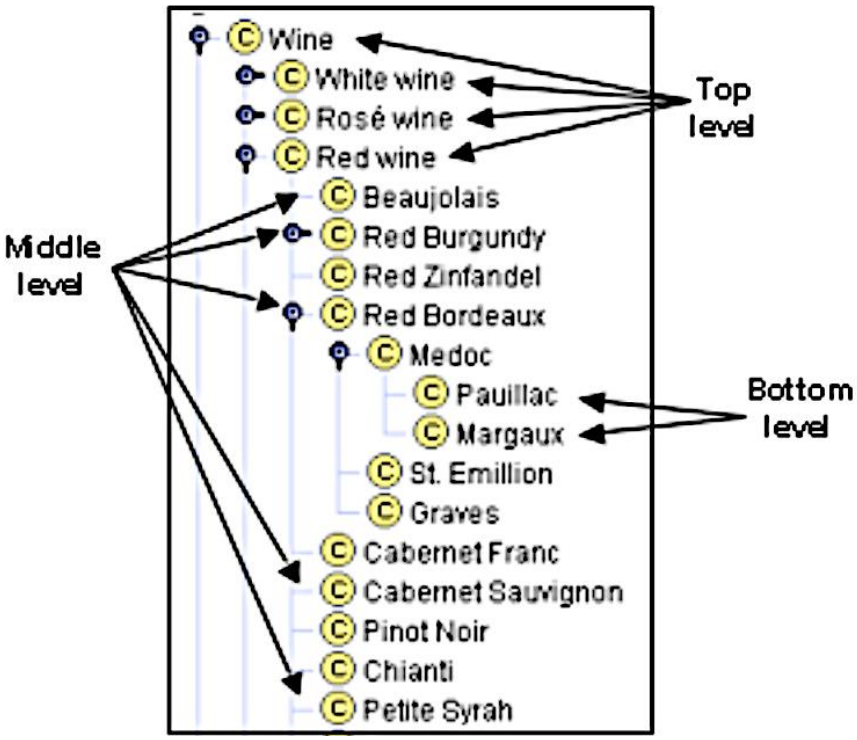


Figura 8 *– Esempio di struttura gerarchica: dalle sottoclassi più specifiche, a quelle più generali, passando per quelle intermedie.*
(Fonte: https://protege.stanford.edu/publications/ontology_development/ontology101.pdf).

3.2.4 Passo 4: definire le proprietà

Della lista dei termini fatta inizialmente rimangono adesso solo proprietà e istanze.
Le proprietà di una classe possono essere intrinseche ed estrinseche, e tutte le sottoclassi ereditano le proprietà della superclasse [25].

In questa fase si possono assegnare alle proprietà valori [19] quali:

- cardinalità: numero di valori che può assumere;
- tipo: numero, stringa, booleano, enumerato, istanza;
- dominio:
- range.

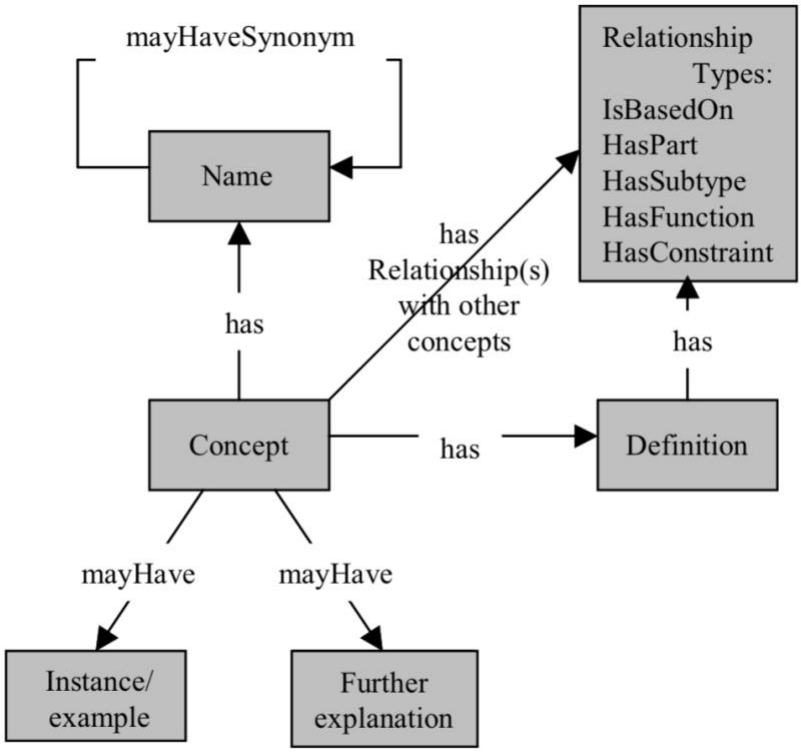


Figura 9 *– Rappresentazione delle possibili proprietà di una classe.*
(Fonte: https://core.ac.uk/download/pdf/11310019.pdf)

3.2.5 Passo 5: dichiarare le istanze

L'ultima cosa che rimane da fare è generare le istanze. Esse rappresentano la realtà che si vuole rappresentare con le ontologie. L'ontologia insieme a una serie di istanze forma la base di conoscenza. Una ontologia senza istanze è dunque incompleta.

3.2.6 Passo 6: Classe o Proprietà? Classe od Istanza?

Altre difficoltà nello sviluppare una ontologia sono riuscire a scegliere se creare una nuova classe oppure aggiungere una proprietà o un'istanza (e viceversa).
In entrambi i casi la risposta a questa domanda è sempre il tipo di ontologia che dobbiamo costruire: alcune proprietà potrebbero essere del tutto marginali come classi. L'istanza è il concetto più specifico che si può rappresentare [23]. Potrebbe non essere utile inserire classi ottenendo istanze ancora più specifiche. Viceversa, se l'istanza esprime concetti che formano una struttura gerarchica aggiungerla come classe renderebbe l'ontologia sintatticamente più corretta.

3.3 Classificazione delle Ontologie

Dalla fine degli anni Novanta, con l'emergente interesse nell'ambito del Web Semantico, ricercatori e sviluppatori si sono dedicati allo studio e alla creazione di ontologie sempre più prestanti e diversificate. Esistono diversi modi in cui poter classificare le ontologie, tuttavia la distinzione tra le varie categorie non è mai netta, poiché una stessa ontologia può appartenere a più categorie. Una prima distinzione può essere fatta considerando il livello di generalità usato per la descrizione dei domini che permette di individuare le seguenti categorie:

- *Ontologie Generiche o Superiori*: descrivono un insieme di oggetti, concetti e relazioni applicabili a diversi domini della conoscenza con lo scopo di creare un collegamento superiore tra tutte le ontologie specializzate[12];
- *Ontologie di Dominio*: descrivono categorie, relazioni e proprietà che sono specifiche di un particolare dominio.

Dai capitoli precedenti abbia visto come l'ontologia sia la "specificazione di una concettualizzazione". Possiamo esprimere la distinzione tra Ontologia Superiore e quella di Dominio anche come una diversa "specificazione del soggetto della concettualizzazione"[13] [27]: nel primo caso è generale, nel secondo è specifica. Sotto questo nuovo criterio di classificazione si possono aggiungere altri due tipi di ontologie:

- *Ontologie di Rappresentazione*: spiegano le concettualizzazioni che stanno alla base della rappresentazione formale della conoscenza, senza però fare nessuna asserzione sulle entità del mondo (come per esempio oggetti fisici, eventi e processi). Sono utilizzate per specificare le categorie di meta-livello usate per modellare la realtà. Questo tipo di ontologie descrivono categorie, relazioni e proprietà usate per esprimere altre ontologie, ovvero sono delle Meta-Ontologie. Perciò le Ontologie di

[12] Di dominio.

[13] Ovvero la specificazione del soggetto del modello formale con cui rappresentiamo la conoscenza (e dunque la realtà).

Dominio e le Ontologie Generiche vengono descritte a partire da primitive[14] [28] fornite dalle Ontologie di Rappresentazione.

- *Ontologie di Applicazione*: contengono tutte le definizioni necessarie a modellare la conoscenza richiesta per una specifica applicazione. A tale scopo possono essere combinate concettualizzazioni proprie sia delle Ontologie di Dominio sia delle Ontologie Generiche tramite un processo di integrazione e fusione. L'ontologia risultante è particolare per una determinata applicazione e non sempre può essere riutilizzata per un compito differente [29].

Le ontologie possono anche essere classificate in funzione della quantità e del tipo di strutture di concettualizzazione permettendo una suddivisione in:

- *Ontologie Terminologiche*, che specificano i termini usati per rappresentare la conoscenza di un determinato dominio di interesse. Un esempio di questo tipo di ontologia in campo biomedico è il network semantico UMLS (*Unified Medical Language System*) [30].
- *Ontologie di Informazione*, che specificano la struttura o l'organizzazione logica dei dati contenuti nei database descritta mediante l'uso di un linguaggio formale. Un esempio di questa classe di ontologie è rappresentato dal primo livello del Modello PEN&PAD [31], un framework sperimentato in UK per generare le cartelle cliniche dei pazienti ricoverati. Il modello è diviso in due fasi. Nella prima è possibile registrare con il medesimo formalismo i dati e le osservazioni dirette di pazienti diversi. Tuttavia, a questo livello non è possibile fare distinzioni tra sintomi, segnali clinici, trattamenti [32]. La seconda fase sfrutta l'utilizzo di Ontologie di Modellazione della Conoscenza.
- *Ontologie di Modellazione della Conoscenza*, che specificano la concettualizzazione della conoscenza, per cui la struttura interna propria queste ontologie è molto più ricca delle Ontologie di Informazione. Inoltre, proprio per il loro scopo la loro conoscenza è specifica per ciò che devono

[14] Le primitive sono proposizioni, predicati, funzioni logiche e operatori, ottenuti da una semantica formale volta a rappresentare la realtà del mondo sia come relazione tra oggetti del mondo reale (entità del mondo) sia come categorie di categorie (categorie di meta-livello) usate per modellare il mondo (e del mondo) sia come categorie di categorie (categorie di meta-livello) usate per modellare il mondo (e le entità che esso contiene), soddisfacendo il significato formale dei postulati (Nicola Guarino,"The Ontological Level").

descrivere. Un esempio è il secondo livello del Modello PEN&PAD [33]. A questo livello le osservazioni fatte nella fase precedente venivano poi usate nel *decision making*. Il modello complessivamente si presenta come una serie di "descrizioni" formulate nel SMK (Structured Meta Knowledge) un network semantico del linguaggio usato per rappresentare formalmente non solo la terminologia linguistica medica ma anche dati medici e conoscenza pragmatica sulla pratica clinica.

Le ontologie possono differire anche per il per il livello di formalismo utilizzato per esprimere i termini e i loro significati. Ad incidere sulla scelta del livello di formalismo, è per gran parte il livello di automazione che l'ontologia deve supportare. Se l'ontologia è utilizzata per comunicare fra persone, la rappresentazione dell'ontologia può essere informale, mentre se l'ontologia è pensata per sistemi automatici, allora assumerà un linguaggio più formale comprensibile appunto dalle macchine oltre che dalle persone [25]. Un livello di formalismo maggiore non implica un'ontologia maggiormente sviluppata bensì una con un livello di specificazione maggiore. Si distinguono così:

- *Ontologie altamente informali*: sono espresse in linguaggio naturale (con l'intrinseca ambiguità nella definizione dei termini tipica di questo linguaggio);
- *Ontologie semi-informali*: sono espresse in una forma più rigida e strutturata del linguaggio naturale, migliorando la chiarezza e riducendo le ambiguità;
- *Ontologie semi-formali*: sono espresse mediante un linguaggio formale pur conservando una parte meno formale che consenta un accesso intuitivo anche alle persone;
- *Ontologie rigorosamente formali*: sono ontologie i cui termini sono precisamente definiti con un linguaggio formale, teoremi e prove di proprietà, sviluppate e pensate specificatamente per uso macchina.

Una ulteriore classificazione può essere fatta tenendo conto della loro espressività [20]. Per essere maggiormente espressiva una ontologia deve prevedere una quantità maggiore di vincoli nelle proprietà aumentando la complessità assiomatica interna. Per esempio, alla proprietà "ossigenazione" si potrebbe associare una restrizione di tipo numerico espressa tramite un range di

valori che essa può assumere oppure tale valore potrebbe essere ricavato mediante equazioni matematiche che usano i valori di altre proprietà. Si distinguono così ontologie semplici da quelle con un grado di espressività elevato. La costruzione di un'ontologia complessa comporta costi molto elevati, talvolta anche proibitivi[15], non solo per la sua creazione ma anche nell'ottica di utilizzi applicativi. Un esempio di ontologia semplice è la Unified Medical Language System (UMLS) [30], già citata in precedenza, è una grande ontologia sulla terminologia medica, mentre la più grande, complessa e completa base di conoscenza di senso comune attualmente disponibile è la CYC[16] del progetto Cycorp [34].

[15] in termini di tempo, risorse ed economico

[16] Generata sulla base di microteorie ognuna delle quali può essere vista come una serie di asserzioni che descrivono un particolare dominio.

3.4 Integrazione e Riuso di Ontologie: alcuni limiti

Una delle proprietà principali delle ontologie risiede nella possibilità di essere riutilizzate od integrate tra di loro a formare una nuova rappresentazione della conoscenza di dominio.

Una ontologia per essere riutilizzabile in più applicazioni non può essere specifica. Dal punto di vista teorico, il riuso è possibile: un concetto può essere utile per una determinata funzione senza essere necessariamente specifico per quel dato compito [35]. Tuttavia, nella pratica il riuso di concetti generali non sempre è funzionale: richiede uno sforzo elevato di traduzione dei termini generali in termini specifici.

