



**UNIVERSITÀ DI PISA**

Corso di Laurea in Informatica Umanistica

RELAZIONE

**Generazione automatica di ontologie**

**Candidato:** *Luana Grillo*

**Relatore:** *Prof. Franco Turini*

Anno Accademico 2008-2009

## RINGRAZIAMENTI

*Desidero ringraziare con immenso affetto mia mamma, Maurizio, Leandro e Giorgio per l'amore e la pazienza con cui mi hanno accompagnato in questo percorso.*

*Ringrazio il Prof. Franco Turini per la professionalità e la disponibilità con cui mi ha seguito.*

*Grazie al Prof. Mirko Tavoni.*

*E ancora un ringraziamento speciale a mia zia Mariangela e a mia zia Lidia per avermi guidato con il loro amore nei momenti difficili.*



# Indice generale

1. Introduzione.....	5
1.1 Il Web Semantico.....	6
1.2 Struttura della tesi.....	10
2. Concetti fondamentali di ontologia.....	11
2.1 A cosa serve un'ontologia.....	12
2.2 Ontologie e Basi di Conoscenza.....	12
2.3 Ontologie e Basi di Dati (Database).....	13
2.4 Dai dati alla conoscenza.....	16
3. Il problema della generazione di ontologie.....	17
3.1 Ontological Engineering: processo di creazione e gestione di ontologie.....	17
3.2 Tesi sostenuta da Aleksander Pivk.....	22
4. Un approccio che sfrutta Open Calais per la ricerca di concetti.....	26
4.1 Introduzione a Open Calais.....	26
4.2 Considerazioni sull'applicazione di Calais: un esempio pratico.....	28
5. Conclusioni e prospettive.....	32
6. Bibliografia.....	33

# 1. Introduzione

Internet ai suoi esordi era costituito unicamente da testi e da ipertesti. Col tempo le cose si sono evolute portando il mondo dell'informazione a un livello di integrazione e condivisione senza precedenti tanto che la mole di informazione condivisa è cresciuta senza controllo. Chiunque abbia accesso a una rete e in particolare a internet, oggi può avere, infatti, a disposizione un'enorme quantità di dati e gli ipertesti che lo compongono possono richiamarsi l'uno con l'altro in modo molto rapido. Il web mette a disposizione documenti e risorse multimediali riguardanti gli argomenti più disparati (Wikipedia, voce *Internet*).

Un soggetto può recuperare dalla rete ciò di cui necessita per il lavoro, lo studio, la comunicazione e gli interessi personali e può anche condividerlo nuovamente, alimentando continuamente il processo di aggiornamento delle informazioni.

Questa situazione assicura una lunga serie di vantaggi, ma apporta anche un certo numero di nuovi problemi da risolvere. Tra questi assume un notevole rilievo quello della selezione delle informazioni che realmente interessano l'utente.

Ma perché l'utente non riesce a trovare con facilità le informazioni desiderate?

La ragione è molto semplice: l'utente si orienta nel web grazie alla sua esperienza di navigazione e alla capacità di evocazione che possono avere parole o espressioni chiave. L'esperienza è un aspetto molto importante che, però, si acquisisce nel tempo in maniera spontanea e non è molto legata ad aspetti tecnici, al codice e alle applicazioni che costituiscono un sito. Per questo motivo, molto spesso, quando viene utilizzato un motore di ricerca, qualsiasi query è sempre soggetta al rischio dell'ambiguità. Il web in particolare può essere visto come un enorme contenitore in cui i dati sono memorizzati e organizzati secondo criteri quasi esclusivamente lessicali. Ciò è dovuto al fatto che le informazioni sono create a uso e consumo dei soli utenti umani, gli unici allo stato attuale in grado di comprendere i contenuti delle pagine che stanno visitando.

L'aspetto più importante è invece la loro organizzazione semantica, poiché è questa che permette a un utente di filtrare e selezionare le informazioni di cui ha bisogno.

Di cosa si necessita, dunque, per sfruttare al meglio le risorse di cui oggi disponiamo?

Per risolvere il problema è necessario un Web in cui vi siano applicazioni capaci di comprendere il significato dei testi presenti sulla rete e che siano anche in grado di guidare l'utente direttamente verso l'informazione desiderata, oppure di sostituirsi a lui nello svolgimento di alcune operazioni.

Il titolo del presente lavoro di tesi, ovvero “generazione automatica di ontologie”, introduce l'argomento trattato. Si cercherà di spiegare l'importanza della generazione automatica di ontologie, oltre a presentare la proposta elaborata da Aleksander Pivk per trovare una soluzione a tali esigenze.

A cosa può servire la generazione automatica di ontologie? Perché è utile la condivisione delle informazioni? A che punto sono gli studi sull'argomento presentato e quali sono gli sviluppi futuri per i quali gli studiosi stanno lavorando? Queste sono alcune delle domande analizzate per sviluppare l'argomento di questa tesi.

## **1.1 Il Web Semantico.**

Il termine Web Semantico è stato proposto per la prima volta da Tim Berners-Lee nel 2001 per identificare un web più funzionale, in grado di fornire regole sintattiche e semantiche per l'elaborazione dei documenti in modo da consentire non solo alle persone, ma anche alle macchine di comprendere i contenuti sul web migliorando i servizi disponibili (Wikipedia, voce *Internet*). Infatti quando vogliamo soddisfare un nostro bisogno informativo, cioè la necessità che ci spinge ad accrescere la nostra conoscenza su un determinato argomento, di solito ci affidiamo ai motori di ricerca, ma non sempre i risultati sono quelli sperati.

Per chiarire questo concetto Andrea Morandini propone un esempio che analizza la fase di ricerca di un qualsiasi utente che utilizza i motori di ricerca (TSW, voce Andrea Morandini). In particolar modo Morandini sottolinea che di solito i motori di ricerca sono poco efficaci in quanto, molto spesso, non otteniamo subito una risposta adeguata alle nostre esigenze, ma siamo costretti a compiere una serie di ricerche prima di ottenerla. Dobbiamo quindi attivare un insieme di processi cognitivi: “Ciò che ci permette di attivare in modo naturale questi processi è quella che Noam Chomsky chiama competenza linguistica: ogni individuo ha delle competenze innate nel capire i suoni della lingua parlata (competenza fonologica), di riconoscere frasi

grammaticalmente corrette (competenza sintattica) e di estrarre significati da queste frasi (competenza semantica) (Morandini 2006)” (TSW, voce Andrea Morandini).

Alcune volte, però, come nel caso in cui vengano utilizzate nuove terminologie di cui non eravamo a conoscenza, il nostro background non ci aiuta nel processo di ricerca. Così è utile affidarsi a sistemi di ricerca non più basati su “parole chiave” ma su “concetti”.

Come dice Morandini: ”Tali sistemi di ricerca si basano sulla supposizione che nel nostro cervello i significati delle parole sono organizzati in una mappa concettuale, dove i termini sono collegati fra loro in base alla prossimità semantica che hanno due termini, cioè l’analogia che le nostre conoscenze le attribuiscono. Queste “mappe” sono chiamate reti semantiche e rappresentano uno dei possibili modelli usati per simulare le nostre competenze linguistiche innate”.

Ciò che è stato appena esaminato, grazie anche all’esempio di Morandini, ci ha permesso di capire che è necessario utilizzare dei sistemi che possano essere inizialmente guidati nella costruzione di un’ontologia e che una ricerca tramite keyword possa essere inefficace in alcuni frangenti, mentre i sistemi che utilizzano reti semantiche sono capaci di colmare queste lacune. E’ fondamentale, quindi, disporre di uno schema concettuale esaustivo e rigoroso che permetta di capire quali sono i criteri di ricerca utilizzati per distinguere se il documento trovato è appropriato o meno e, soprattutto, che nel tempo si possa arricchire dei nuovi concetti acquisiti.

Un sistema così istruito sarà in grado di fornire risposte adeguate all’utente scegliendo le fonti più autorevoli e pertinenti fra i risultati delle ricerche effettuate.

Attualmente si sta sviluppando l’idea di utilizzare schemi per descrivere domini d’informazione, in cui un dominio viene descritto da un particolare schema e i dati di questo schema di dominio vengono mappati da dei metadati rispetto a classi, o concetti. Quindi, tutto dipende dallo schema, dall’insieme di regole che si è deciso di utilizzare per archiviare le informazioni. Il modo di organizzare e classificare queste informazioni è definito appunto ontologia, il cui compito è quello di legare i concetti fra loro evidenziandone le caratteristiche e le relazioni.

L’obiettivo è quello di basarsi su ontologie condivise per rendere la rete un’enciclopedia utilissima che abbracci tutti gli ambiti del sapere.

