Word Integration

Amedeo Leo[1] Alessio Petrozziello[1]
Simone Romano[1]

[1]Università degli studi di Salerno

Presentazione progetto Gestione Avanzata dei Dati
2014/2015

1 Introduzione

2 Fonti

3 Wrapper

4 Schema globale

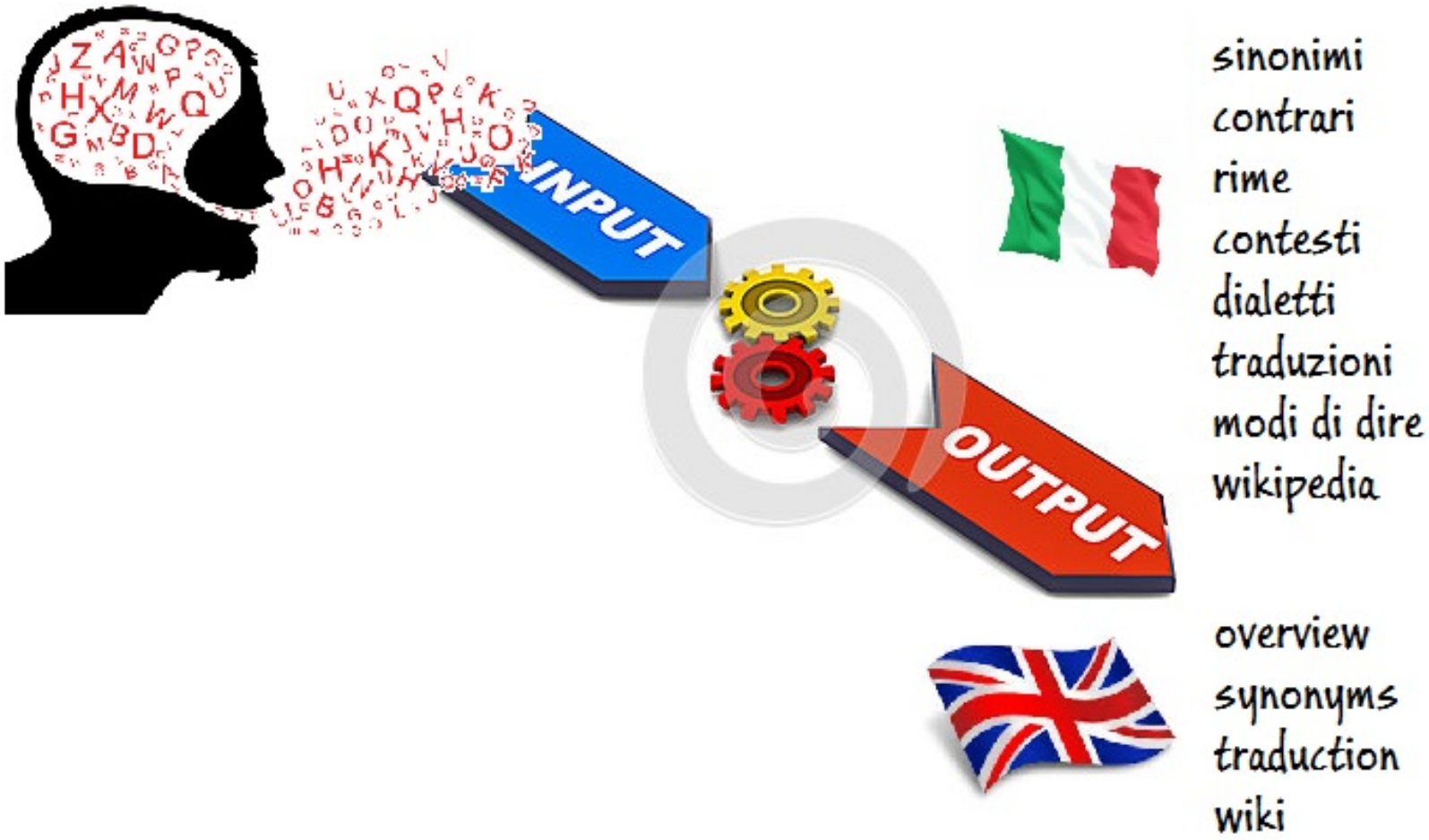
5 Tecnologie

Outline

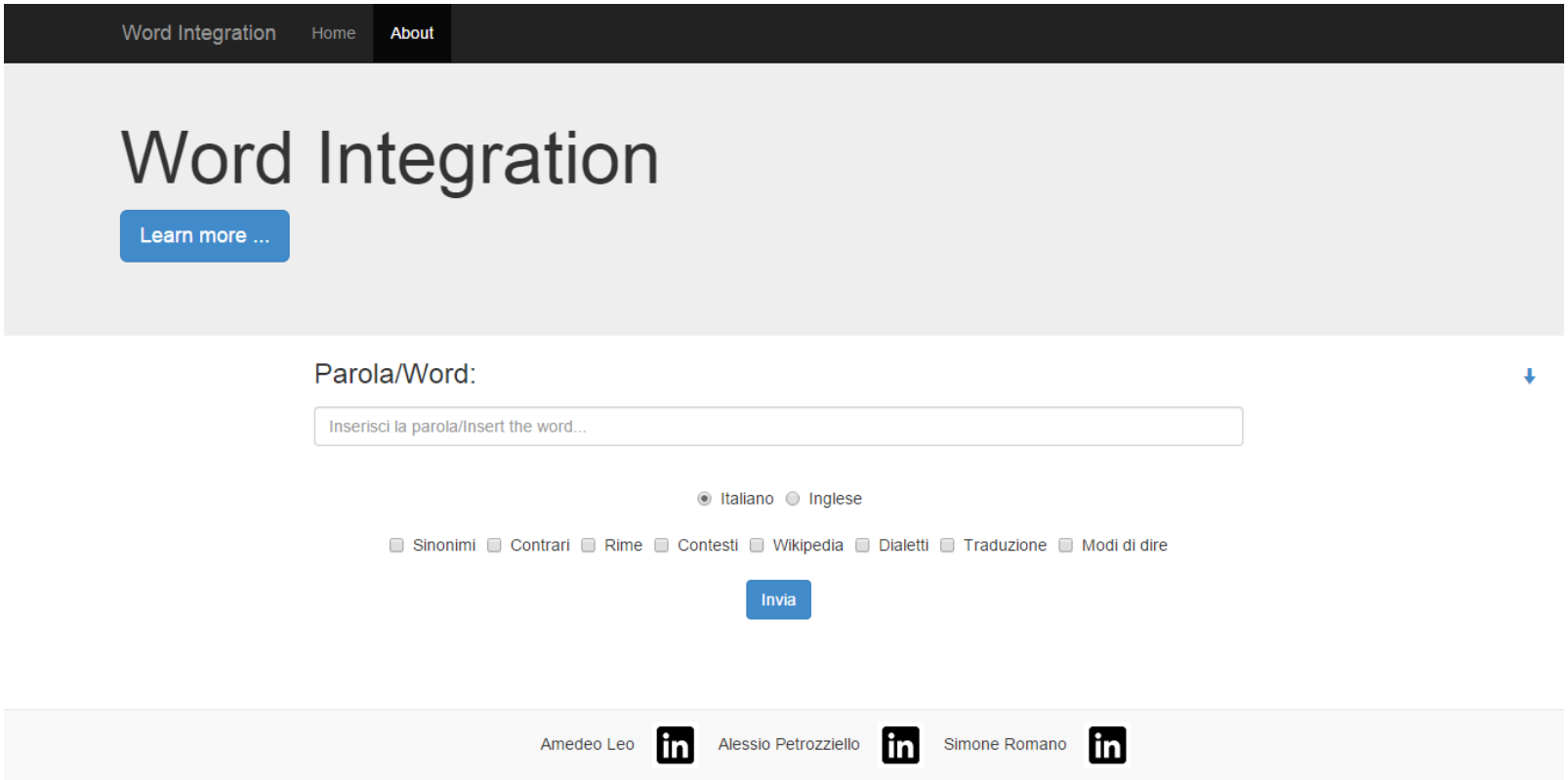
Specifica

Obiettivo:

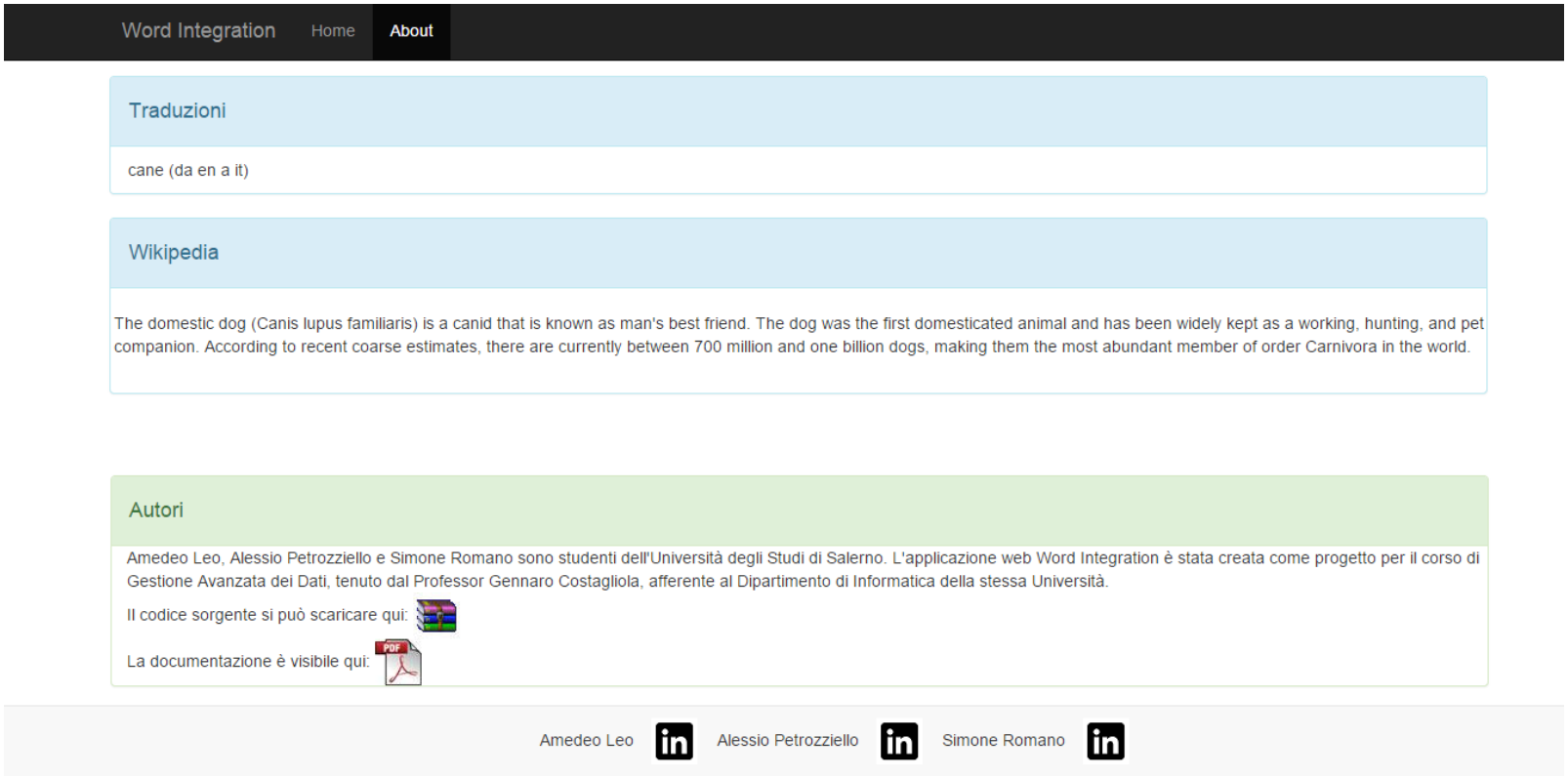
- Generare informazioni di diverso tipo su una parola(italiana/inglese) data in input.