Una possibile alternativa è utilizzare un'ontologia generica come base per costruire una ontologia più specifica, estendendo classi, proprietà e assiomi dell'ontologia di partenza [1]. Per integrazione o *merging,* si intende il riuso di più ontologie diverse combinate per creare una nuova ontologia [36]. La generazione di queste ontologie prevede un processo delicato di allineamento ontologico per rendere il risultato semanticamente e sintatticamente corretto.

Sono preliminarmente individuate le aree di sovrapposizione delle ontologie, si procede collegando i concetti che sono semanticamente chiusi (attraverso relazioni di equivalenza e l'inclusione in una categoria più vasta che compongono i veri e propri processi di allineamento), ed infine si effettuano i processi di verifica e validazione canonici. Dal *merging* possono nascere problematiche dovute alla difficoltà o l'impossibilità di due o più ontologie di essere unite insieme. La maggior parte di questi problemi sono a livello sintattico, logico, semantico ed espressivo [16]. Per esempio, possono essere presenti termini in più ontologie ma con significati diversi, termini da ontologie diverse con stesso significato, termini con significato simile ma strutturati diversamente nella rispettiva ontologia (diversi percorsi dalla radice alla classe, diverse proprietà associate), concetti simili ma non esattamente corrispondenti a quelli di altre ontologie. Sebbene esistano diversi tool per agevolare il processo di integrazione, sono costosi e non

incrementabili. Inoltre quando si cercano di integrare diverse ontologie, il risultato spesso non trova un riscontro applicativo utile[17][1].

[17] Il risultato è qualcosa che non nasce da una scelta progettuale risultando un misto di termini e proprietà aspecifici per qualsiasi applicazione.

4. Ontologie Biomediche

Lo sviluppo di ontologie nasce come risposta alla necessità di concettualizzare la conoscenza in ogni sua forma, rendendola al tempo stesso omogenea e interoperabile. Le ontologie biomediche forniscono una conoscenza specifica di dominio permettendo la gestione formale di annotazioni e integrazione di dati, il recupero dati (*query*), il supporto alle decisioni [37]. Il rapido aumento di informazioni e dati reperibili è sempre meno compatibile con l'umana gestione di ricercatori e sviluppatori, limitando le capacità sia di analisi che di ricerca. La semantica trasmessa dalle ontologie biomediche fornisce vantaggi quali formalismi per rappresentare la conoscenza e metodi di modellazione concettuale per database. Negli anni, le ontologie biomediche hanno continuato ad ampliarsi e differenziarsi, diventando uno dei principali settori di sviluppo delle ontologie [38]. Con la collaborazione delle comunità scientifiche e di Consorzi, sono stati creati specifici *repository* che raccolgono le ontologie biomediche.

BioPortal è più ricco e noto archivio di ontologie; vanta un totale di 890 ontologie biomediche, biologiche, bioinformatiche, e di scienze naturali in genere [39]. È stato sviluppato dal *National Center fo Biomedical Ontology* (NCBO), uno dei *National Centers for Biomedical Computing* fondati stotto la *NHI Roadmap Initiative*. In *BioPortal* sono state riportate ontologie provenienti da diversi gruppi tra i quali il *Consultative Group on International Agricoltural Research* [40], OBO (*Open Biological and Biomedical Ontology Foundry*) [41] e OBI (*Ontology for Biological Investigation*) [42], *GO Consortium* [43] , la *Proteomics Standards Initiative* [44], la *Unified Medical Language System* (UMLS) [30], e la *Word Health Organization* (WHO) [37]. La ricerca all'interno di questo portale è molto intuitiva. È possibile filtrare la ricerca per linguaggio, nome, dominio scientifico di appartenenza, ottenendo intere ontologie con il relativo file scaricabile, oppure si possono estrarre specifiche annotazioni e *mapping*[18], postare e ricevere commenti mediante interfacce Web (*Annotator, Mappings, Recommender*) [39] (Error! Reference source not found.).

[18] Permette di ricercare e visualizzare il *mapping di classi tra diverse ontologie.*

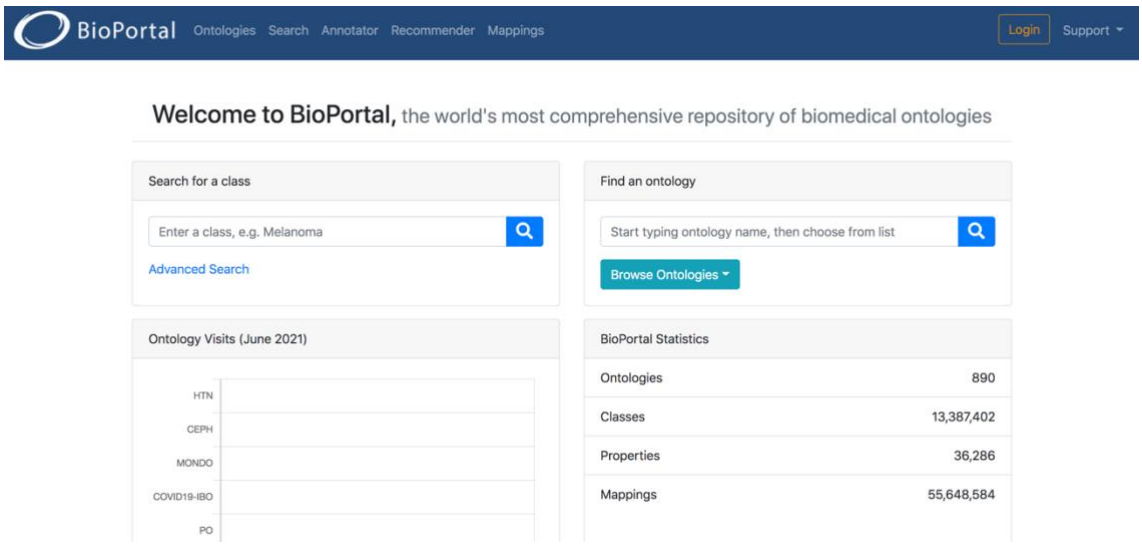


Figura 10 - BioPortal (Fonte: **https://bioportal.bioontology.org/ontologies**).

Alcune delle ontologie biomediche più note ed importanti che si possono trovare all'interno di BioPortal sono:

- *Unified Medical Language System Metathesauru* (UMLS)[30], un compendio di circa 100 vocabolari combinati 2 milioni di nomi e 900.000 concetti;
- *Systematized Nomenclature Of Medicine Clinical Terms* (SNOMED CT) [45], una raccolta di nomenclatura medica clinica, 361.042 classi;
- *Gene Ontology* (GO)[46], contiene annotazioni sui geni rispetto alla funzione molecolare, ai componenti cellulari, e al ruolo biologico che ricoprono, 50.515 classi;
- *Protein Ontology* (PR)[47], rappresentazione di entità relative alle proteine mediante definizioni e relazioni esistenti tra di loro, 331.980 classi;
- *Chemical Entities of Biological Interest Ontology* (ChEBI) [48], classificazione di composti chimici di rilevanza biologica, 156.098 classi;

- *Human Phenotype Ontology* (HPO) [49], vocabolario standardizzato di anormalità fenotipiche riscontrate in malattie umane, 19.618 classi e più di 150.00 annotazioni per malattie ereditarie;
- *Human Disease Ontology* (DOID) [50], vocabolario controllato e organizzato gerarchicamente, 17.375 classi;
- *Drug Ontology* (DrOn) [51], comprende una classificazione dei farmaci esistenti e relazioni sulla loro interazione, 578,205 classi.

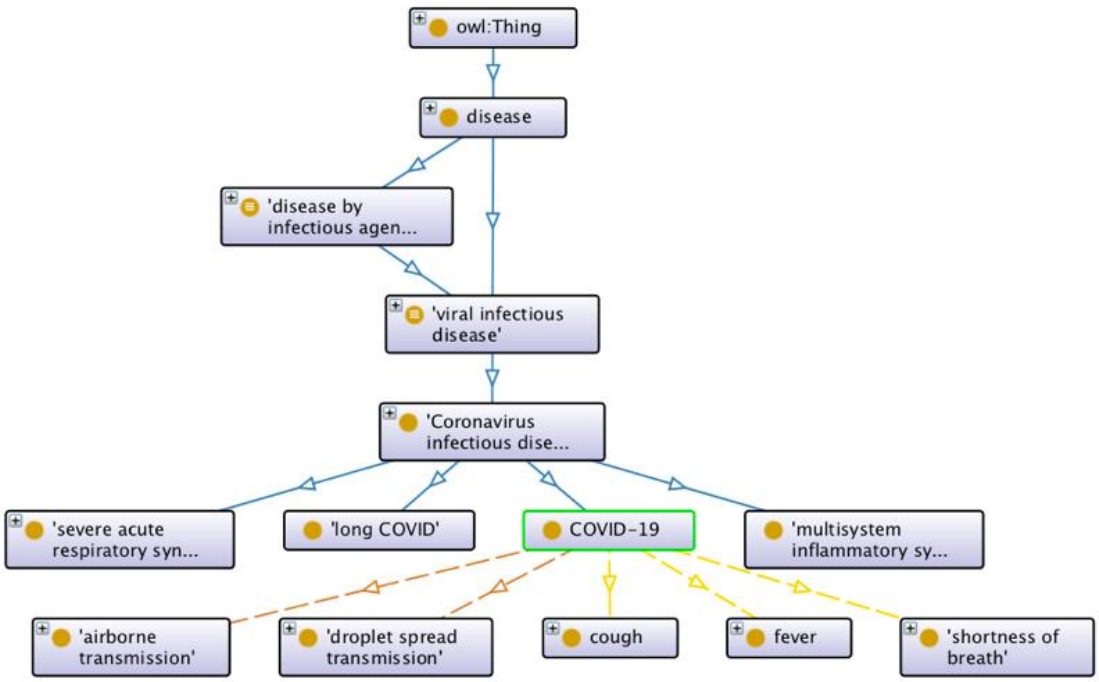


***Figura 11** – Estratto della Human Disease Ontology (DOID) tramite Protégé*[19] *(https://protege.stanford.edu/products.php).*

[19] Editor per ontologie.

5. Applicazioni in ambito Biomedico

Tra i principali temi che si riscontrano in ambito ingegneristico per costruire sistemi basati sulla conoscenza, vi sono problemi tecnici relativi al definire le modalità con cui acquisire, rappresentare ed infine utilizzare in modo appropriato la conoscenza stessa [52].

Ad alimentare la necessità, in ambito biomedico, di un modello quanto più formale per rappresentare la conoscenza vi sono altre questioni principali:

1. un enorme quantitativo di dati e informazioni[20] da analizzare ed interrogare disponibili sotto forma di articoli, brevetti e report scritti [53];
2. una difficoltà intrinseca nel descrivere i meccanismi che stanno alla base di tutti i sistemi biologici;
3. l'incapacità di poter riutilizzare e sviluppare tool e applicazioni già esistenti per mancanza di omogeneità nell'espressione della conoscenza, rendendo difficile o impossibile l'interoperabilità tra i vari sistemi.

Con il rapido sviluppo di nuove tecnologie di sequenziamento, è possibile analizzare e testare un grande quantitativo di dati in un tempo molto limitato [54]. A beneficiarne particolarmente sono campi quali la biologia molecolare (analisi di proteomica), la genetica (sequenziamento del genoma umano), la biochimica e la farmacologia (*drug discovery*). Il prodotto di tali applicazioni in questi ambiti è a sua volta un'enorme quantità di dati eterogenei. Per questi motivi, centinaia di ontologie sono state sviluppate su quasi tutti i domini biologici e biomedici, rappresentando di fatto enormi database in cui dati e informazioni sono concettualizzati. I fenomeni biologici vengono descritti formalmente mediante un vocabolario controllato[21] all'interno di un dominio specifico e collegati ad altri domini vicini, riducendo l'eterogenità e favorendo l'interoperabilità tra sistemi di informazione biomedica. Attraverso il processo di annotazione all'interno delle ontologie, dati e descrizioni di metadati sono identificati da ID unici e

[20] Espressi in formati eterogenei.

[21] Un vocabolario controllato è un elemento basilare di ogni ontologia, mentre la possibilità di collegare entità tra ontologie di domini vicini non è obbligatorio ma generando collegamenti si amplifica il valore stesso delle ontologie creando una rete sempre più ricca di entità e proprietà collegate tra loro.

corrispondenti ad etichette. Le ontologie esplicitano le relazioni trai i tipi di dati nei *database*, provvedendo a una chiara specificazione dei termini usati per esprimere l'informazione biomedica, rendendo possibile alle applicazioni di fare assunzioni tra le classi [55]. Perciò, ontologie come *Gene Ontology* [46], *Human Phenotype Ontology* [56], e *Chemical Entities of Biological Interest* [48], e tante altre risultano particolarmente utili per la creazione di tool e nuovi metodi per l'analisi dell'informazione biomedica [54].