La realizzazione del Semantic Web necessita di algoritmi e strumenti per la costruzione automatica di ontologie. Questo perché la costruzione manuale di ontologie non può gestire, sia per questioni di costo che di tempo, l’enorme quantità

di conoscenza disponibile attualmente sul Web. Per rendere operativo il Web semantico e per attuare le potenzialità del web sono molte le tecnologie che devono lavorare insieme: il Web semantico si basa sullo standard Uri (Uniform Resource Identifiers) per la definizione univoca di indirizzi Internet, sullo standard Xml (Extensible Markup Language) e sullo standard Rdf (Resource Description Framework) nonché sull'uso delle ontologie.

XML è un meta-linguaggio che fornisce un insieme standard di regole sintattiche per modellare la struttura di documenti e dati senza troppi vincoli. Questa libertà lo rende, però, poco adatto a definire completamente la struttura e l'interscambio di informazioni tra diverse realtà, perciò è stato creato un nuovo linguaggio derivato da Xml: Rdf (Resource Description Framework).

RDF è un'applicazione di XML, una base per il trattamento dei metadati, che permette interoperabilità tra le applicazioni che scambiano informazioni comprensibili alle macchine sul Web. Con RDF si ha l'intento di realizzare un meccanismo per la descrizione di risorse che sia adattabile a qualsiasi dominio e che, quindi, non si basi su un particolare dominio di applicazione definendone a priori la sua semantica (w3c, voce RDF).

Dato che la quantità di informazioni disponibili nelle intranet delle società aumenta costantemente, la necessità di funzionalità di ricerca avanzate cresce allo stesso passo. Tuttavia, sembra che le funzionalità di ricerca avanzata presuppongano che le informazioni siano strutturate, di conseguenza, in riferimento a determinate categorie. Gli strumenti di ricerca possono, quindi, trarre profitto da risorse di informazioni indicizzate, in riferimento a queste categorie, per fornire le funzionalità di ricerca avanzate: ricerca per categoria, recupero di documenti con categorie simili, ecc.

L'attività di apprendimento dell'ontologia può essere applicata, ad esempio, per ricavare automaticamente le categorie dei dati sottostanti, che possono poi essere utilizzati per indicizzare le risorse informative e, conseguentemente, per favorire la loro ricerca e il riutilizzo. L'indicizzazione delle risorse rispetto ad una data tassonomia delle categorie può essere fatta manualmente o mediante l'applicazione di tecniche di estrazione di testo. Recentemente è stato anche dimostrato che le tecniche di estrazione dal testo, come classificazione o raggruppamento di testo, possono essere migliorata integrando informazioni nell'ontologia.

Uno dei problemi principali con cui le persone hanno lottato per il reperimento delle informazioni è il cosiddetto problema di incongruenza del vocabolario, che consiste essenzialmente nel fatto che, in molti casi, i termini utilizzati in una query non corrispondono alle parole presenti in un documento, sebbene entrambi si adattano l'uno all'altro da un punto di vista semantico. Nell'estrazione di testo, siamo di fronte a una situazione analoga.

I risultati recenti nel campo della classificazione e del raggruppamento di documenti hanno dimostrato che le gerarchie di concetto che apprendono automaticamente possono essere applicate con successo per superare parzialmente il problema di incongruenza di vocabolario.

Le principali tendenze che possono essere identificate nel campo dell'apprendimento dell'ontologia sono di occuparsi dell'incertezza delle previsioni degli algoritmi di apprendimento e di creare delle strutture flessibili per l'apprendimento dell'ontologia. D'altronde, la maggioranza dei ricercatori ha capito che l'output degli algoritmi di apprendimento dell'ontologia è ben lontano dall'essere perfetto. Di conseguenza, per rendere il processo controllabile, vi è la necessità di valutare quanto è certo un algoritmo nelle sue previsioni. I valori numerici che riguardano la certezza di una previsione di un algoritmo potrebbero, quindi, essere usati come base per combinare diversi algoritmi in modo da compensare gli svantaggi e le false previsioni.

Anche se diverse proposte sono emerse recentemente, il problema è lontano dall'essere risolto. In particolare, arrivare a dare una base solida per la combinazione di algoritmi, con un'interpretazione ben definita di certezza, sembra un problema non banale. A ogni fase del processo di apprendimento dell'ontologia dovrebbe avvenire una valutazione automatica o manuale dell'ontologia "istruita" per evitare la propagazione di errori e per consentire una riconfigurazione degli algoritmi dall'utente o dal sistema stesso. Tale valutazione, per essere efficiente, deve tener conto delle prove o delle certezze generate dai singoli algoritmi di apprendimento dell'ontologia, delle esigenze specificate dall'utente, come pure dei vincoli formali della lingua di modellazione dell'ontologia di destinazione.

## 1.2 Struttura della tesi

Le attività preliminari svolte per questo lavoro di tesi sono state lo studio delle informazioni e degli articoli riguardanti le ontologie e soprattutto la generazione automatica di ontologie.

L'obiettivo principale di questa tesi è quello di affrontare il problema della generazione automatica di ontologie seguendo gli sviluppi degli studi condotti sull'argomento. Come conclusione di tale studio un esempio pratico mostrerà l'elaborazione di un articolo di giornale analizzato semanticamente tramite Open Calais. Inoltre, sempre utilizzando Calais, verrà applicata una parte del metodo descritto da Aleksander Pivk.

Il presente lavoro è organizzato in cinque sezioni, il cui contenuto è riassunto di seguito al fine di offrire una panoramica della trattazione:

- La prima e la seconda sezione introducono rispettivamente i concetti fondamentali del Web Semantico e delle ontologie.
- Nella terza sezione viene illustrato il problema della generazione di ontologie e l'approcci proposto da A. Pivk per affrontare tale questione.
- Nella quarta sezione, all'introduzione di Open Calais segue un esempio pratico d'uso di tale programma.
- L'ultima sezione è dedicata alle valutazioni conclusive sul lavoro svolto e ad alcune considerazioni per eventuali sviluppi futuri.

## 2. Concetti fondamentali di ontologia

Le origini dell'ontologia vanno ricercate in ambito filosofico e solo negli ultimi anni si è avuto un largo impiego nell'ambito dell'IA (Intelligenza Artificiale) e nelle discipline informatiche.

In ambito informatico un'ontologia è definita da Gruber (1993) come una *specificazione di una concettualizzazione*, ovvero una descrizione formale di un insieme di concetti e delle relazioni che intercorrono tra essi.

Un'ontologia è uno schema, strutturato gerarchicamente, di termini e relazioni che descrivono un certo dominio. I linguaggi per definire le ontologie sono sintatticamente e semanticamente ricchi di informazioni e utilizzano una terminologia comune e concordata per poter essere usata e riusata nella condivisione e nello scambio di informazioni.

Le ontologie oggi sono utilizzate per applicazioni che richiedono molti dati, servizi e risorse in generale, per essere interoperabili e comprensibili alle macchine e alle persone. Queste applicazioni possono essere, per esempio, l'individuazione e l'uso di servizi web, l'integrazione di informazioni attraverso database, ecc.

L'idea generale è che per consentire la condivisione e il riutilizzo dell'informazione è necessario fornire una descrizione semantica dei dati e dei servizi. Le nuove ricerche, si sono concentrate proprio verso lo sviluppo di sistemi capaci di favorire la condivisione e il riuso delle informazioni, in modo da creare un sapere comune e condiviso di un certo dominio.

I dati, i servizi e le risorse sono, quindi, annotati semanticamente attraverso metadati che descrivono il loro significato o le loro capacità.

L'annotazione dei dati, riguardo a un'ontologia, permette oltre alla condivisione anche un loro maggior reperimento, in modo che la ricerca del servizio desiderato possa essere resa più efficace e che vari servizi possano essere generati sulla base delle loro descrizioni semantiche per raggiungere un obiettivo più complesso.

Le ontologie, infatti, permettono una migliore strutturazione del sapere per la ricerca e per i metodi di ragionamento automatico che possono aiutare a raggiungere risultati in modo più veloce.

## 2.1 A cosa serve un'ontologia

L'introduzione delle ontologie nel Web consente la strutturazione delle informazioni e permette di superare alcuni aspetti critici del Web tradizionale.

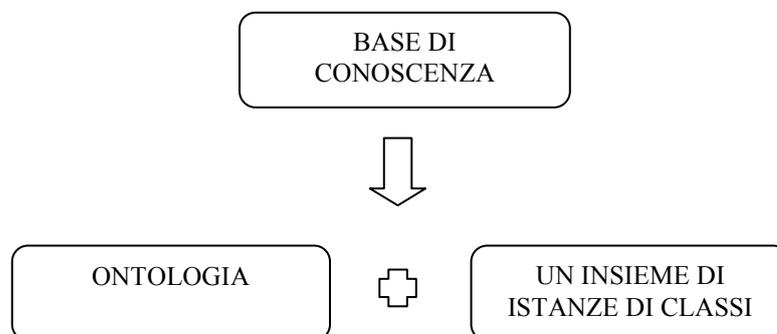
In primo luogo, un'ontologia descrive in modo formale i contenuti di ogni sorgente che genera informazione e per questo motivo è in grado di aiutare un utente nella ricerca dell'informazione desiderata. Di conseguenza, viene risolto anche il problema che riguarda la mancanza di struttura del Web, dovuta al fatto che HTML è un linguaggio di formattazione e non si occupa di gestire le informazioni per facilitare il loro reperimento. Un'ontologia descrive il dominio dal punto di vista strutturale, definendone così i componenti e i loro legami.

Le ontologie rappresentano oggi uno strumento chiave per risolvere i problemi di integrazione e di interoperabilità semantica di dati e conoscenze.

La navigazione e l'interrogazione di un'ontologia costituiscono, attualmente, il metodo più efficace per orientarsi all'interno della conoscenza complessa che caratterizza un dominio. Infatti, definendo mediante un'ontologia il contesto in cui è inserito un certo documento, i termini utilizzati sono propri di quell'ambito e quindi il loro significato non è più ambiguo.

## 2.2 Ontologie e Basi di Conoscenza

Una base di conoscenza è un particolare tipo di database per la gestione della conoscenza per scopi aziendali, culturali o didattici, ovvero è un'ontologia popolata di istanze di classi e completata con delle regole di inferenza, dove per classi si intende una descrizione formale di concetti in un dominio.



Come schematizzato da Daniele Nardi, professore presso l'Università di Roma "La Sapienza" (Dipartimento di Informatica e Sistemistica dell'Università "La Sapienza" di Roma, voce *Daniele Nardi*):

**Base di conoscenza**

Una **base di conoscenza** è una rappresentazione delle conoscenze sul mondo (problema).

- **conoscenza intensionale** esprime delle leggi generali sul dominio di interesse (regole)
- **conoscenza estensionale** esprime la conoscenza sul particolare caso (situazione) in esame (fatti)

La costruzione della base di conoscenza si effettua tramite l'asserzione delle frasi che rappresentano la conoscenza intensionale ed estensionale (Tell).

Il linguaggio usato per esprimere la conoscenza si chiama **linguaggio di rappresentazione della conoscenza**.

Rappresentazione della Conoscenza – Daniele Nardi, 2008Lezione 1 5|

### 2.3 Ontologie e Basi di Dati (*Database*)

Un database (DB) si può definire come una collezione di dati strutturati, gestita da un software che memorizza e organizza le informazioni, progettata per essere accessibile ad applicazioni e utenti differenti.

Alla base di un database c'è quello che viene definito il DBMS (Data Base Management System), ovvero l'insieme dei software necessari per la gestione e la manutenzione di grandi quantità di dati memorizzati. Questa si presenta come una struttura generica, indipendente da uno specifico dominio applicativo, quindi estremamente flessibile ed adattabile alle più disparate situazioni.

DB e DBMS sono legati da una forte interazione: il DBMS è il sistema attivo, mentre il DB è l'oggetto passivo sul quale opera il primo (Wikipedia, voce *Database*).

Esistono diversi modelli di database (gerarchici, relazionali e orientati agli oggetti), la cui classificazione è legata alla rappresentazione logica dei dati.

Il modello che meglio si presta alla rappresentazione di un'ontologia è quello dei database orientati agli oggetti, un'estensione del modello relazionale basato sui

paradigmi della programmazione ad oggetti. Essi rappresentano la realtà analizzata sotto forma di classi e sottoclassi. Si tratta di un database flessibile che supporta, appunto, l'uso di tipi di dati astratti, oggetti e classi e che può memorizzare una vasta gamma di dati, spesso includendo suono, video e grafica, oltre a testo e numeri.

L'obiettivo di questo modello di database è quello di attribuire maggior significato ai dati incapsulando in essi gli algoritmi necessari alla loro interpretazione.

Tuttavia tale modello di database risulta ancora oggi poco conveniente sia per la mancanza di un modello universalmente accettato per gli oggetti, che per la mancanza di un linguaggio di interrogazione standard. Oltre a ciò influisce negativamente anche l'elevato *overhead* computazionale sia in termini di sviluppo di progetti che di gestione del sistema.

Ad oggi, quindi, il modello di database più utilizzato è quello relazionale, considerato quello più semplice e più efficace in quanto è più vicino al modo consueto di pensare i dati e si adatta in modo naturale alla classificazione e alla strutturazione dei dati (Polo Didattico e di Ricerca di Crema, voce *Paolo Ceravolo*).

Nel modello relazionale i dati sono rappresentati tramite tabelle chiamate relazioni sulle quali si possono effettuare operazioni e tra le quali possono essere stabilite delle associazioni. Le tabelle identificano le entità e le relazioni di tipo "molti a molti" che intercorrono tra di esse ed ogni relazione è definita dalle colonne (campi) e da una serie di righe (record), (Politecnico di Torino, voce *corsi a distanza*).

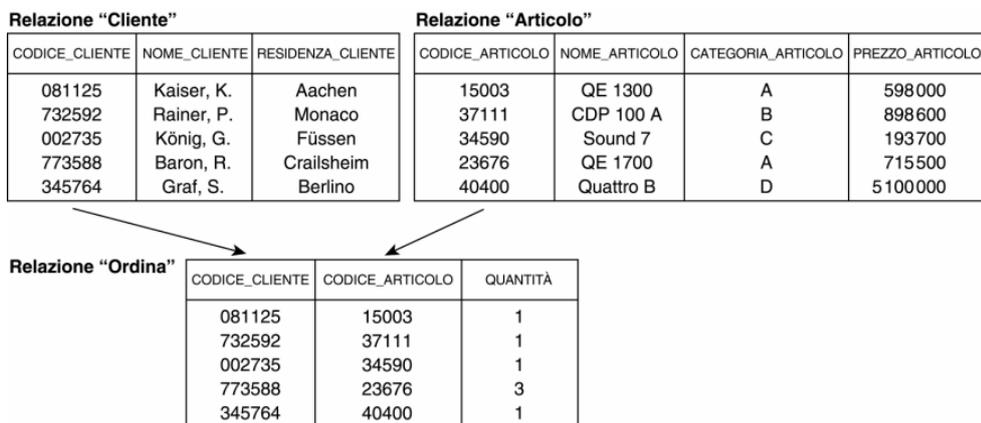
**Relazione “Articolo”**

CODICE_ARTICOLO	NOME_ARTICOLO	CATEGORIA_ARTICOLO	PREZZO_ARTICOLO
15003	QE 1300	A	598000
37111	CDP 100 A	B	898600
34590	Sound 7	C	193700
23676	QE 1700	A	715500
40400	Quattro B	D	5100000

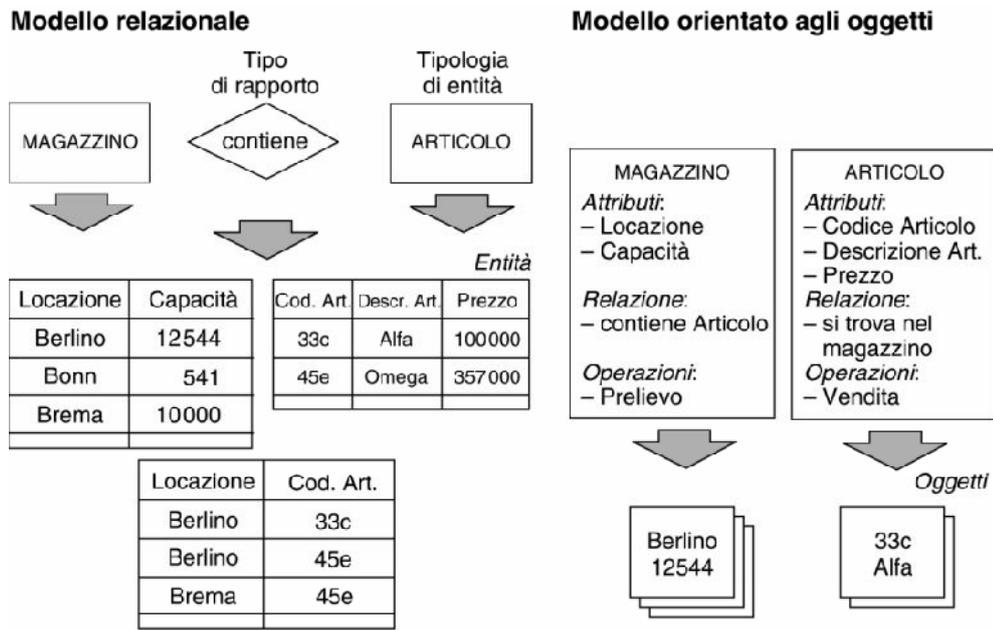
**Figura 1:** esempio di database relazionale (Politecnico di Torino, voce *corsi a distanza*).