La *Gene Ontology* [46] rappresenta una delle ontologie di maggiore successo. All'interno di un progetto internazionale (*GO annotation* [57]) sono stati estratti mediante le annotazioni dell'ontologia liste di termini (*GO term)* per essere utilizzati da software e tool di vario genere. Ogni annotazione rappresenta una tripla del tipo: *UniProtKB:P04637 "involved in" GO:0045944.* UniProtKB:P04637 rappresenta l'antigene della cellula tumorale p53 per l'essere umano e denota il processo di regolazione positiva della trascrizione dell'RNA polimerasi II (GO:0045944). Con le annotazioni si associano a un termine (*GO term*) la funzione di un particolare gene o di un suo prodotto. Per la sua utilità, un procedimento analogo è stato effettuato utilizzando la *Human Phenotype Ontology* [56] in cui la terminologia che descrive le anomalie fenotipiche riscontate in malattie umane, è usata in combinazione con sistemi per la diagnosi di malattie genetiche rare[22]. Alcuni esempi di *Web* tool che beneficiano dell'estrazione di *terms* [56] dalle annotazioni ontologiche sono:

- *Phenomizer* [58], usa le anomalie fenotipiche per la diagnostica clinica;
- *PhenoGramViz* [59], analizza e grafica le interazioni tra gene e fenotipo;
- *Exomiser* [60], capace di *prioritizzare* variazioni dell'esoma usando un algoritmo di *matching* fenotipico tra le specie (generalmente uomo e topo);
- *Genomiser* [61], analizza le varianti non codificanti dell'genoma umano la cui informazione costituisce ancora una sfida nei casi di sequenziamento del genoma in malattie Mendeliane, specialmente nei

[22] Patologie Mendeliane.

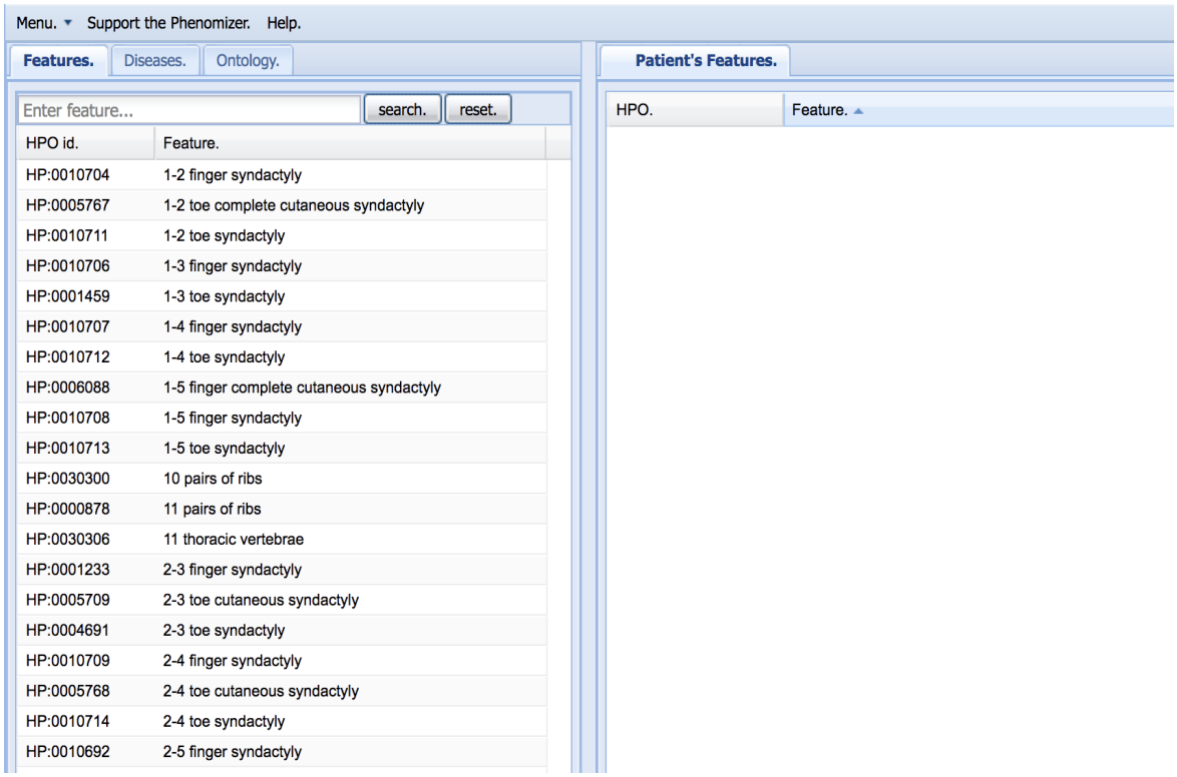
casi di variazione un singolo neuclotide[23] o altre piccole varianti non codificanti.

Il funzionamento di alcuni dei sopracitati strumenti è mostrato in **Figura 12**.

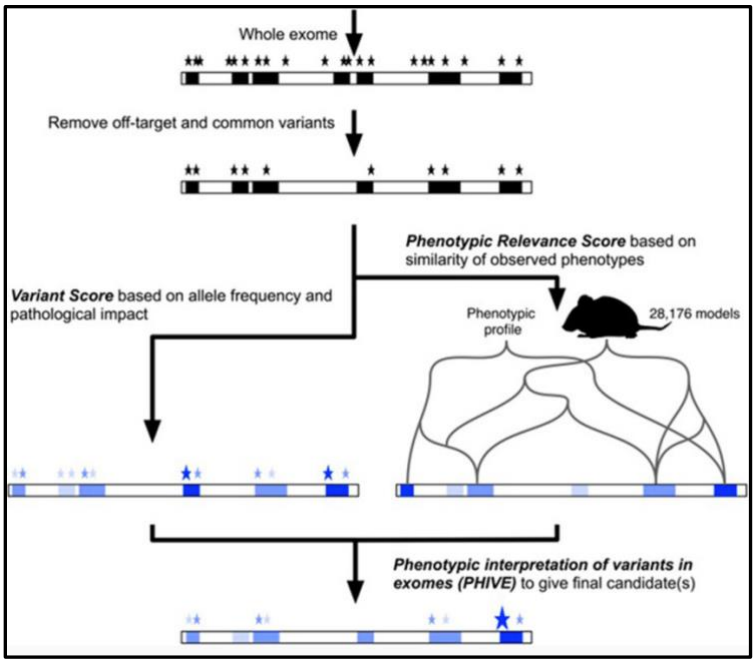
Quindi se da un lato con le ontologie si assume un metodo per formalizzare l'informazione biomedica, dall'altro si genera anche una base di conoscenza interoperabile ed utilizzabile per integrare sistemi di diverso tipo (tool, applicazioni, sistemi di deep learning) che necessitano, in un quadro euristico, di una conoscenza specifica di dominio strutturata. Alcuni interessanti studi mostrano come sia possibile integrare le ontologie all'interno sia di sistemi di analisi, quali modelli di machine learning per *text meaning* [62], misure di somiglianza semantica, e riconoscimento di dati (immagini, audio, colori), sia di *tool* che sfruttino le annotazioni per minimizzare i tempi di attesa in diagnosi e terapie (per esempio *tool* di prioritizzazione genetica) [63]. Nei successivi capitoli verrà presentato un approfondimento su come l'utilizzo di ontologie biomediche a supporto di tecniche di analisi affermate possa portare ad interessanti vantaggi in diversi settori della biomedica.

[23] unità ripetitive costitutive degli acidi nucleici (DNA e RNA).

a)


Menu.
Support the Phenomizer.
Help.
Features.
Diseases.
Ontology.
Patient's Features.
Enter feature...
search.
reset.
HPO id.
Feature.
HP:0010704 1-2 finger syndactyly
HP:0005767 1-2 toe complete cutaneous syndactyly
HP:0010711 1-2 toe syndactyly
HP:0010706 1-3 finger syndactyly
HP:0001459 1-3 toe syndactyly
HP:0010707 1-4 finger syndactyly
HP:0010712 1-4 toe syndactyly
HP:0006088 1-5 finger complete cutaneous syndactyly
HP:0010708 1-5 finger syndactyly
HP:0010713 1-5 toe syndactyly
HP:0030300 10 pairs of ribs
HP:0000878 11 pairs of ribs
HP:0030306 11 thoracic vertebrae
HP:0001233 2-3 finger syndactyly
HP:0005709 2-3 toe cutaneous syndactyly
HP:0004691 2-3 toe syndactyly
HP:0010709 2-4 finger syndactyly
HP:0005768 2-4 toe cutaneous syndactyly
HP:0010714 2-4 toe syndactyly
HP:0010692 2-5 finger syndactyly
HPO.
Feature.

b)


Whole exome
Remove off-target and common variants
Phenotypic Relevance Score based on similarity of observed phenotypes
Variant Score based on allele frequency and pathological impact
Phenotypic profile
28,176 models
Phenotypic interpretation of variants in exomes (PHIVE) to give final candidate(s)

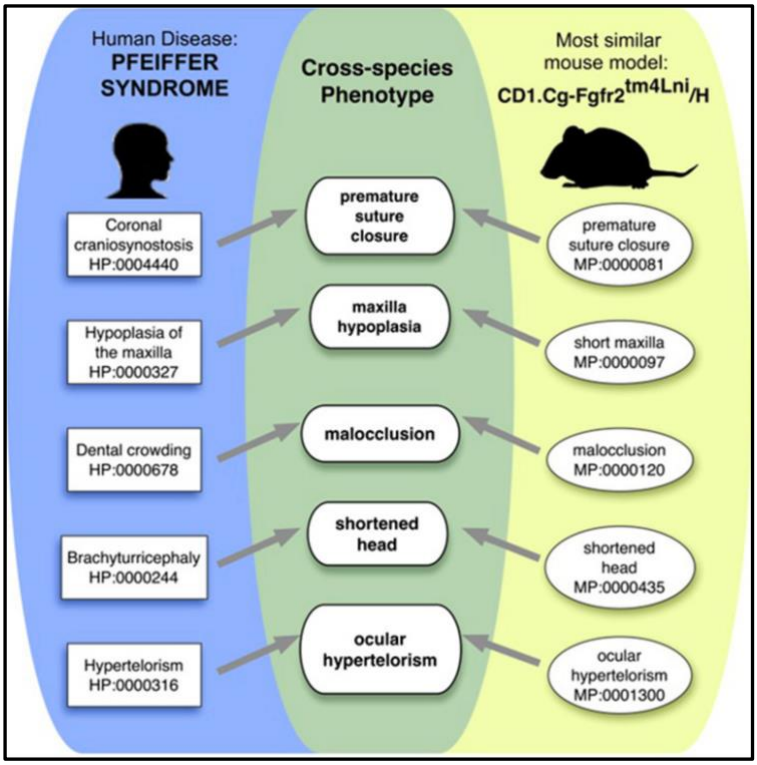
c) 

Figura 12 *- a) Phenomiser: in base ai termini HPO inseriti vengono rilevate il numero e il tipo di malattie ad essi collegate. b)-c) Exomiser: passaggi principali dell'algoritmo di matching, l'utente seleziona un fenotipo umano dalla lista di termini HPO e tutti i geni con varianti che superano la prime fasi di filtraggio vengono confrontati i modelli creati sul topo associando poi l'effetto fenotipico del modello trovato alla malattia umana. (Fonti: https://hpo.jax.org/app/tools/exomiser, https://hpo.jax.org/app/tools/phenomizer).*

5.1 Ontologie Biomediche nei processi di Machine Learning

Il volume di informazioni attualmente esistenti su Web rappresenta un limite umano sia per l'analisi che la ricerca, anche qualora venisse ristretto ad un dominio molto specifico. La letteratura biomedica scritta rappresenta il metodo standard con cui i ricercatori e sviluppatori scambiano dati e nuove scoperte.
Poter analizzare, dunque, la mole di informazioni mediante tecniche di *text meaning*[24] risulta l'unico modo possibile [64]. Tuttavia, molti dei sistemi affermati nello svolgere questo compito risultano deboli sull'aspetto semantico della terminologia e perciò rischiano di risultare ambigui anche nel riconoscere relazioni tra entità. Risulta quindi necessario associare a sistemi di rappresentazione del linguaggio, già allo stato dell'arte[25], strutture semantiche derivanti da risorse esterne di conoscenza, come ontologie specifiche di dominio, rappresentazioni di dati mediante grafi, o misure di somiglianza semantica [53].
Tra le tre forme appena citate di conoscenza, le ontologie sono quelle che risultano più complete. Infatti, sono rappresentabili mediante chiari grafi aciclici in cui si possono distinguere le entità (nodi) e sono relazionate tra loro mediante specifiche proprietà e assiomi (collegamenti), definendo una struttura semantica priva di ambiguità. Per questo motivo si prestano particolarmente come supporto ai modelli machine learning. In particolare, attraverso gli *embeddings* dell'ontologia[26] si possano sfruttare la conoscenza di dominio delle ontologie biomediche, e come gli assiomi[27] possano essere usati come vincoli per l'ottimizzazione dei modelli di machine learning potendo ridurre lo spazio di ricerca [62]. Attualmente questi modelli e metodi sono in costante sviluppo rappresentando una novità nel panorama bioinformatico. La conoscenza di

[24] Tecniche di machine learning atte alla comprensione del significato del testo e alla sua classificazione.

[25] Come BERT, o BioBERT (precedentemente allenato in ambito biomedico), o ELMo.

[26] Per *embeddings* dell'ontologia (Ontology Embeddings) si intendono strutture che preservano la mappatura dell'ontologia all'interno di vettori spaziali e forniscono un importante metodo per l'uso delle ontologie nel machine learning in quanto queste strutture mantengono invariate i diversi elementi che compongono un'ontologia inclusi grafi strutturali, regolarità sintattiche e modelli teorici semantici.

[27] delle ontologie.

dominio può essere utilizzata per vincolare la ricerca e trovare soluzioni ottimali più velocemente, oppure per identificare soluzioni migliori. Questa osservazione portò E. A. Feigenbaum nel 1977 a suggerire che il potere dei sistemi di Intelligenza Artificiale risiedesse proprio nella conoscenza specifica di dominio che essi stessi codificavano [52]. La conoscenza così come è espressa dalle ontologie può essere usata all'interno dei modelli di machine learning per due diversi scopi [62]:

1. espandere ed arricchire le*features* usate;
2. vincolare la ricerca di soluzioni ottimali per problemi di apprendimento.