**Figura 2:** rappresentazione di una relazione n:1 tra la relazione “Articolo” e la relazione “Categoria” (Politecnico di Torino, voce *corsi a distanza*).



**Figura 3:** rappresentazione di un rapporto n:m tra la relazione tabella “Cliente” e la relazione tabella “Articolo” con l’aiuto di una connessione tra relazioni-tabelle (Politecnico di Torino, voce *corsi a distanza*).



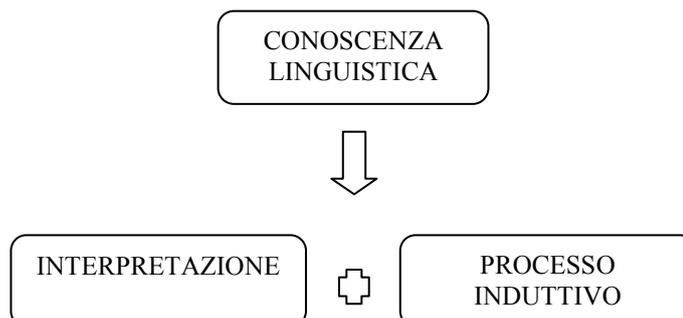
**Figura 4:** rapporti in database relazionali e orientati agli oggetti (Politecnico di Torino, voce *corsi a distanza*).

## 2.4 Dai dati alla conoscenza

Descrivere un significato etichettandolo fuori da un testo (annotazioni, regole o basi di conoscenza) è utile per consolidare un processo di interpretazione ma non si applica naturalmente al riconoscimento linguistico.

Due parole apparentemente diverse (bere  $\neq$  prendere) possono assumere lo stesso significato in un determinato contesto (es. bere del caffè / prendere del caffè).

L'interpretazione, piuttosto, ha origine dall'esperienza dei fatti linguistici che condividono lo stesso contesto. E' una forma di induzione, cioè apprendimento (Astrea, voce *dr. Paolo Naggari* - *prof. Roberto Basili*).

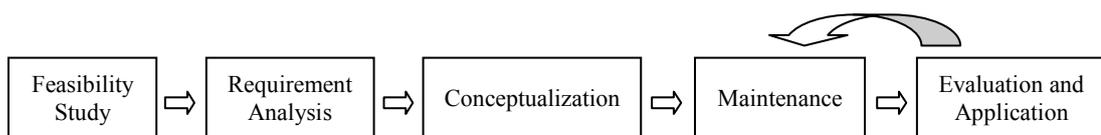


### 3. Il problema della generazione di ontologie

#### 3.1 Ontological Engineering: processo di creazione e gestione di ontologie.

Il problema della rappresentazione della conoscenza assume un ruolo centrale nello sviluppo del Web Semantico. In particolar modo, poiché la generazione di ontologie ha un costo relativamente elevato, sono emerse diverse proposte per l'apprendimento delle ontologie da risorse strutturate e non strutturate. Le ricerche attuali sono rivolte a sviluppare strumenti, algoritmi, linguaggi e modelli che siano in grado di gestire le enormi quantità di informazioni e risorse disponibili tramite l'estrazione, la rappresentazione e la distribuzione della conoscenza mediante servizi avanzati e flessibili.

Una questione cruciale nella visione di un web semantico è il come fare a sostenere e facilitare l'Ontological Engineering, cioè il processo di creazione e gestione di ontologie. Infatti, ogni attività di progettazione e di mantenimento delle ontologie, comporta un elevato sforzo e una certa metodologia ben definita consistente in almeno uno studio di fattibilità, da una fase di analisi dei requisiti, da una fase di formalizzazione e di concettualizzazione, così come una fase di applicazione, di valutazione e di perfezionamento (Cimiano, Philipp, Johanna Volker, e Rudi Studer, 2006).



**Figura 5:** processo dell'Ontological Engineering (Cimiano, Philipp, Johanna Volker, e Rudi Studer, 2006).

Le ontologie sono state concepite per rappresentare una concettualizzazione di un dominio nel quale si suppone che riflettano l'accordo di una certa comunità o gruppo di persone. Nel caso più semplice, le ontologie sono costituite da un insieme di concetti o classi, che sono rilevanti per il dominio di interesse, e da una serie di

relazioni definite su questi concetti distinte generalmente in relazioni tassonomiche e relazioni non tassonomiche.

I rapporti della sottoclasse, o tassonomici, stabiliscono un ordine gerarchico fra i concetti, che è definito semanticamente tramite l'inclusione dell'insieme dei membri di una classe o di un concetto. Le relazioni non tassonomiche, invece, sono relazioni definite su concetti, ma non hanno lo scopo di ordinare i concetti gerarchicamente. I concetti e le relazioni possono essere definiti assiomaticamente, specificando proprietà aggiuntive, come la transitività o la cardinalità, come in sistemi di database. Idealmente, le ontologie dovrebbero essere abbastanza astratte al fine di essere riutilizzate in diverse applicazioni.

La questione cruciale è certamente il modo in cui tali ontologie possono essere costruite. Come illustrato precedentemente, la maggior parte delle metodologie di Ontology Engineering distinguono almeno le seguenti fasi: studio di fattibilità, analisi dei requisiti, concettualizzazione e, infine, la distribuzione, che in genere consiste in un ciclo di applicazione, valutazione e manutenzione dell'ontologia (vedi figura 5).

Queste fasi sono talvolta chiamate diversamente e, talvolta, suddivise in sottofasi.

La concettualizzazione, ad esempio, può essere separata in almeno tre sottoattività:

- lo sviluppo del modello di dominio;
- la formalizzazione del modello;
- la sua attuazione in un certa lingua dell'ontologia.

Le tecniche di apprendimento di un'ontologia possono essere, ad esempio, applicate nella fase di concettualizzazione come modello d'inizio per poter essere utilizzate come base per la discussione.

Una volta che questo modello è formalizzato e realizzato nella lingua formale dell'ontologia, le tecniche di apprendimento dell'ontologia possono essere applicate nella fase di mantenimento per ampliare, perfezionare o modificare il modello.

Per alleviare le spese sostenute per l'attività di Ontological Engineering, che è un processo costoso, sono emerse varie proposte per l'apprendimento automatico di dati da parte delle ontologie.

A tale scopo, negli ultimi anni vi è stato un grande aumento di interesse per i metodi che riguardano l'apprendimento delle ontologie da risorse testuali.

I metodi più importanti applicati nel processo di apprendimento dell'ontologia riguardano, in particolar modo, le seguenti attività (Cimiano, Philipp, Johanna Volker, e Rudi Studer, 2006):

- Estrarre da una raccolta di testo la terminologia e i sinonimi relativi al dominio.
- Scoprire concetti che possono essere considerati come astrazioni del pensiero umano.
- Derivare una gerarchia di concetto che organizza questi concetti.
- Ampliare una gerarchia di concetto esistente con nuovi concetti.
- Apprendere relazioni non tassonomiche fra concetti.
- Popolare l'ontologia con istanze di relazioni e concetti.
- Scoperta di altre relazioni o regole assiomatiche che implicano i concetti e le relazioni.

**Estrarre da una raccolta di testo la terminologia e i sinonimi relativi al dominio:**

estrarre la terminologia relativa ad un dominio, da una raccolta di testo, è un passo fondamentale all'interno del processo di apprendimento di un'ontologia. Di solito l'ipotesi asserita è che alcuni termini riferiscono chiaramente, senza ambiguità, un concetto a un dominio specifico. Al fine di individuare i sinonimi, gli approcci più comuni applicano tecniche per raggruppare insieme parole simili, o usano qualche misura di associazione per individuare coppie di termini correlati statisticamente. L'individuazione di sinonimi può aiutare i termini a costituire dei gruppi di termini, a condividere, o quasi, lo stesso significato, rappresentando così classi ontologiche.

In generale, i metodi per estrarre i termini e i sinonimi hanno conseguito una maturità ragionevole. I metodi di estrazione dei sinonimi, ad esempio, hanno dimostrato di raggiungere risultati quasi umani.

**Apprendere i concetti e le gerarchie di concetto:** la struttura portante di qualsiasi ontologia è costituita da un insieme di relazioni tassonomiche tra classi. Ognuna delle classi può essere definita, per esempio, da un'etichetta descrittiva o dalle sue relazioni con altre classi.

Dato che il nucleo della tassonomia di un'ontologia è di fondamentale importanza per l'uso delle ontologie come mezzo di astrazione, finora la maggior parte degli approcci si sono concentrati sulla formazione dei concetti e gerarchie di concetto.

I principali metodi, che sono stati applicati all'attività di apprendere relazioni di sottoclasse, sono tecniche di raggruppamento gerarchico non controllate riconosciute dal processo di apprendimento della macchina. Così, alcuni ricercatori hanno pensato di introdurre un controllo nel processo di raggruppamento, coinvolgendo direttamente l'utente a convalidare o rifiutare alcuni raggruppamenti o includendo informazioni esterne per guidare il processo di raggruppamento. Altri, invece, si basano sul concetto che alcuni modelli indicano, in modo affidabile, una relazione di interesse tra termini. Ad esempio un modello del tipo "X come Y" indica che Y è una sottoclasse di X. Sebbene questi approcci siano più o meno affidabili, non si verificano abbastanza spesso in dati di testo.

La soluzione proposta a questo problema è di accoppiare queste strutture sul web. Gli approcci basati sull'unione di tali modelli possono essere attuati in modo relativamente facile utilizzando le espressioni regolari che sono in genere abbastanza efficienti, in quanto fondamentalmente esse devono essere eseguite solo una volta durante la raccolta di testo.

Lo svantaggio concettuale di tali metodi è che essi scoprono essenzialmente rapporti lessicali fra parole, ma non tra concetti, che si suppone che siano astrazioni e non solo semplici parole.

**Ampliare una Gerarchia di concetto esistente:** questa attività consiste nell'ampliare una gerarchia di concetto con nuovi concetti. Ciò è possibile aggiungendo un nuovo concetto nella posizione appropriata della tassonomia già esistente. Sia i metodi controllati che quelli non controllati possono essere applicati per questo scopo. Nel caso di un approccio controllato, i classificatori, che predicano l'appartenenza per ogni concetto alla gerarchia di concetto esistente, hanno bisogno di essere addestrati.

Gli approcci non controllati assumono una funzione di somiglianza che calcola una misura di aggancio tra il nuovo concetto e i concetti che esistono nell'ontologia. Tali metodi si basano su un'appropriata rappresentazione contestuale di vari concetti resa possibile dalla similarità che può essere calcolata.

Alcuni di questi approcci, cioè quelli che si basano su un determinato insieme di istanze di classi inizialmente sconosciute, affrontano sia il problema della costruzione che quello di popolazione della tassonomia.

**Apprendere relazioni non tassonomiche fra concetti:** una volta stabilita una gerarchia tassonomica, molti degli attuali strumenti utilizzati nel processo di apprendimento dell'ontologia cercano di apprendere una serie di relazioni non tassonomiche che sono essenziali per esprimere le proprietà specifiche del dominio sia delle classi che delle istanze.

Alcuni ricercatori hanno sfruttato soprattutto i verbi che compaiono nel testo come indicatori di una relazione tra le loro argomentazioni.

In generale, sebbene di solito la qualità di questi approcci sia ragionevole, le relazioni dovranno essere ispezionate e convalidate da un tecnico che si occupa della progettazione delle ontologie. Un problema importante per l'estrazione di relazioni, tuttavia, è trovare un adeguato livello di generalizzazione di queste relazioni.

**Popolazione di ontologie:** consiste essenzialmente nell'aggiungere all'ontologia istanze di concetti e relazioni. Per la popolazione di ontologie con istanze di concetto, gli approcci si basano sull'abbinamento di certe strutture lessico-sintattiche. Per l'attività di apprendimento delle relazioni di istanze, sono stati principalmente esplorati gli approcci *bootstrapping*, come vengono definiti da Cimiano, Volker e Studer (2006, p. 317), che raccolgono le relazioni tuple sul web. Una tupla è una riga di una matrice bidimensionale delle relazioni di un database relazionale. Essa rappresenta l'entità base che deve essere memorizzata nel database. Più tabelle di un database possono, ad esempio, essere messe in relazione definendo i campi che, se contengono valori uguali nelle due tabelle, generano un corrispondente blocco di tuple, formato dall'unione dei dati delle due tabelle.

Solitamente, sembra che gli approcci di popolazione delle ontologie abbiano acquisito una certa maturità e funzionino bene.

### 3.2 Tesi sostenuta da Aleksander Pivk.

La tesi presentata da Aleksander Pivk propone un nuovo metodo per la generazione automatica di modelli di conoscenza, come le ontologie, da strutture tabellari arbitrarie. Queste strutture sono direttamente applicabili a vari scenari, per esempio per rispondere alle query (v. figura 6).

In particolar modo, il suo studio riguarda l'analisi di strutture tabellari che mirano a sfruttare la loro struttura parziale e le abitudini di modellazione cognitive degli esseri umani. Infatti, come afferma nella discussione del suo articolo, la comprensione del contenuto della tabella richiede un'interpretazione semantica, oltre alla conoscenza della struttura della tabella.

Il suo approccio al problema della generazione automatica di ontologie si basa sul modello di Hurst ed ha come obiettivo, appunto, l'analisi e la trasformazione automatica di strutture tabellari arbitrarie in modelli di conoscenza, cioè in ontologie. Questo modello analizza le tabelle sotto l'aspetto grafico, fisico, strutturale, funzionale e semantico.

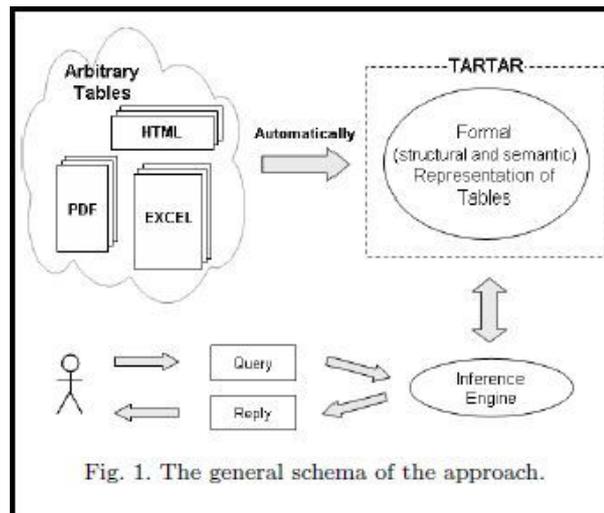
Prendendo spunto da tale modello, il metodo di Aleksander Pivk pone l'attenzione in particolar modo sulla dimensione fisica, strutturale, funzionale e semantica:

- **Primo passo (dimensione fisica):** prevede che una tabella sia estratta, pulita, canonizzata e trasformata in una matrice di forma regolare.
- **Secondo passo (dimensione strutturale):** ha l'obiettivo di rilevare la struttura della tabella. Questo è un compito molto complesso, in quanto ci sono moltissime varianti di layout di tabella. Le tre sottoattività più importanti sono: (a) determinare l'orientamento di lettura di tabella, scoperto misurando la distanza tra celle e perciò tra righe e colonne; (b) smembrare una tabella in unità logiche e ulteriormente in singole sezioni, in modo che le sezioni siano composte solo da celle di attributo o di istanza; (c) stabilire il tipo di tabella, che deve appartenere a uno dei cinque tipi predefiniti.
- **Terzo passo (dimensione funzionale FTM):** viene costruito il modello di tabella funzionale (FTM) rappresentato come un grafico diretto aciclico, che riordina i campi della tabella in modo da produrre un percorso globale per

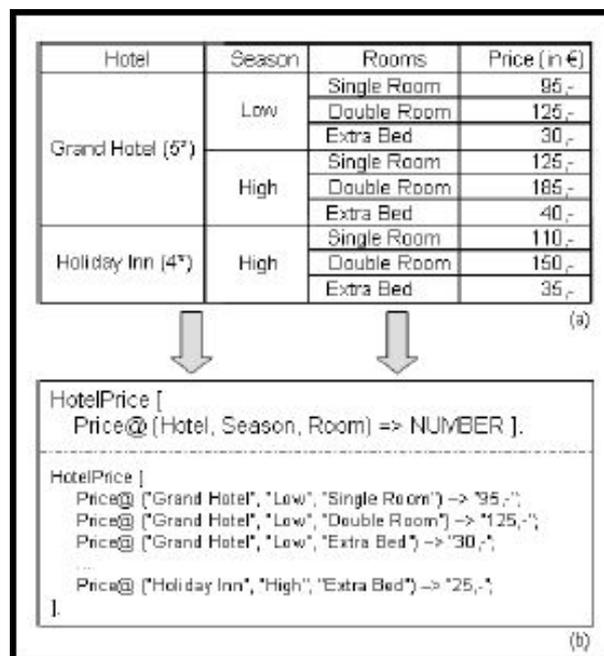
ogni singola cella. Dopo avere completato la struttura FTM viene eseguito anche il suo riepilogo, con l'obiettivo di ridurre al minimo il modello.