Nel primo caso, la possibilità di ampliare il gruppo delle *features* permette di fornire informazioni ai modelli di *machine learning* a cui probabilmente non avrebbero accesso se non mediante le ontologie. Le ontologie biomediche rappresentano quindi una conoscenza supplementare a quella fornita nelle fasi di allenamento degli algoritmi, e risultano essenziali qualora il dominio specifico in questione sia povero di dati già classificati e catalogati [65]. Ontologie biomediche come *Gene Ontology* (GO) [46] e *Human Phenotype Ontology* (HPO) [56] contengono ognuna migliaia di termini e annotazioni e possono apportare una grande quantità di informazioni utili per riconoscere ed estrarre relazioni dai testi [53]. Per esempio, il poter collegare fenotipi come "cardiomiopatia" alle strutture anatomiche che sono affette (il "cuore") crea nuove e dirette associazioni con altri dataset che altrimenti non esisterebbero. In particolare, nel precedente esempio della cardiomiopatia, il legame con il cuore in quanto struttura anatomica può essere usato per collegare il fenotipo all'espressione genetica nel tessuto cardiaco o nei cardiomiciti. Questo vincolo è fornito a priori mediante gli assiomi dell'ontologie relative al fenotipo, all'anatomia, alle cellule. Per inferenza sulle ontologie il vincolo non necessita di essere scoperto ma viene gratuitamente fornito dalle strutture assiomatiche che caratterizzano le ontologie stesse [62].

Un secondo modo per poter usare la conoscenza espressa tramite ontologie è la possibilità di poter vincolare la ricerca di soluzioni per problemi di ottimizzazione. Un semplice esempio del ruolo che questi vincoli rappresentano è il seguente: se il prodotto di un gene G può potenzialmente essere coinvolto in un processo P_1

e ogni processo P_1 è parte di un processo P_2, allora G è anche coinvolto nel processo P_2. Questo vincolo è "rigido" in quanto non è né una legge né un'osservazione di tipo empirico, ma esiste in virtù delle definizioni di P_1 e P_2 , per cui per G sarebbe impossibile essere parte di un solo processo e non di entrambi [62]. Il fatto stesso di appartenenza a uno o più "processi" nelle ontologie è possibile anche solo definendo le classi al suo interno: un gene coinvolto nello *sviluppo di crescita cellulare* (GO:0048588) è coinvolto anche nello *sviluppo cellulare* (GO:0048468) semplicemente avendo definito classe e sottoclasse nella medesima ontologia, *Gene Ontology* [46].

Metodi e modi di combinazioni di machine learning con ontologie biomediche mediante tool e applicazioni sono riassunti in **Tabella 4**.

Type	Method/Tool	Description
Processing and preprocessing ontologies	OWLAPI	Reference library to process OWL ontologies, supports most OWL reasoners.
	funowl	Python library to process OWL ontologies.
	owlready2	Python library to process OWL ontologies.
	Apache Jena	RDF library with OWL support.
	rdflib	Python RDF library with OWL support.
	Protégé	Ontology editor and knowledge engineering environment.
Computing entailments, reasoning	ELK	Very fast reasoner for the OWL 2 EL profile with polynomial worst-case time complexity.
	HermiT	Automated reasoner supporting most of OWL axioms with exponential worst-case complexity.
	Pellet	OWL reasoner supporting most of the OWL constructs and supporting several additional features.
Generating graphs from ontologies	OBOGraphs	Syntactic conversion of ontologies to graphs, targeted at OBO ontologies
	Onto2Graph	Semantic conversion of OWL ontologies to graphs, following the axiom patterns of the OBO Relation Ontology.
Computing Semantic Similarity	Semantic	Comprehensive Java library to compute semantic similarity measures over ontologies.
	Measures Library sematch	Python library to compute semantic similarity on knowledge graphs.
	DiShIn	Python library for semantic similarity on ontologies.

Embedding graphs	OWL2Vec, DL2Vec	Method that combines generation of graphs from ontologies, random walks on the generated graphs, and generation of embeddings using Word2Vec. Syntactically processes most OWL axioms.
	RDF2Vec	Method to embed RDF graphs.
	Node2Vec	Method to embed graphs using biased random walks.
	Walking RDF&OWL	Method that combines generation of graphs from ontologies, random walks on the generated graphs, and generation of embeddings using Word2Vec. Only considers the ontology taxonomy.
	PyKEEN, BioKEEN,	Toolkit for generating knowledge graph embeddings using several different approaches.
	OpenKE	Library and toolkit for generating knowledge graph embeddings.
	PyTorch Geometric	Library for graph neural networks which can be used to generate graph embeddings.
Embedding axioms	Onto2Vec	Embeddings based on treating logical axioms as a text corpus.
	OPA2Vec	Embeddings that combine logical axioms with annotation properties and the literature.
	EL Embed- dings	Embeddings that approximate the interpretation function and preserve semantics for intersection, existential quantifiers, and bottom.
Ontology-based constrained learning	DeepGO	Implements an ontology-based hierarchical classifier for function prediction. The hierarchical classification module is generic and can be used with other ontologies and applications.
	DEEPred	Automated Protein Function Prediction with Multitask Feed-forward Deep Neural Networks.
	DeepMiR2GO	Inferring Functions of Human MicroRNAs Using a Deep MultiLabel Classification Model.

Tabella 4 *- Una panoramica di software, tool ed applicazioni coinvolti nei sistemi di machine learing integrati con ontologie biomediche*.

(Fonte: http://biorxiv.org/lookup/doi/10.1101/2020.05.07.082164).

5.1.1 Modelli Predittivi di Machine Learning e Deep Learning

I modelli di machine learning più comunemente legati al mondo scientifico in senso lato sono di tipo predittivo, ovvero sistemi in grado di rispondere a problemi di ottimizzazione ricercando soluzioni migliori per funzioni descritte in uno spazio continuo o discreto. I modelli predittivi sono il risultato di un'applicazione di *machine learning* conosciuta come apprendimento supervisionato [66].

L'obbiettivo dell'apprendimento supervisionato è riuscire a prendere decisioni sequenziali usando una fase di allenamento, *training*. Nella fase di *training* vengono passati ad un algoritmo sia una grande quantità di dati (che l'algoritmo deve analizzare) sia i risultati che ci si aspetta che l'algoritmo restituisca per quegli specifici input. In questo modo l'algoritmo vede sia i dati, sia la risposta e si deve allenare per arrivare il più possibile vicino a quella che è la risposta corretta. Per far ciò, vengono assegnati dei pesi[28] ai vari input in ingresso inizialmente a caso: l'uscita risultante in prima analisi sarà errata perché frutto di un assegnamento di pesi randomico. Viene poi calcolata la differenza tra l'output restituito e quello atteso noto: questa differenza viene utilizzata per modificare e correggere i pesi. L'algoritmo nei successivi calcoli fornirà uscite sempre più precise con pesi più accurati, fino a restituire uscite molto vicine a quelle attese [66]. Una volta concluso il *training*, i valori dei pesi vengono salvati in un modello, ovvero il risultato finale dell'addestramento. L'algoritmo addestrato potrà adesso essere interrogato con nuovi dati di cui vogliamo ottenere e conoscere la risposta. L'apprendimento supervisionato è di tipo *task-driven,* è perciò capace di eseguire uno specifico compito predicendo classi (attività di classificazione) o variabili continue[29] (attività di regressione).

Le applicazioni relative alla capacità di classificazione includono [66]:

- validazione di firma;
- riconoscimento facciale o di oggetti;
- individuazione di frodi secondo pattern di riconoscimento.

[28] Ovvero quanto pesa l'input 1,2,3…n sull'output

[29] A seconda di come viene allenata la rete in ambiente supervisionato quello che si può predire in output sono o risposte di tipo variabile continua (il cui valore è un insieme di numeri o un intervallo di numeri reali) o di tipo classe (variabile categorica i cui valori possibili è costituito da un numero finito di categorie, esprimibile con un array). La nostra rete può quindi fare una classificazione, cioè stimare se un certo dato appartiene a una classe piuttosto che ad un'altra.

La regressione, invece, è principalmente coinvolta in gestione di previsione come:

- pronostici sui prezzi di mercato;
- domanda sui prodotti;
- dati ambientali;
- processi di ottimizzazione dei parametri.

Questi sistemi si inseriscono all'interno del panorama scientifico principalmente con due scopi:

- analizzare ed interrogare un enorme quantitativo di dati (articoli scientifici, brevetti, report);
- comprendere i complessi meccanismi che stanno alla base dei sistemi biologici mediante specifiche tecniche di *relation extraction*.

La *relation extraction* o *relation classification* è il processo tramite cui si stabiliscono le relazioni (se ci sono) tra le entità trovate all'interno per esempio di un articolo scientifico [67].In questo caso le *features* coinvolte passate come input al modello di machine learning possono essere l'entità sorgente[30], l'entità di destinazione[31], simboli interni, frasi intere, precedentemente riconosciute mediante una fase preliminare di riconoscimento delle entità[32] e delle possibili relazioni. Dunque, il risultato del processo di estrazione in uscita dal modello è una lista di entità riconosciute e una di relazioni stabilite tra di esse. Un ramo del *machine learning* molto in uso e sviluppo negli ultimi decenni è il *deep learning,* una rete che emula il funzionamento neurale di alto livello del cervello umano interponendo una serie di *hidden layer* intermedi tra input e output del sistema [68]. Parlare di *deep learning* soprattutto correlato alle ontologie è particolarmente importante perché differentemente dal *machine learning* in cui le *features* devono essere selezionate manualmente, nel *deep learning* è esso stesso a sceglierle. Questo significa che applicando modelli di *deep learning* per *relation extraction* associati ad ontologie biomediche esterne potrebbe permette di estrarre relazioni ignote e ampliare la conoscenza biomedica. Inoltre, rispetto al *machine learning,* è in grado di poter usare un grande quantitativo di features

[30] sorgente della relazione

[31] destinazione della relazione (se esiste) stabilita dalla *relation extraction* stessa

[32] *Entity Recognition*

e variabili in quanto il passaggio di estrazione delle caratteristiche più utili e significative viene effettuato dal modello stesso. I sistemi di *deep learning* sono in grado di analizzare e predire con un'accuratezza ben più ampia e veloce di quella umana: eventuali complessità aggiuntive provenienti dall'ambiente sono percepiti come ulteriori input e il sistema si adatta continuamente nel rispondere al problema per cui "tutto influenza tutto" [69].
Due reti di *deep learning* note sono:

- RNN (*Recurrent Neural Network*): reti neuronali che analizzano ricorsivamente sequenze che possono essere discorsi o testi [70].

- LSTM (*Long-Short Term Memory*): è un tipo di rete neurale ricorsiva e si basa sul tenere memoria degli input anche a lungo termine, permettendo di stabilire dipendenze più complesse ed eliminare tutto ciò che risulta ininfluente al processo di predizione (congiunzioni, articoli, simboli) [71].

5.1.2 BiOnt e BO-LSTM

La descrizione dei meccanismi responsabili del comportamento dei sistemi biologici è qualcosa di estremamente complesso e ogni passo verso la comprensione di questi sistemi costituisce a tutti gli effetti un traguardo per la scienza. Per esempio, descrivere malattie associate con i meccanismi che originano anomalie fenotipiche come risultato di una modifica dell'espressione genetica, oppure descrivere l'azione di medicinali in queste malattie sono attualmente risultati non banali da ottenere. Un metodo altamente efficace per poter comprendere i meccanismi intrinsechi dei sistemi biologici è estrarre e classificare le relazioni che esistono tra diverse entità biomediche come composti chimici, malattie, geni e fenotipi [65].
In letteratura questo è solitamente un compito di *relation extraction* che, come abbiamo visto nel capitolo precedente, è una tecnica di machine learning in grado di estrarre entità e classificare relazioni da articoli scientifici o medici. Una peculiare caratteristica della *relation extraction* risiede nella sua capacità di portare alla luce associazioni sconosciute tra entità biomediche estremamente utili per ricercatori e medici. La maggior parte di questi sistemi non ricorrono a risorse esterne di conoscenza, tuttavia, negli ultimi anni sono stati sviluppati innovativi metodi di *deep learning* integrati con ontologie biomediche capaci di fronteggiare problemi comuni di *relation extraction*. In generale i sistemi di apprendimento (come il *word embedding*[33]) possono imparare come riconoscere le relazioni tra le entità ma trovano difficoltà nel cogliere la semantica e lo specifico dominio di ogni entità [53]. L'aggiunta delle ontologie biomediche come *features* in un classificatore permette il raggiungimento di risultati più accurati.

BO-LSTM e BiOnt sono due esempi di rete neurale LSTM bidirezionale che usano Gene Ontology (GO) [46], Human Phenotype Ontology (HPO) [56],

[33] Il word embedding permette di memorizzare le informazioni sia semantiche che sintattiche delle parole costruendo uno spazio vettoriale in cui i vettori delle parole sono più vicini quanto più le parole sono semanticamente simili.

Chemical Entities of Biological Interest (ChEBI) [48] e Human Desease Ontology (DOID[34]) [50], come conoscenza integrata nei processi di *relation extraction*. In **Figura** 13 si possono vedere combinate le quattro ontologie in BiOnt con lo scopo di estrarre le relazioni di dieci diverse combinazioni di entità biomediche, diversamente dalle due sole coppie di relazioni che si ottengono con BO-LSTM [72].

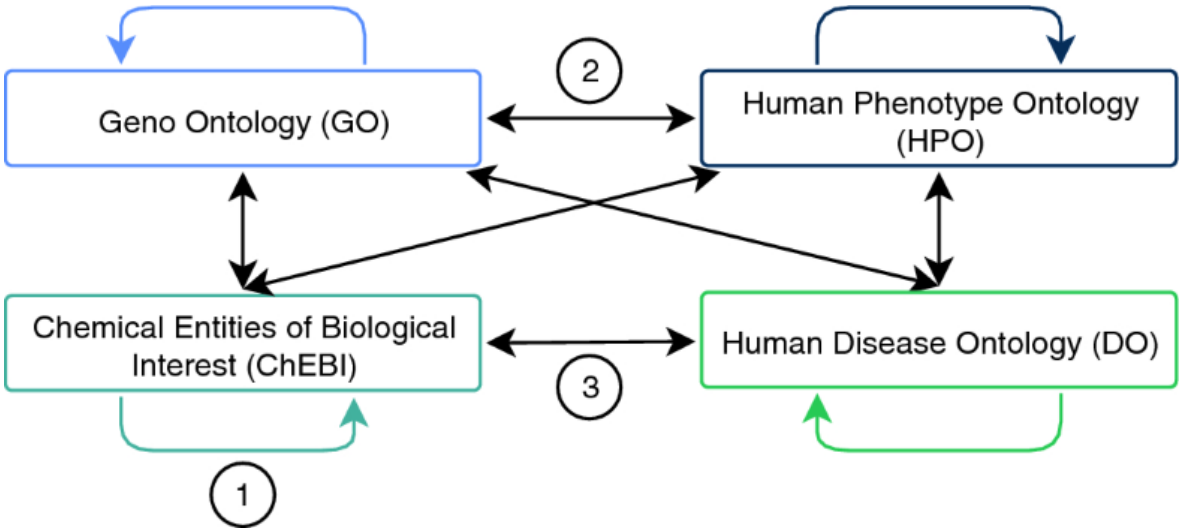


Figura 13 *- Le dieci possibili combinazioni tra le quattro ontologie biomediche. I numeri 1,2,3, rappresentano rispettivamente i tre corpus usati per testare le capacità di BiOnt: DDI, PGR e BC5CDR (Fonte: https://link.springer.com/chapter/10.1007%2F978-3-030-45442-5_46#Sec2)*

Sebbene i due progetti usino un numero di ontologie integrate simili, non sono in grado di estrarre lo stesso numero di relazioni. Con il sistema BO-LSTM si possono estrarre relazioni solo di tipo *drug-drug* e *phenotype-gene*, mentre con il BiOnt si possono classificare fino a quattro diverse interazioni: *drug-drug*, *phenotype-gene*, *chemical-induced disease* rispettivamente con un miglioramento, in termini di F-score, del 4.93%,4.99% e 2,21% in più rispetto allo stato dell'arte [72].

[34] DOID presente solo nel progetto BiOnt

Questi risultati sono stati ricavati mediante l'utilizzo di tre dataset allo stato dell'arte (come indicato in

Figura 13) creati *ad hoc* per valutare questo sistema integrato in mancanza di un corpo gold standard di riferimento:

- *Drug - Drug Interactions* (DDI): corpus che descrive interazioni farmaco-farmaco (5028 interazioni) focalizzate sull'aspetto farmacocinetico e farmacodinamico, totalmente annotato a mano;
- *Phenotype - Gene Relations* (PGR): corpus che descrive relazioni fenotipo-gene umano (4283 relazioni), creato in modo totalmente automatico[35];
- *Chemical- Iduced Disease Relations* (BC5CDR): il Corpus BioCreative V CDR è un corpus di relazioni in malattie chimicamente indotte (3116 relazioni) annotate da articoli su PubMed.

Il sistema BiOnt è stato sviluppato a partire BO-LSTM e come quest'ultimo incorpora al suo interno la rete Word2Vec[36] per il *word embedding.* Il *word embedding* permette di rappresentare una frase di lunghezza variabile in un vettore la cui lunghezza è fissa e ogni elemento del vettore codifica la semantica della frase originale. L'innovazione di BiOnt e BO-LSTM risiede nell'associare alle entità riconosciute in ogni vettore la semantica delle ontologie biomediche [53]. Sfruttare gli *embeddings* di più ontologie come uno dei primari strati di conoscenza in processi di *relation extraction* permette di generare candidate relazioni ignote e indirizzare la ricerca verso percorsi più plausibili [72] **(Figura 14**). Infatti, rappresentando ogni entità come sequenza dei suoi predecessori, è possibile riconoscere nuove relazioni non evidenti tra entità usando solo i dati della fase di *training*.

[35] Le aspettative di correttezza diminuiscono rispetto a un corpus totalmente annotato a mano.

[36] Word2Vec è una rete neurale artificiale a due strati progettata per elaborare il linguaggio naturale, l'algoritmo richiede in ingresso un corpus e restituisce un insieme di vettori che rappresentano la distribuzione semantica delle parole nel testo. Per ogni parola contenuta nel corpus, in modo univoco, viene costruito un vettore in modo da rappresentarla come un punto nello spazio multidimensionale creato. In questo spazio le parole saranno più vicine se riconosciute come semanticamente più simili.

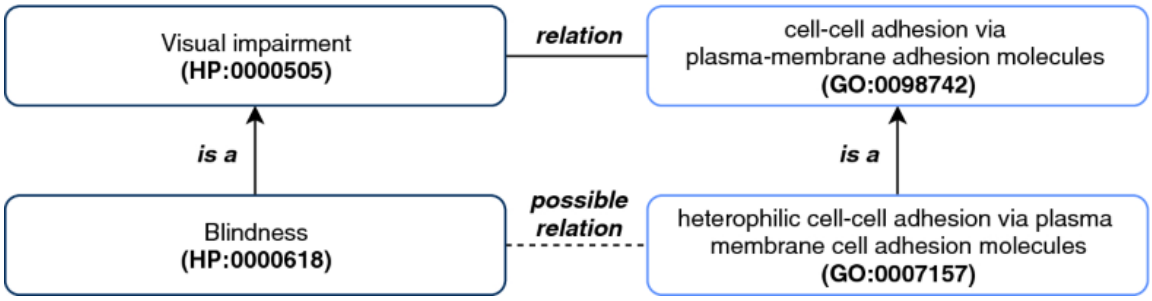


Figura 14 *- Esempio di ontology embedding in BiOnt basato sulle ontologie HPO e GO, per la relazione candidata tra il fenotipo umano "cecità" e il gene CRB1 (rappresentato dal termine GO:0007157 " heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules"). (Fonte: https://link.springer.com/chapter/10.1007%2F978-3-030-45442-5_46#Sec2).*

Per esempio, la dopamina (CHEBI:18243), un composto chimico con diversi ruoli importanti su corpo e cervello, può essere caratterizzata nell'essere una catecolamina (CHEBI:33567), un composto alchilamminico (CHEBI:64365) e un composto organico aromatico (CHEBI:33659) (**Figura 15**). Nel predire una certa interazione di una medicina con la dopamina, i suoi predecessori possono aggiungere informazioni in più, non necessariamente espresse in un testo [65].

Per poter estrarre nuovi tipi di relazioni dalle quattro ontologie a favore della diversità, BiOnt introduce una novità rispetto al precedente modello BO-LSTM, permettendo la concatenazione di canali di *ancestors.* Così facendo da un canale comune di predecessori si possono ricavare relazioni tra entità biomediche dello stesso tipo, mentre dalla loro concatenazione si possono estrarre relazioni tra differenti entità biomediche provenienti dalle diverse ontologie coinvolte.

I risultati di questi sistemi integrati per ogni *dataset* si possono osservare in **Tabella 5.** La rete di *deep learning* risulta più performante quando è supportata dalla conoscenza ontologica ("+ *Ontologies"* in Tabella 5), rispetto all'uso di sistemi allo stato dell'arte quali *word embeddings* o WorldNet ("State-of-the-art" in Tabella 5). Sebbene il numero di relazioni corrette identificate sia aumentato per tutti e tre i corpus con l'uso delle ontologie, per il DDI corpus si nota una piccola perdita di precisione, assente invece negli altri due corpus in cui si riscontrano solo valori positivi [72].

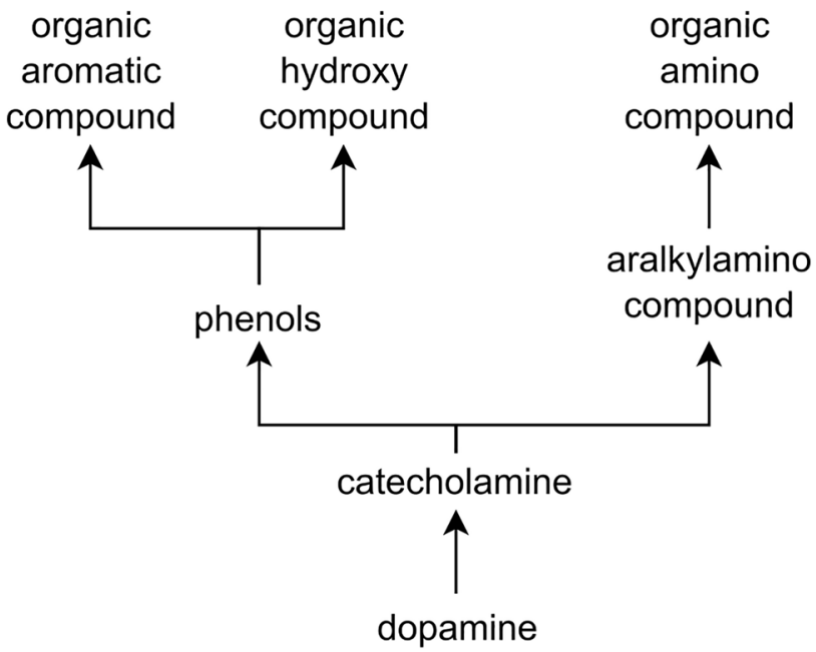


Figura 15 *- Un estratto dell'ontologia ChEBI che mostra I primi predecessori della dopamina usando la sola relazione "is-a".*
(Fonte: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2584-5).

Data Set	Configuration	Precision	Recall	F-score
DDI Corpus	State-of-the-art	0.7134	0.6410	0.6753
	+ Ontologies	0.6784	0.7775	**0.7246**
PGR Corpus	State-of-the-art	0.8421	0.6666	0.7442
	+ Ontologies	0.8438	0.7500	**0.7941**
BC5CDR Corpus	State-of-the-art	0.5371	0.7264	0.6175
	+ Ontologies	0.5770	0.7173	**0.6396**

Tabella 5 - *Risultati di Relation Extraction con il sistema BiOnt (+ Ontologies) confrontato con lo Stato dell'arte, per ogni corpus : DDI per le interazioni drug – drug, PGR per le relazioni phenotype- gene, BC5CDR per le relazioni chemical - induced disease. (Fonte: https://link.springer.com/chapter/10.1007%2F978-3-030-45442-5_46#Sec2)*

5.3 HPO e Tool di Prioritizzazione Genetica

L'intero contenuto genetico del DNA umano ammonta a più di 3 miliardi di nucleotidi [73]. I sistemi tradizionali di diagnosi di una presunta patologia monogenica si basano su test di sequenziamento di nuova generazione (NGS), di cui fanno parte il sequenziamento dell'intero genoma (WGS) e dell'intero esoma (WES). Il range diagnostico del sequenziamento dell'esoma va circa dal 25% al 51%, dimostrando di essere una scelta ottimale nei primi passi verso la diagnosi [74]. Tuttavia, sequenziando solo le regioni codificanti[37] (esomiche), rimangono escluse dall'analisi le regioni intergeniche (regioni non codificanti composte da introni e zone tra gene e gene) che compongono il 98% del DNA. Si stima che, sebbene l'85% delle variazioni che generano una malattia nei tratti dell'esoma, alcune mutazioni patogene possono cadere anche in regioni introniche profonde (come nel caso di distrofia retinica) o in regioni a monte o a valle dei geni a volte non evidenziabili dall'analisi dell'esoma [75]. Non tutte le varianti sono patogene e definiscono semplicemente la variabilità fenotipica interindividuali. L'espressione fenotipica[38] è il risultato dall'interazione del patrimonio genetico con l'ambiente e da un fattore di casualità[39][76]. Analizzare l'enorme quantitativo di dati sequenziali generati dall'analisi dell'intero genoma è ancora ad oggi una sfida, in termini sia di risorse che di tempo, ed uno ostacolo all'utilizzo clinico [74]. Una conoscenza preliminare biomedica come informazioni fenotipiche esterne possono aiutare l'individualizzazione dei geni che contribuiscono al manifestarsi di una malattia [63]. Dalla necessità di migliorare i sistemi diagnostici di prioritizzazione genetica sono stati sviluppati molti tool che combinano l'utilizzo dei termini contenuti nella *Human Phenotype Ontology* (*HPO terms*) per effettuare una ricerca su grandi database disponibili online come *OMIM*[40], *Phenolyzer*, *Mendelian* [77]. Alcuni di questi tool computazionali che utilizzano termini HPO sono: Phen-Gen, eXtasy, PhenIX,

[37] Gli esoni sono le regioni codificanti di ogni gene.

[38] Manifestazione osservabile fisicamente del genotipo.

[39] Relativo ad eventi che possono verificarsi in modo casuale durante lo sviluppo.

[40] Online Mendelian Inheritance in Man.

Exomiser, Phenomizer, AMELIE 2, DeepPVP, GADO, Phenolyzer, Phen2Gene [78].