- **Quarto passo (dimensione semantica):** vengono individuate etichette semantiche per i campi della tabella, dove sono utilizzati l'ontologia lessicale WordNet e il servizio GoogleSets. Queste etichette semantiche servono come annotazioni di nodi FTM e sono poi utilizzate anche all'interno di strutture formali in uscita.

Alla fine del processo di analisi e di trasformazione, un frame e un'ontologia sono estratti da un FTM. Il frame rende esplicito il significato del contenuto della cella (dimensione funzionale) e il significato della tabella in base alla sua struttura. Nell'ontologia i concetti sono disposti in un grafico diretto aciclico, dove gli archi rappresentano le relazioni tra i concetti e anche i tipi di relazioni. Infine, dalla formalizzazione dei contenuti della tabella, secondo le strutture formali appena generate, viene creata una base di conoscenza (vedi figura 7).



**Figura 6:** schema generale dell'approccio proposto (A. Pivk, Automatic ontology generation from web tabular structures).



**Figura 7:** esempio di trasformazione di una tabella (a) e una corrispondente parte di una base di conoscenza (b). (A. Pivk, Automatic ontology generation from web tabular structures).

Il metodo di trasformazione proposto da Pivk è attuato all'interno di un sistema chiamato TARTAR (Transforming ARbitrary TABLES into fRAMES), composto da 15.000 righe di codice scritto in linguaggio di programmazione Java (Department of Intelligent Systems, voce A. Pivk). Un aspetto fondamentale di questo metodo è che non è vincolato ad un unico tipo di documento né di dominio, quindi può essere applicato a qualsiasi formato di tabella (html, excel, pdf, ecc.). Questo rende l'approccio molto flessibile e facilmente estendibile. Infatti, non ci sono domini specifici per le operazioni e nessuna conoscenza specifica è inclusa nell'approccio.

La valutazione empirica di questo metodo fornisce buoni risultati in termini di efficienza e usabilità ed è quindi applicabile in pratica (Pivk 2006). I risultati ottenuti dalla valutazione hanno dimostrato oltre l'80% del tasso di successo della trasformazione automatica di tabelle in rappresentazioni semantiche e la precisione del 100% nell'attività di risposta automatica alle query sul contenuto della tabella.

Nella sezione 4.2 verrà messo in pratica una parte del procedimento descritto da Pivk utilizzando Calais (v. 4.2).

## 4. Un approccio che sfrutta Open Calais per la ricerca di concetti.

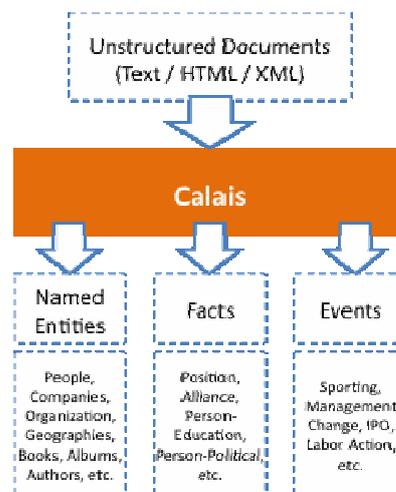
### 4.1 Introduzione a Open Calais



Calais è un progetto open source finanziato dalla famosa agenzia di stampa Reuters e realizzato dal team di ClearForest.

E' un Web Service in grado di analizzare semanticamente un documento, contenente testo non strutturato, e restituire metadati in formato XML/RDF.

Sostanzialmente è uno strumento di tagging, che arricchisce semanticamente contenuti Web testuali mediante la generazione di metadati, permettendo così di estrarre dai documenti informazioni su luoghi, eventi, persone, quasi in real time e di metterle poi in correlazione tra loro, creando categorie e rendendo i contenuti più semplici da trovare e da leggere. Questi metadati arricchiti semanticamente sono utili, ad esempio, per la costruzione di mappe per collegare documenti a persone o a società. Queste mappe possono essere utilizzate per migliorare la navigazione nel sito, per la diffusione dei contenuti, analizzare il contenuto per vedere se contiene ciò che interessa e altro ancora.



**Figura 10:** processo di elaborazione eseguito da Calais.

Sono numerosi i tool costruiti a partire da Calais tra cui Gnosis, un plugin per Firefox, che analizza automaticamente la pagina Web visualizzata dal browser e fornisce alcuni strumenti per cercare persone, compagnie, luoghi e cose.

Di seguito vengono illustrate le tre componenti principali di Calais:

- Il *Calais Web Service*, che si occupa di generare automaticamente metadati in formato RDF.
- Una serie di semplici applicazioni che servono come punto di partenza per dimostrare le potenzialità di Calais e come stimolo agli sviluppatori per la creazione di nuove applicazioni. Il progetto Calais è infatti completamente gratuito e open. Basta registrarsi e richiedere una *developer key* per ottenere le API (*Application Programming Interface*, ovvero Interfaccia di Programmazione di un'Applicazione). Consente di creare applicazioni e plugins per piattaforme di blogging e per CMS (*Content management system*, cioè un sistema di gestione dei contenuti).
- Un supporto attivo per gli sviluppatori tramite un *forum*.

E' disponibile anche un *viewer* molto semplice in grado di visualizzare i metadati semantici di un file RDF, il Calais Document Viewer.

Ed ecco come funziona il servizio:

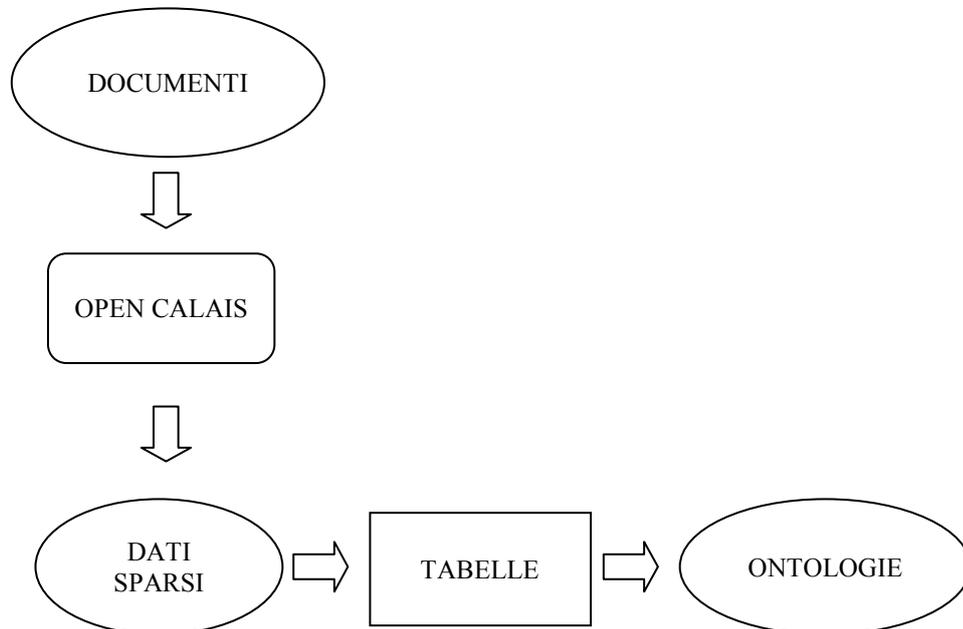
- 1) L'utente fornisce a Calais dei contenuti testuali da esaminare, presenti, ad esempio, all'interno di pagine HTML;
- 2) Analizzando questi contenuti, nonché sulla base di elaborazioni condotte utilizzando il linguaggio naturale ed altri metodi, Calais estrapola informazioni strutturate dal testo, essendo capace di riconoscere persone, organizzazioni, luoghi, fatti ed eventi;
- 3) Successivamente, Calais genera dei metadati in formato RDF, cioè dei dati descrittivi di ogni entità, evento o fatto riconosciuto all'interno del testo esaminato, associandoli permanentemente al documento Web al quale essi appartengono, documento a cui viene inoltre assegnato un identificativo unico (il Globally Unique Identifier - GUID);

- 4) Utilizzando il GUID, qualunque applicazione semantica sul Web delegata dall'utente, e conforme agli standard RDF, potrà di conseguenza leggere e comprendere questi metadati, potendo poi utilizzarli per i più svariati scopi.