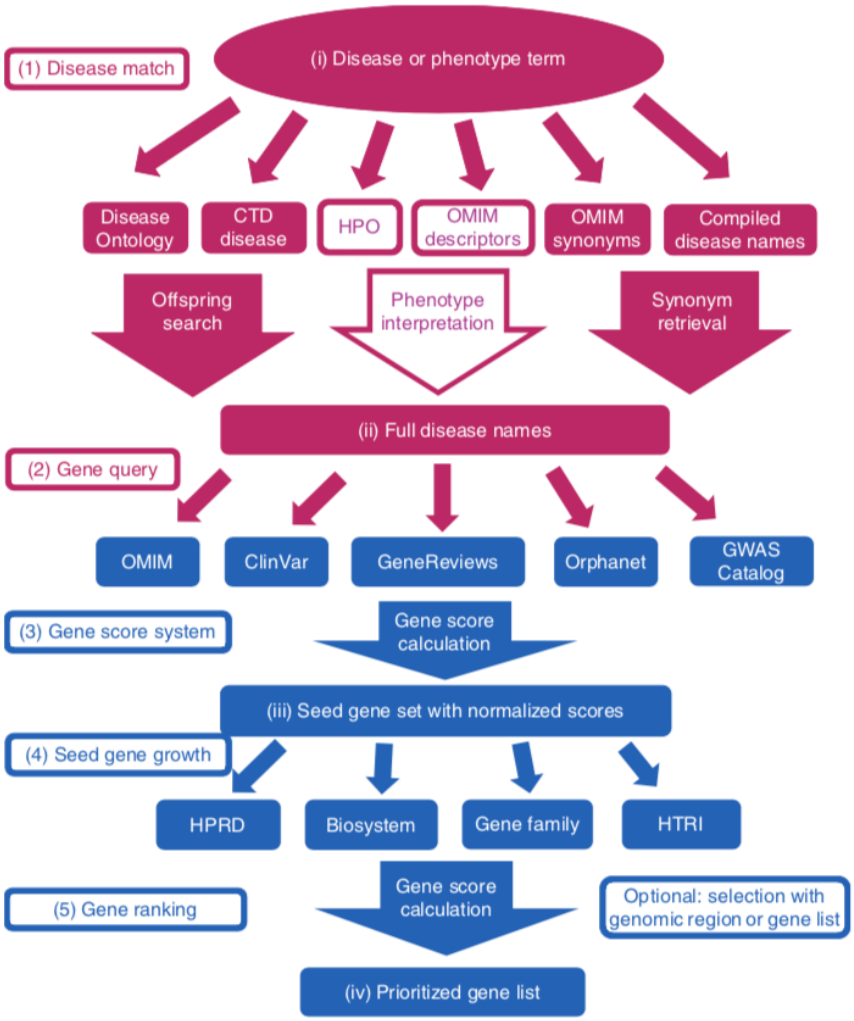


Figura 16 *- Esempio di prioritizzazione genetica mediante Phenolyzer. (1) Disease match: ogni termine relativo alla malattia o al fenotipo sono separatamente tradotti in una serie di nomi di malattie mediante matching delle parole, ricerca per discendenti, sinonimi e interpretazioni fenotipiche in nomi di malattie nei database. (2) Gene query: ogni nome di malattia riconosciuto è interrogato nei database gene-malattia per ottenere una lista di geni. (3) Gene score system: viene generato un punteggio per ogni gene corrispondente ad ogni nome di malattia, basato sul tipo e la confidenza della relazione gene-malattia. Poi, il loro punteggio viene normalizzato. (4) Seed gene growth: la selezione dei geni candidati è terminata, la priorizzazione avviene rapportandoli a 4 diversi dataset che esprimono relazioni di tipo gene-gene. (5) Gene ranking: tutte le informazioni acquisite permettono di associare ad ogni gene di ogni lista un peso finale e di prioritizzare i geni che con più probabilità sono coinvolti. (Fonte: https://www.nature.com/articles/nmeth.3484).*

Tutti questi sistemi richiedono l'inserimento di note cliniche, sintomi e segni, o direttamente termini HPO, e in alcuni casi anche l'analisi di geni, varianti o entrambi per poter generare una lista prioritizzata di geni candidati. Inoltre, fanno riferimento a database di conoscenza diversi. La forza di questi strumenti risiede nell'utilizzo di una robusta ontologia (HPO) per trasformare le osservazioni mediche effettuate sul paziente in termini concettualizzati di informazione usabili da sistemi automatici. Le annotazioni della Human Phenotype Ontology sono aggiornate frequentemente e possono provvedere a una conoscenza fenotipica dettagliata in molte malattie umane. Come ontologia, la HPO permette inferenze computazionali e algoritmi sofisticati che combinano l'analisi genomica a quella fenotipica [78]. I termini HPO sono mappati per trovare geni causali con relazioni binarie. Dei tool citati in precedenza solo Phen2Gene è in grado di leggere direttamente i termini HPO, le altre applicazioni devono convertire il linguaggio naturale in termini HPO tramite sistemi di elaborazione del linguaggio naturale (NLP) [63](**Figura 17**). Tuttavia, studi in merito riportano grandi assonanze nei risultati; se in alcuni i risultati sono interessanti in altri deludenti. Scarsi risultati sono stati riscontrati in studi in cui erano stati esclusi dalla valutazione soggetti con gestalt fenotipici specifici (per cui una sola descrizione fenotipica sarebbe stata sufficiente in una fedele prioritizzazione) [77].
Altri limiti riscontrati nei risultati si suggerisce siano legati:

- alla consistenza dei database online utilizzati dai tool di prioritizzazione per generare le liste di geni candidati;
- alla presenza o meno nei campioni di studio di anomalie congenite multiple[41].

[41] Questi sistemi sono pensati per il riconoscimento di malattie Mendeliane, in cui la variazione che genera l'anomalia fenotipica è localizzata in un solo gene. Nelle malattie congenite multiple i siti di variazione genetica sono multipli, e dunque questi sistemi potrebbero non essere i migliori.

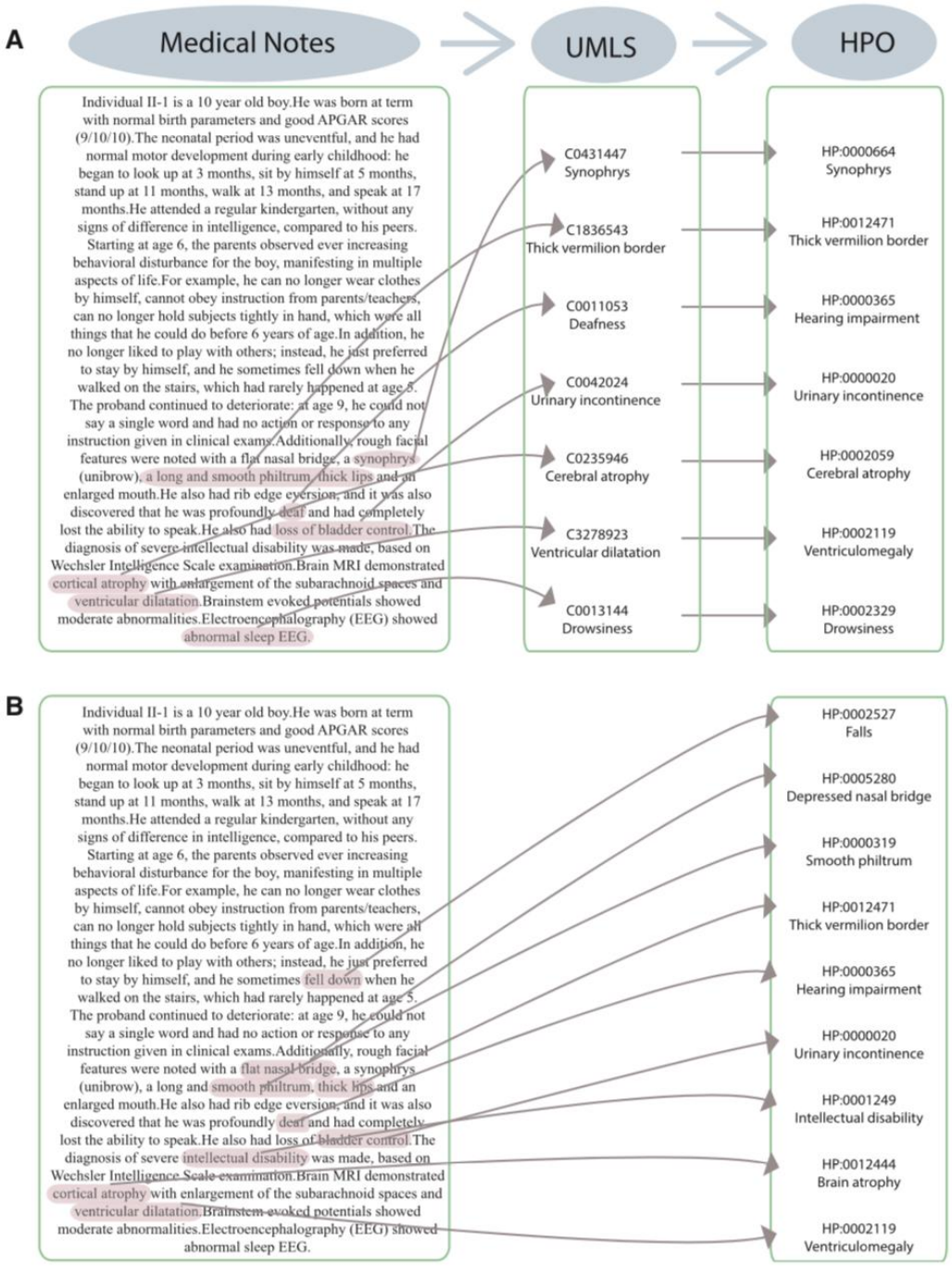


Figura 17 *- A) e B) rappresentano due modi di estrazione di termini HPO. Nel primo caso si sfrutta un'altra ontologia UMLS, un enorme dizionario di termi medici. Nel secondo caso la conversione è diretta.*

(Fonte: https://www.sciencedirect.com/science/article/pii/S000292971830171X).

5.3.1 Phen2Gene

Phen2Gene è uno strumento per la prioritizzazione genetica rapida per malattie rare basata sul fenotipo, decisamente più innovativo rispetto ai precedenti. Infatti:

- sfrutta una versione di Phenolyzer potenziato;
- è in grado di leggere in ingresso sia dati in linguaggio naturale che termini HPO;
- genera una lista di geni candidati usando un database pre-computato;
- non richiede preliminarmente la conoscenza del sequenziamento genetico;
- specifico per il riconoscimento di malattie rare;
- impiega in media 0.94 secondi per elaborare il risultato.

Diagnosi rapide e accurate di malattie Mendeliane[42] sono fondamentali per per incrementare la precisione medica e individuare in modo tempestivo terapie migliori. Il funzionamento di Phen2Gene è riassunto in **Figura 18.** Le note sul paziente scritte dal dottore possono essere in linguaggio naturale, tradotte poi da uno specifico tool Doc2HPO in termini HPO, o direttamente in termini specifici [64].

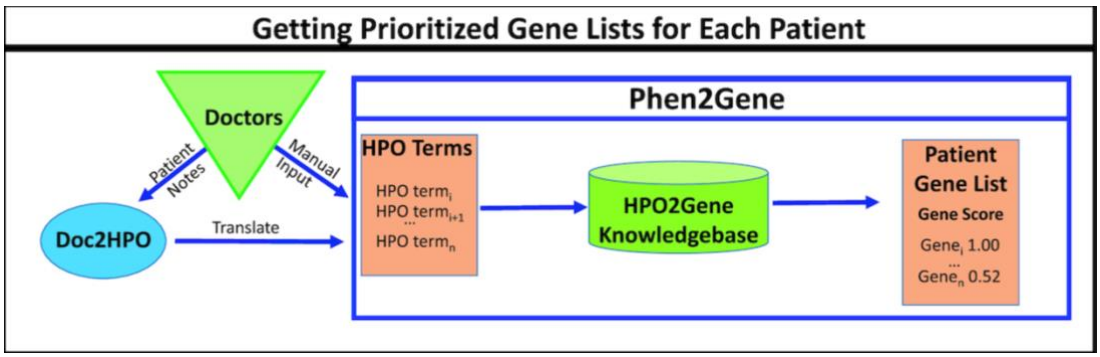


Figura 18 *- Schema riassuntivo del funzionamento di Phen2Gene.*
(Fonte: https://doi.org/10.1093/nargab/lqaa032).

In particolare, Phen2Gene riconosce tutti i termini HPO sotto la radice 'Phenotypic abnormality' (HP:0000118). Uno dei passaggi fondamentali che

[42] Riconducibili a variazioni di un unico gene.

rende questo strumento davvero efficiente è la *HPO2Gene Knowledge base* (**Figura 19**). Per migliorare la diagnosi di malattie rare le annotazioni HPO sono state incorporate con i databases gene-malattia e gene-gene[43], e con un modello probabilistico. Per l'elaborazione di questa base di conoscenza è stato usato *Enhanced Phenolyzer*, una versione potenziata del precedentemente citato Phenolyzer, capace di elaborare per ogni termine HPO una lista di candidati geni [78]. In questa fase, vengono associati ogni termine HPO:

- una lista di geni causali candidati, poi salvata nel HPO2Gene Knowledge base (H2GKB);
- un peso che tiene conto della specificità del contenuto informativo sul fenotipo[44].

Le liste sono stoccate all'interno della H2GKB. I brevi tempi di risposta agli input sono legati all'avere questa parte pre-computata. Per ogni serie di termini HPO dati in ingresso (1,2,3,….,n) il punteggio dei geni candidati è basato sul peso associato ai termini HPO e la loro ricorrenza negli n-termini. All'interno della lista prioritizzata il punteggio che i vari geni assumono è un valore compreso tra 0 e 1 e sono elencati in ordine discendente.

Per provare l'efficienza di questo strumento di prioritizzazione in una situazione realistica è stata effettuata un'analisi retrospettiva su un precedente caso pubblicato [63]. Un soggetto probando[45], è sospettato di avere una malattia Mendeliana. Dai precedenti studi è noto che egli sia affetto dalla Sindrome KBG relativa alla mutazione del nucleotide *ankyrin repeat domain 11* (ANKRD11).

Sono stati messi a confronto Phen2Gene, Phenolyzer, Exomiser, DeepPVP (**Figura 20**). Exomiser e DeepPVP necessitano come prerequisiti sia delle note cliniche che dell'analisi esomica del paziente. Le note cliniche del dottore sulle condizioni fenotipiche del paziente sono state convertite in termini HPO come input per Phen2 Gene e Phenolyzer.

[43] I molteplici database usati hanno tutti strette relazioni con le ontologie. OMIM per esempio è una parte di conoscenza usata per la costruzione stessa della HPO, ma anche della DOID.

[44] Un termine HPO più generico avra meno peso di uno più specifico (contenuto informativo maggiore).

[45] Primo elemento della famiglia a manifestare una malattia dovuta a variazione genetica.

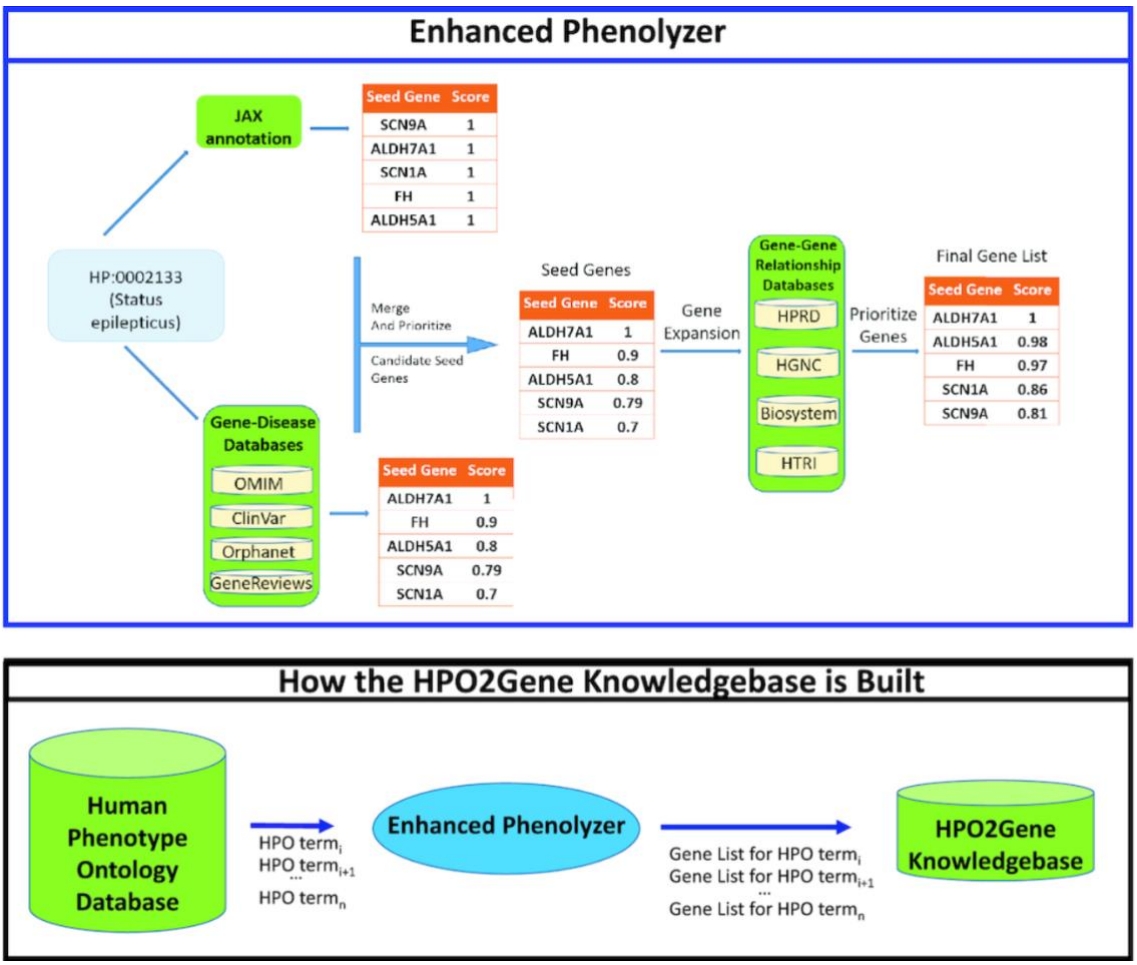


Figura 19 *- Schema rappresentativo della generazione della H2GKB mediante Enhanced Phenolyzer (Fonte: https://doi.org/10.1093/nargab/lqaa032).*

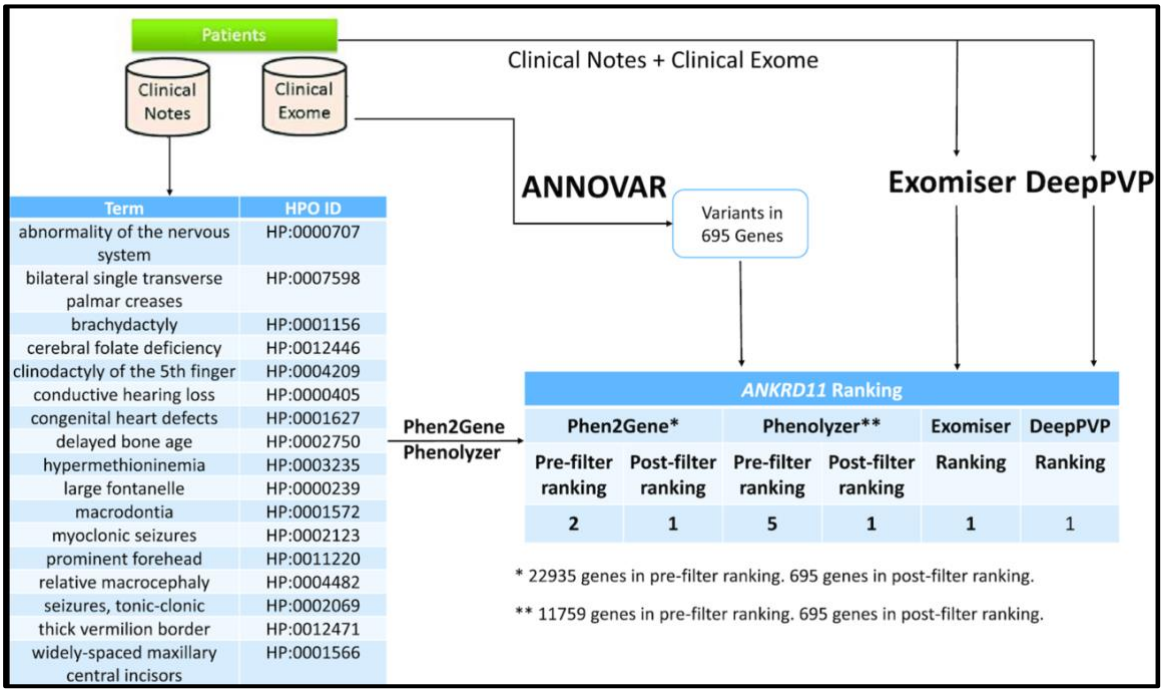


Figura 20 *– Confronto tra Phen2Gene, Phenolyzer, Exomiser, DeepPVP nell'individuazione del gene coinvolto in un probando con la Sindrome KBG.*
(Fonte: https://doi.org/10.1093/nargab/lqaa032).

Questi tool sono in grado di priorizzare migliaia di geni, in particolare si può notare come Phen2Gene esibisca un numero maggiore di geni analizzati nel processo. Exomiser e DeepPVP[46] prioritizzano il gene responsabile dell'anomalia fenotipica in un solo ciclo impiegando però moltissimo tempo e richiedendo grandi spazi di memoria. Dopo un primo filtraggio Phen2Gene cataloga come secondo il gene coinvolto, Phenolyzer quinto. Sfruttando uno specifico software per la rilevazione delle varianti genetiche (ANNOVAR) dall'esoma, è stato possibile eliminare i geni non catalogati tramite ANNOVAR e quindi ripetere la prioritizzazione dei soli geni in cui è stata riscontrata una effettiva variazione. Da questo secondo ciclo sia Phenolyzer sia Phen2Gen mostrano risultati precisi, con tempi decisamente più ristretti di Exomiser e DeepPVP[64]. Questi risultati confermano in parte alcune riflessioni già affrontate nel capitolo precedente. Come nel caso appena visto,

[46] DeepPVP impiega in media un giorno solo per scaricare il database per il suo funzionamento.

possibili strategie future potrebbero essere l'utilizzo combinato di tool integrati con termini fenotipici associati ad una analisi incrociata con software capaci di sequenziare le variazioni esomiche o genomiche. Così facendo si potrebbero fornire ai dottori strumenti adatti ai tempi clinici ottimizzandone in una seconda fase i risultati di prioritizzazione. Un altro approccio potrebbe essere quello di continuare a lavorare sulla valorizzazione dei termini HPO in strategie strutturate per la loro selezione [77].

6. Conclusioni

Le ontologie per loro natura sono state caratterizzate affinchè sappiano comprendere, reperire e gestire informazioni. La rappresentazione della conoscenza che forniscono garantisce l'interoperabilità e genera una *knowledge base* interpretabile in modo automatico dalle macchine. Nella realizzazione di applicazioni biomediche, trovano utilizzo come conoscenza di dominio esterna integrabile in sistemi già allo stato dell'arte. Il tentativo è quello di colmare l'assenza di una conoscenza strutturata, *machine readable* e con valenza semantica e al tempo stesso sintattica. Questo è un limite alle performance di questi sistemi. Nel caso di sistemi di deep learning come BiOnt, l'utilizzo di quattro ontologie interconnesse tra loro e con la rete neurale permette di aumentare il grado di inferenza, riuscendo a dedurre possibili relazioni ignote dalle ontologie. Questo genere di applicazione risulta trarre un chiaro vantaggio dalle ontologie rispetto ai sistemi allo stato dell'arte.

Nel caso dei tool di prioritizzazione genetica, le conclusioni non sono altrettanto nette. In generale, questi tool forniscono strumenti di interesse clinico notevole. Il tool Phen2Gene risulta fornire risultati migliori rispetto al predecessore Phenolyzer, ma questo potrebbe non essere necessariamente dovuto ad un utilizzo migliore della *Human Phenotype Ontology.* Alcuni studi presentano delle perplessità sull'autonomia di questi strumenti che usano annotazioni fenotipiche da ontologie per la prioritizzazione genetica: si ritiene che per avere risultati sicuri sia necessario un confronto incrociato dei risultati con sistemi di sequenziamento genomico. Nell'analisi comparativa tra più sistemi di prioritizzazione è stato necessario l'uso di ANNOVAR per selezionare tutte le varianti esomiche, questo confermerebbe la supposizione precedentemente fatta. Tuttavia, Phen2Gene non è stato incluso nei suddetti studi e in letteratura non sono ancora presenti trattazioni che provino le performance di Phen2Gene su una popolazione di campioni numerosa.

Nonostante queste limitazioni, i sistemi di prioritizzazione genetica e i modelli di *deep learning* hanno ottenuto risultati promettenti. La ricerca futura potrà concentrarsi sia sul migliorare e validare ulteriormente questi approcci sia

nell’identificare nuovi ambiti applicativi biomedici in cui sfruttare il potenziale delle ontologie.

Bibliografia

[1] L. Ding, P. Kolari, Z. Ding, e S. Avancha, «Using Ontologies in the Semantic Web: A Survey», Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems vol. 14, MA: Springer US, 2007. Disponibile su: http://link.springer.com/10.1007/978-0-387-37022-4_4

[2] W3C, «XML Essentials».
Disponiile su: https://www.w3.org/standards/xml/core

[3] W3C, «Universal Resource identifiers in WWW».
Disponibile su: https://www.w3.org/Addressing/URL/uri-spec.html

[4] W3C, «Unicode in XML and other Markup Languages».
Disponibile su: https://www.w3.org/TR/unicode-xml/

[5] Wikipedia, «Resource Description Framework».
Disponibile su:
https://it.wikipedia.org/w/index.php?title=Resource_Description_Framework&oldid=116821640

[6] D. Fensel, J. A. Hendler, e H. Lieberman, «Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential», 2005, ISBN: 978-0-262-56212-6.