#### 4.2 Considerazioni sull'applicazione di Calais: un esempio pratico.

Open Calais pone le basi per mettere in pratica le soluzioni proposte da alcuni studiosi sul problema della generazione automatica di ontologie. Infatti, elaborando dei documenti con Open Calais si ottengono dei dati sparsi arricchiti semanticamente attraverso metadati. Se a questi dati viene applicata, per esempio, la tesi proposta da Aleksander Pivk (v. 3.1), essi subiscono una serie di elaborazioni che danno come risultato la generazione di un'ontologia.

Rappresentazione schematica del processo appena descritto:



L'esempio proposto di seguito vuole dimostrare le potenzialità di Calais illustrate precedentemente.

A tale scopo un articolo del “The New York Times” è stato elaborato con Calais e successivamente ai dati ottenuti è stato applicato una parte del metodo di Pivk:

Microsoft and Yahoo Executives Meet  
By MIGUEL HELFT  
Published: January 16, 2009

SAN FRANCISCO — Steven A. Ballmer, Microsoft’s chief executive, met in New York earlier this week with Roy J. Bostock, Yahoo’s chairman, according to a person briefed on the meeting.

It was the second high-level contact between Microsoft and Yahoo to come to light this week, and suggests that negotiations over a Web search deal between the two companies could be restarting. On Wednesday, Carol Bartz, Yahoo’s new chief executive, told employees she had had an informal conversation with Mr. Ballmer since she was selected by the board to lead Yahoo.

Details of the conversations could not be learned, and spokesmen for Yahoo and Microsoft declined to comment.

In recent weeks, Mr. Ballmer has said repeatedly that Microsoft remains interested in doing a deal with Yahoo involving its Web search and advertising business, and that he favored coming to an agreement sooner rather than later.

The appointment of Ms. Bartz on Tuesday renewed speculation that Yahoo might be more willing to consider a deal to hand over its search business to Microsoft. Last year, after merger talks between the two companies collapsed, Microsoft and Yahoo discussed various ways to combine their search assets, including an outsourcing agreement or an outright sale of Yahoo’s search business to Microsoft, but the negotiations did not bear fruit.

Some on Yahoo’s board and the company’s top brass, including Jerry Yang, then the chief executive, and Susan L. Decker, who was then president, were cool to the idea of giving up control over the search business. They said that advertisers increasingly wanted to buy search-related ads and display ads together, and that separating the two would not be in Yahoo’s long-term interest.

After the appointment of Ms. Bartz, Mr. Yang returned to his role of “chief Yahoo” and Ms. Decker resigned, though she remains with the company during a transition period.

But many Yahoo shareholders say they believe that short of an outright merger, a search deal would be the best outcome for Yahoo, and have pressed the company to pursue one.

On Tuesday, Ms. Bartz acknowledged that deciding whether a search deal made sense was one of the big strategic questions she faced, but she declined to give her opinion about it. During the employee meeting the next day, she said her “gut” suggested that it was not a good idea, but she added that she would spend time studying the matter. News reports of her remarks caused Yahoo shares to drop more than 6 percent on Thursday. They fell another 2 cents Friday, to close at \$11.59.

Word of a meeting between Mr. Ballmer and Mr. Bostock was first discussed by Valleywag, a Silicon Valley gossip blog.

If any negotiations or discussions between the two companies are restarting, they are likely to be preliminary. Two people close to the companies said that currently no investment bankers were involved in any discussions.

Yahoo has also been in on-and-off discussions about a possible merger with the AOL Internet unit of Time Warner, which that company has been trying to sell. The conversations are continuing, according to people close to the companies.

A Time Warner spokesman declined to comment.

**Figura 11:** articolo scritto da Miguel Helft, pubblicato il 16/01/2009 sul quotidiano “The New York Times”.

Le figure seguenti illustrano il risultato ottenuto elaborando, con Calais, l'articolo estratto dal quotidiano "The New York Times":

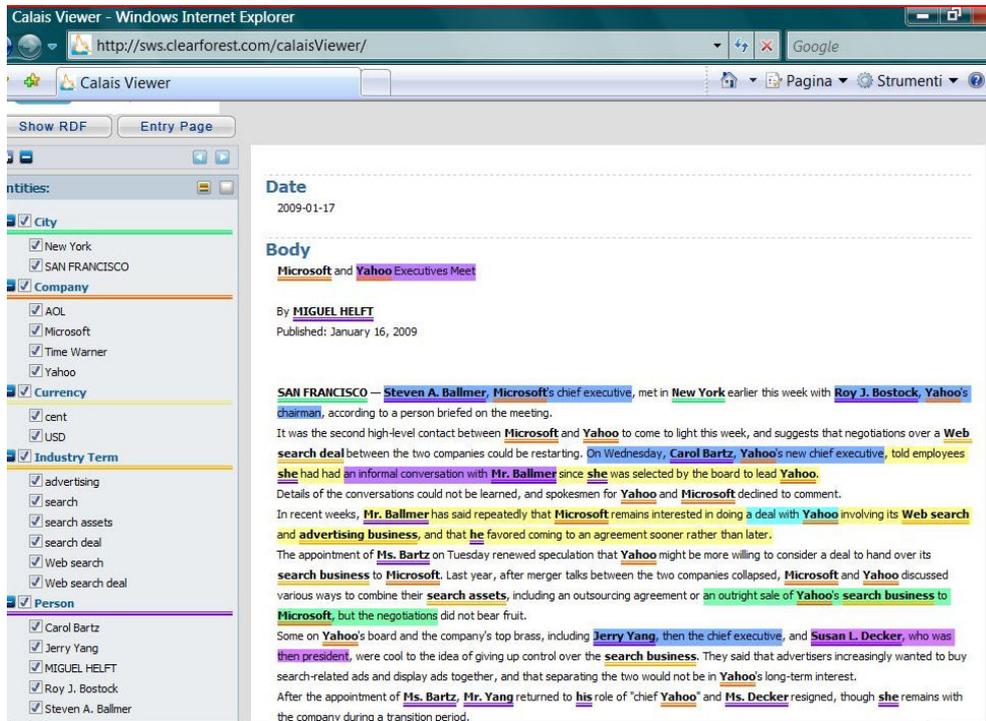


Figura 12: estrapolazione delle informazioni strutturate utilizzando Calais.

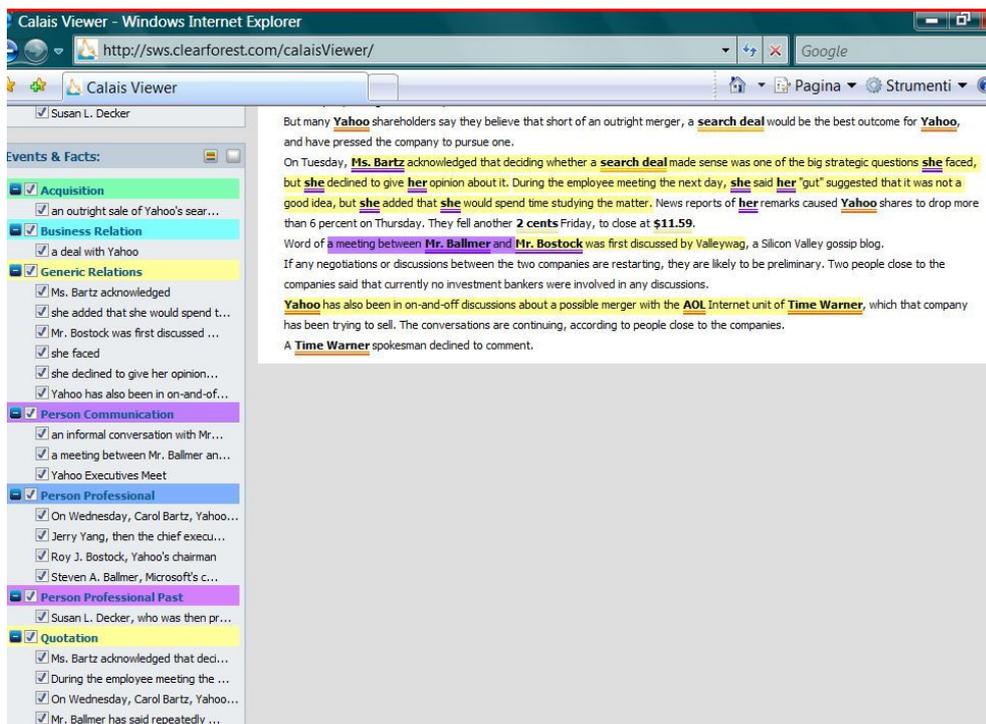


Figura 13: estrapolazione delle informazioni strutturate utilizzando Calais.