[7] Wikipedia, «HTML».
Disponibile su:
https://it.wikipedia.org/w/index.php?title=HTML&oldid=121580548

[8] Wikipedia, «RDF Schema».
Disponibile su:
https://it.wikipedia.org/w/index.php?title=RDF_Schema&oldid=118733638

[9] W3C, «OWL - Semantic Web Standards».
Disponibile su: https://www.w3.org/OWL/

[10] W3C, «SPARQL Query Language for RDF».
Disonibile su: https://www.w3.org/TR/rdf-sparql-query/

[11] Wikipedia, «Structured Query Language».
Disponibile su:
https://it.wikipedia.org/w/index.php?title=Structured_Query_Language&oldid=120555041

[12] C. Bizer, T. Heath, e T. Berners-Lee, «Linked Data - The Story So Far».
Disponibile su:
https://www.researchgate.net/publication/225070216_Linked_Data_The_Story_so_Far

[13] Vocabolario Treccani, «ònto-».
Disponibile su: https://www.treccani.it/vocabolario/onto

[14] VocabolarioTreccani, «lògos».
Disponibile su: https://www.treccani.it/vocabolario/logos

[15] Thomas R.Gruber, «A translation approach to portable ontology specifications». Disponibile su:
https://www.sciencedirect.com/science/article/abs/pii/S1042814383710083

[16] G. Canfora, D. D. Fatta, e G. Pilato, «Ontologie e Linguaggi Ontologici per il Web Semantico», TechReport, 2004.
Disponibile su:
https://intranet.icar.cnr.it/wp-content/uploads/2016/11/TechReport-04-06.pdf

[17] M. R. Genesereth e N. J. Nilsson, «Logical Foundations of Artificial Intelligence», 1987.

[18] N. Guarino, «Formal Ontology in Information Systems: Proceedings of the First International Conference »,1998.

[19] Stanford University, «Protégé 5 Documentation: View». Disponibile su:
http://protegeproject.github.io/protege/views/

[20] D. Mcguinness, «Ontologies Come of Age.», 2003.

[21] S. Pieroni, M. Franchini, F. Mariani, L. Fortunato, S. Molinaro, «Ontologie e modellazione di dati sanitari Attività di ricerca nell'ambito del progetto ODINET». Disponibile su: https://core.ac.uk/download/pdf/37830677.pdf

[22] M. Fernandez, A. Gomez-Pearez, e N. Juristo, «Methontology: From Ontological Art Towards Ontological Engineering». Disponibile su:
http://oa.upm.es/5484/1/METHONTOLOGY_.pdf

[23] N. F. Noy e D. L. McGuinness, «Ontology Development 101: A Guide to Creating Your First Ontology».
Disponibile su:
https://protege.stanford.edu/publications/ontology_development/ontology101.pdf

[24] M. Grüninger e M. S. Fox, «The Role of Competency Questions in Enterprise Engineering», MA: Springer US, 1995. Disponibile su:
https://doi.org/10.1007/978-0-387-34847-6_3

[25] M. Uschold e M. Grüninger, «Ontologies: Principles, methods and applications»,1996. Disponibile su:
https://www.researchgate.net/publication/302937543_Ontologies_Principles_methods_and_applications

[26] S. Boyce e C. Pahl, «Developing Domain Ontologies for Course Content». Disponibile su: https://core.ac.uk/download/pdf/11310019.pdf

[27] G. van Heijst, A. Th. Schreiber, e B. J. Wielinga, «Using explicit ontologies in KBS development»,1997. Disponibile su: https://linkinghub.elsevier.com/retrieve/pii/S1071581996900907

[28] N. Guarino, «The Ontological Level»,1998. Disponibile su: https://www.researchgate.net/publication/2603589_The_Ontology_Level

[29] D. Mcguinness, R. Fikes, J. Rice, e S. Wilder, «An Environment for Merging and Testing Large Ontologies.», 2000. Disponibile su: https://www.researchgate.net/publication/221393548_An_Environment_for_Merging_and_Testing_Large_Ontologies

[30] U.S. National Library of Medicine, «Unified Medical Language System (UMLS)». consultato lug. 04, 2021). Disponibile su: https://www.nlm.nih.gov/research/umls/index.html

[31] H. A. Heathfield, N. R. Hardiker, e J. Kirby, «Using the PEN&PAD information model to support hospital-based clinical care», 1994. Disponibile su: https://pubmed.ncbi.nlm.nih.gov/7949968/

[32] A. L. Rector, W. A. Nowlan, S. Kay, C. A. Goble, e T. J. Howkins, «A framework for modelling the electronic medical record»,1993. Disponibile su: https://pubmed.ncbi.nlm.nih.gov/8321129/

[33] W. Nowlan e A. Rector, «Medical Knowledge Representation and Predictive Data Entry», 1991. Disponibile su: https://www.researchgate.net/publication/302229914_Medical_Knowledge_Representation_and_Predictive_Data_Entry

[34] Cycorp, «Cyc | The Next Generation of Enterprise AI». Disponibile su : https://cyc.com/

[35] N. Guarino, «Understanding, building and using ontologies»,1997. Disponibile su: https://www.sciencedirect.com/science/article/pii/S1071581996900919

[36] A. Maier, J. Aguado , A. Bernaras, I. Laresgoiti, C. Pedinaci, P. Nieves, T. Smithers, «Integration with Ontologies.», 2003. Disponibile su: researchgate.net/publication/220927290_Integration_with_Ontologies

[37] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, M.A. Musen, «BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications». Disponibile su: https://doi.org/10.1093/nar/gkr469

[38] C. Pesquita, D. Faria, A. O. Falcão, P. Lord, e F. M. Couto, «Semantic Similarity in Biomedical Ontologies», 2009. Disponibile su:

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000443

[39] National Center for Biomedical Ontology, «BioPortal». Disponibile su: https://bioportal.bioontology.org/ontologies

[40] CGIAR: Science for humanity's greatest challenges. Disponibile su: https://www.cgiar.org/

[41] The OBO Foundry. Disponibile su: http://www.obofoundry.org/

[42] OBI, Ontology for Biomedical Investigation. Disponibile su: http://obi-ontology.org/

[43] GO Consortium, «Gene Ontology». Disponibile su: http://geneontology.org/docs/go-consortium/

[44] HUPO - Proteomics Standards Initiative. Disponibile su: https://www.hupo.org/Proteomics-Standards-Initiative

[45] SNOMED CT, NCBO BioPortal. Disponibile su: https://bioportal.bioontology.org/ontologies/SNOMEDCT

[46] Gene Ontology (GO), NCBO BioPortal. Disponibile su: https://bioportal.bioontology.org/ontologies/GO

[47] Protein Ontology, NCBO BioPortal. Disponibile su: https://bioportal.bioontology.org/ontologies/PR

[48] Chemical Entities of Biological Interest Ontology, NCBO BioPortal. Disponibile su: https://bioportal.bioontology.org/ontologies/CHEBI

[49] Human Phenotype Ontology, NCBO BioPortal. Disponibile su: https://bioportal.bioontology.org/ontologies/HP

[50] Human Disease Ontology, NCBO BioPortal. Disponibile su: https://bioportal.bioontology.org/ontologies/DOID

[51] The Drug Ontology, NCBO BioPortal. Disponibile su: https://bioportal.bioontology.org/ontologies/DRON

[52] E. A. Feigenbaum, «The Art of Artificial Intelligence. 1. Themes and Case Studies of Knowledge Engineering», 1977. Disponibile su: https://apps.dtic.mil/sti/citations/ADA046289

[53] D. Sousa, «Deep Learning System for Biomedical Relation Extraction Combining External Sources of Knowledge», 2021. Disponibile su: https://www.researchgate.net/publication/350569542_Deep_Learning_System_for_Biomedical_Relation_Extraction_Combining_External_Sources_of_Knowledge

[54] H. Pan, Y. Zhu, S. Yang, Z. Wang, W. Zhou, Y. He e X. Yang, «Biomedical ontologies and their development, management, and applications in and beyond China», 2019. Disponibile su: https://journals.lww.com/jbioxresearch/fulltext/2019/12000/biomedical_ontologies_and_their_development,.5.aspx

[55] D. Rubin, N. Shah, e N. Noy, «Biomedical ontologies: A functional perspective», 2008. Disponibili su: https://www.researchgate.net/publication/5772043_Biomedical_ontologies_A_functional_perspective

[56] Human Phenotype Ontology. Disponibile su: https://hpo.jax.org/app/

[57] GO annotations. Disponibile su: http://geneontology.org/docs/go-annotations/

[58] Phenomiser. Disponibile su: https://hpo.jax.org/app/tools/phenomizer

[59] PhenoGramViz. Disponibile su: https://hpo.jax.org/app/tools/phenogramviz

[60] Exomiser. Disponibile su: https://hpo.jax.org/app/tools/exomiser

[61] Genomiser. Disponibile su: https://hpo.jax.org/app/tools/genomiser

[62] M. Kulmanov, F. Z. Smaili, X. Gao, e R. Hoehndorf, «Machine learning with biomedical ontologies», 2020. Disponibile su: http://biorxiv.org/lookup/doi/10.1101/2020.05.07.082164

[63] H. Yang, P. N. Robinson, e K. Wang, «Phenolyzer: phenotype-based prioritization of candidate genes for human diseases», 2015. Disponibile su: https://www.nature.com/articles/nmeth.3484

[64] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah khan, «A Review of Machine Learning Algorithms for Text-Documents Classification»,2021. Disponibile su: http://www.jait.us/index.php?m=content&c=index&a=show&catid=160&id=856

[65] A. Lamurias, D. Sousa, L. A. Clarke, e F. M. Couto, «BO-LSTM: classifying relations via long short-term memory networks along biomedical ontologies», 2019. Disponibile su: https://doi.org/10.1186/s12859-018-2584-5

[66] T. M. Mitchell, «Machine Learning», McGraw-Hill, 1997. Disponibile su: https://akum.pw/mkk00nyuy5.pdf

[67] N. Bach e S. Badaskar, «A Review of Relation Extraction». Disponibile su: https://www.cs.cmu.edu/~nbach/papers/A-survey-on-Relation-Extraction.pdf

[68] Y. LeCun, Y. Bengio, e G. Hinton, «Deep learning», 2015. Disponibile su: https://www.nature.com/articles/nature14539

[69] Ontologies In Agriculture, Webinar : Machine Learning and ontology. Disponibile su: https://www.youtube.com/watch?v=eVs0KjV9LhU

[70] Wikipedia, «Rete neurale ricorrente». Disponibile su: https://it.wikipedia.org/w/index.php?title=Rete_neurale_ricorrente&oldid=120617040

[71] Alfredo Canziani, «Architettura delle RNNs e modelli LSTM · Apprendimento Profondo». Disponibile su: https://atcold.github.io/pytorch-Deep-Learning/it/week06/06-3/

[72] D. Sousa e F. M. Couto, «BiOnt: Deep Learning Using Multiple Biomedical Ontologies for Relation Extraction», 2020. Disponibile su: https://link.springer.com/chapter/10.1007%2F978-3-030-45442-5_46#Sec2

[73] Breda Gentics srl, «Analisi del genoma (whole genome sequencing)», 2017. Disponibile su: https://bredagenetics.com/analisi-del-genoma-whole-genome-sequencing/?lang=it

[74] J. H. Son, G. Xie, C. Yuan, L. Ena, Z. Li et al., «Deep Phenotyping on Electronic Health Records Facilitates Genetic Diagnosis by Clinical Exomes», 2018. Disponibile su : https://www.sciencedirect.com/science/article/pii/S000292971830171X

[75] Breda Genetics srl, «Analisi dell'esoma (exome sequencing)», 2016. Disponibile su: https://bredagenetics.com/analisi-dellesoma-exome-sequencing/?lang=it

[76] Wikipedia , «Fenotipo». Disponibile su: https://it.wikipedia.org/w/index.php?title=Fenotipo&oldid=121037454

[77] A. Fellner, N. Ruhrman-Shahar, N. Orenstein, G. Lidzbarsky, A. R. Shuldiner, «The role of phenotype-based search approaches using public online databases in diagnostics of Mendelian disorders», 2021. Disponibile su: https://www.nature.com/articles/s41436-020-01085-7

[78] M. Zhao, J. M. Havrilla, L. Fang, Y. Chen, J. Peng, C. Liu, C. Wu, M. Sarmady, P. Botas, J. Isla, G. J. Lyon, C. Weng, K. Wang, «Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases», 2020. Disponibile su: https://doi.org/10.1093/nargab/lqaa032

Ringraziamenti

Per chi mi conosce da tempo sa benissimo che non sono affatto persona da ringraziamenti. Neanche da dedica, ma quella mi sono presa lo sfizio di non farla per davvero.

Però, sempre gli stessi di prima, sanno che non sono così carogna da non scrivere due righe di ringraziamenti, nonostante scrivere sia l'attività più terrificante seconda solo al viaggiare su un regionale Siena-Firenze senza aria condizionata, d'estate e con 50 gradi all'ombra.

Ci tengo particolaramente a ringraziare il Professor Roffia, per l' entusiasmo contagioso che mette nel suo lavoro che ha raggiunto e coinvolto anche me.

Grazie mamma e babbo per la pazienza di questi lunghi anni.

Grazie Chia di essere sempre l'incarnazione (o reincarnazione) del lavoro e della dedizione, no matter what. Nonostante sia impossibile starti dietro, sei una risorsa.

Grazie ai nonni per il conforto e l'amore incondizionato.

Grazie Dani, per la presenza costante e senza riserve.

Grazie a tutti i coinquilini avuti negli anni, nella maggior parte dei casi supporter morali numeri uno di crisi nevrotiche e serate in allegria. Molto più amici che flatmate. Anche se sono conosciuta per essere molto come Dory prometto che avrete sempre un angolo di memoria e di affetto tutto per voi.

Grazie a tutti gli amici, gli amici di amici, gli amici di amici di amici che erano amici prima di essere amici. Grazie. E basta, il resto lo sapete da soli. Che sia chiaro: grazie anche a tutti quelli che sono stati amici con il verbo al passato, non fa niente, è andata così ma grazie uguale del pezzetto di cammino insieme.

Infine, concedetemi un piccolo momento di autoreferenzialità.

Grazie anche a me.

Dai ho finito, potete smettere di fare i frignoni.

Elisa.