CALAIS  
 Powered by Thomson Reuters  
 Show Results    Entry Page

```

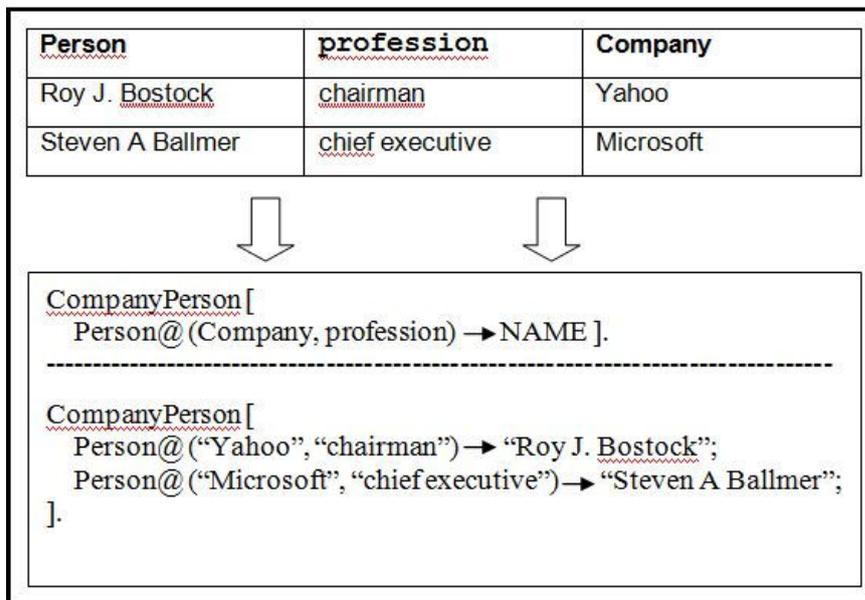
<!--Use of the Calais Web Service is governed by the Terms of Service located at
http://www.opencalais.com. By using this service or the results of the service you agree to
these terms of service.--><!--Relations: Acquisition, BusinessRelation, GenericRelations,
PersonCommunication, PersonProfessional, PersonProfessionalPast, Quotation

City: New York, SAN FRANCISCO
Company: AOL, Microsoft, Time Warner, Yahoo
Currency: USD, cent
IndustryTerm: Web search, Web search deal, advertising, search, search assets, search deal
Person: Carol Bartz, Jerry Yang, MIGUEL HELFT, Roy J. Bostock, Steven A. Ballmer, Susan L.
Decker--><rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:c="http://s.opencalais.com/1/pred/"><rdf:Description c:allowDistribution="true"
c:allowSearch="true" c:calaisRequestID="0a45dd5e-6bbd-4f62-a30e-elc229618756"
c:externalID="calaisbridge" c:id="http://id.opencalais.com/FMLLTrI3VrbfSXnDgQ6Ldg"
rdf:about="http://d.opencalais.com/dochash-1/1395ed5c-9bef-342d-bc7c-20d9c26f173e"><rdf:type
rdf:resource="http://s.opencalais.com/1/type/sys/DocInfo"/><c:document><![CDATA
[<Document><Date>2009-01-17</Date><Body>Microsoft and Yahoo Executives Meet

By MIGUEL HELFT
Published: January 16, 2009
  
```

**Figura 14:** metadati in formato RDF generati da Calais.

Applicando una parte della procedura del metodo illustrato da A. Pivk, ai dati estratti ed elaborati da Calais dell'articolo di Miguel Helft, otteniamo una tabella con la corrispondente parte della base di conoscenza:



**Figura 15:** rappresentazione di una parte del procedimento descritto da A. Pivk applicato ai dati estratti da Calais dall'articolo scritto da Miguel Helft, pubblicato il 16/01/2009 sul quotidiano "The New York Times".

## 5. Conclusioni e prospettive

Le ontologie, come abbiamo visto, sono uno strumento essenziale della nuova rete ma crearle non è semplice. Occorre una buona conoscenza del dominio in cui si opera e spesso la vastità dei dati di partenza rende il lavoro alquanto complicato. Molti studiosi si sono messi all'opera per realizzare dei programmi che svolgano per noi questo compito.

Dagli studi sull'argomento abbiamo potuto vedere che le tecniche di apprendimento automatico dell'ontologia sono ancora agli inizi, dato che in molti casi le ontologie generate non soddisfano le esigenze delle applicazioni previste. Esse, comunque, hanno già il potenziale per migliorare alcune applicazioni classiche, come il recupero delle informazioni, e per contribuire a strutturare enormi raccolte di risorse applicando tecniche di estrazione di testo.

Per garantire l'applicabilità del processo di apprendimento dell'ontologia in varie aree di applicazione saranno, dunque, necessari sia metodi e strutture altamente configurabili, sia una stretta integrazione con approcci di valutazione delle ontologie, manuali o automatici. Ciò sembra fondamentale per investire nello sviluppo di nuove metodologie che siano in grado di integrare i risultati dei sistemi di apprendimento dell'ontologia, riducendo al minimo l'intervento dell'utente e massimizzando la qualità delle ontologie per quanto riguarda un particolare dominio o applicazione.

Per concludere, l'indagine svolta ha inoltre evidenziato che, ad oggi, non esiste un programma in grado di costruire in maniera automatica delle ontologie.

Ad ogni modo, è stata analizzata una proposta di notevole interesse che rappresenta un buon punto di partenza nella generazione automatica di ontologie e fa ben sperare riguardo a futuri miglioramenti.

## 6. Bibliografia

- Capuano, Nicola. 2005. *Ontologie OWL: Teoria e Pratica*. “Computer Programming”, 148, pp. 59-64.
- Cimiano, Philipp, Johanna Volker, e Rudi Studer. 2006. *Ontologies on Demand? A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text*. In *Information, Wissenschaft und Praxis* 57 (6-7): 315-320.
- Esposito, Floriana, Nicola Fanizzi, Luigi Iannone, Ignazio Palmisano, Domenico Redavid, e Giovanni Semeraro. 2005. *Un Algoritmo per l'Apprendimento di Concetti Basato su Controfattuali*.
- Gruber T.R. 1993. *A translation approach to portable ontology specifications*. *Knowledge Acquisition*, 5(2), pp. 199-220.
- Pivk, Aleksander. 2006. *Automatic Ontology Generation from Web Tabular Structures*. “*AI Communications*”, vol. 19(1), pp. 83-85.
- Shun-Hong Sie, Jian-Hua Yeh. 2006. *Automatic Ontology Generation Using Schema Information*. *International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*.

### Siti consultati

Astrea, voce *dr. Paolo Naggar - prof. Roberto Basili*

[http://astrea.cineca.it/stdoc/astrea\\_dss\\_naggar\\_basili.pdf](http://astrea.cineca.it/stdoc/astrea_dss_naggar_basili.pdf)

Dipartimento di Informatica e Sistemistica dell'Università “La Sapienza” di Roma, voce *Daniele Nardi*

<http://www.dis.uniroma1.it/~nardi/Didattica/RC/lezioni/sbc.pdf>

Laboratorio di Ontologia Applicata (LOA)

[http://www.loa-cnr.it/index\\_ita.html](http://www.loa-cnr.it/index_ita.html)

Open Calais

<http://www.opencalais.com/>

Pivk Aleksander

<http://dis.ijs.si/sandi/index.html>

Politecnico di Torino, voce *corsi a distanza*

<http://corsiadistanza.polito.it/corsi/pdf/01BHKH/Capitolo4.pdf>

Polo Didattico e di Ricerca di Crema, voce *Paolo Ceravolo*

[ra.crema.unimi.it/turing/MATERIALE/paolo/allegati/Ontometro.pdf](http://ra.crema.unimi.it/turing/MATERIALE/paolo/allegati/Ontometro.pdf)

Semantic Web

[http://semanticweb.org/wiki/Main\\_Page](http://semanticweb.org/wiki/Main_Page)

TSW, voce Andrea Morandini

[http://blog.tsw.it/search-engine-optimization/il-reperimento-delle-informazioni-sul-web-dalle-keyword-alle-reti-semantiche\\_000388.html](http://blog.tsw.it/search-engine-optimization/il-reperimento-delle-informazioni-sul-web-dalle-keyword-alle-reti-semantiche_000388.html)

Web Semantico

<http://www.websemantico.org/index.php>

Wikipedia

[http://it.wikipedia.org/wiki/Pagina\\_principale](http://it.wikipedia.org/wiki/Pagina_principale)

World Wide Web Consortium

<http://www.w3.org>

<http://www.w3c.it/> (w3c - Ufficio Italiano)