Screen - 1

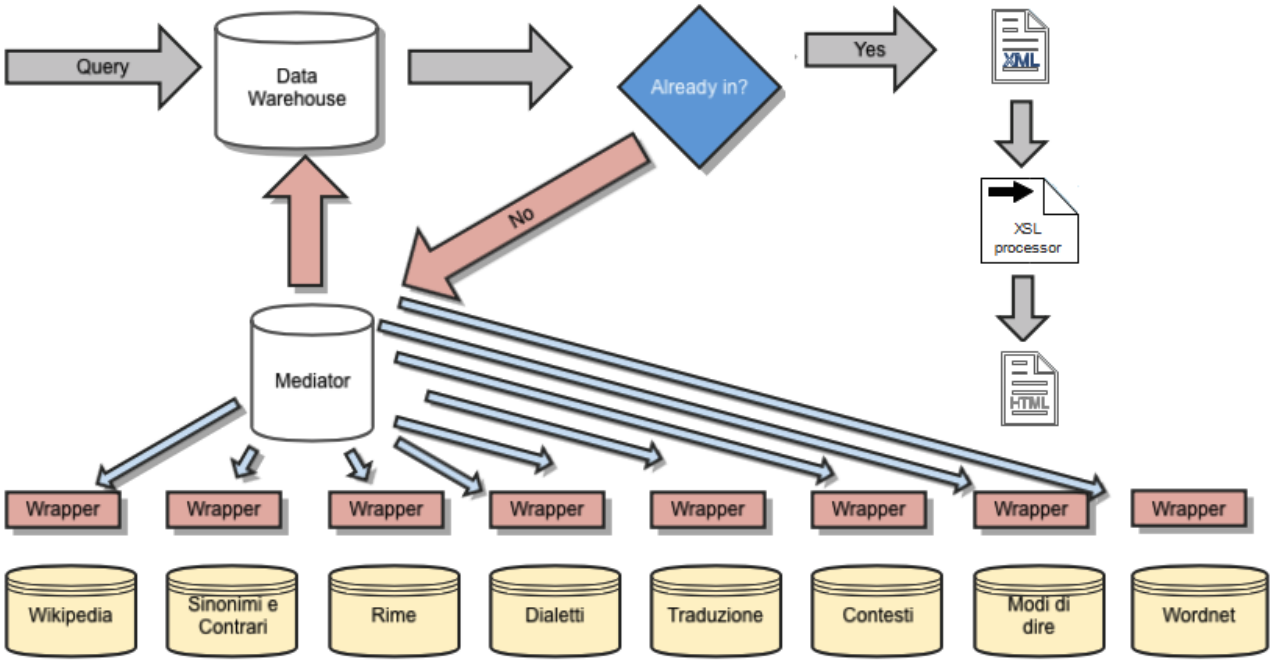

Word Integration
Home
About
Word Integration
Learn more ...
Parola/Word:
Inserisci la parola/Insert the word...
Italiano
Inglese
Sinonimi
Contrari
Rime
Contesti
Wikipedia
Dialetti
Traduzione
Modi di dire
Invia
Amedeo Leo
Alessio Petrozziello
Simone Romano

Screen - 2


Word Integration
Home
About
Traduzioni
cane (da en a it)
Wikipedia
The domestic dog (Canis lupus familiaris) is a canid that is known as man's best friend. The dog was the first domesticated animal and has been widely kept as a working, hunting, and pet companion. According to recent coarse estimates, there are currently between 700 million and one billion dogs, making them the most abundant member of order Carnivora in the world.
Autori
Amedeo Leo, Alessio Petrozziello e Simone Romano sono studenti dell'Università degli Studi di Salerno. L'applicazione web Word Integration è stata creata come progetto per il corso di Gestione Avanzata dei Dati, tenuto dal Professor Gennaro Costagliola, afferente al Dipartimento di Informatica della stessa Università.
Il codice sorgente si può scaricare qui:
La documentazione è visibile qui:
Amedeo Leo
Alessio Petrozziello
Simone Romano

Architettura

Architettura ibrida



Outline

Overview

Possiamo raggruppare le fonti utilizzate in 2 categorie:

1. **Italiane**
   - http://www.sinonimi-contrari.it STATICA
   - http://www.cercarime.it/?rima STATICA
   - http://www.dialettando.com/dizionario/ CAMBI PERIODICI
   - Google translator API STATICA
   - Wikipedia API STATICA
   - Modi di dire STATICA
   - http://web-ngram.research.microsoft.com/similarity/ CAMBI PERIODICI
2. **Inglesi**
   - Google translator API STATICA
   - WordNet API STATICA
   - Wikipedia API STATICA

Datalog

- SinonimiContrari(Parola, Sinonimo, Contrario)
- Cercarime(Parola, Rima)
- Contesti(Parola, Contesto)
- Dialetti(Parola, Dialetto, Regione)
- GoogleTranslator(Parola, Da, A, Traduzione)
- WordNet(Parola, Overview, Synonims)
- Wikipedia(Parola, Descrizione)
- Modi(Parola, Modo)

Outline

Scraping

Come si evince dall'architettura mostrata, a runtime avvengono i seguenti passaggi:

- Ricerca dei dati nel Database:
  - **Dati presenti**: Generazione output
  - **Dati non presenti**: scraping che mira a popolare il database e generazione output

Strumenti usati

Gli strumenti usati per ricavare le *xpath* sono i seguenti:

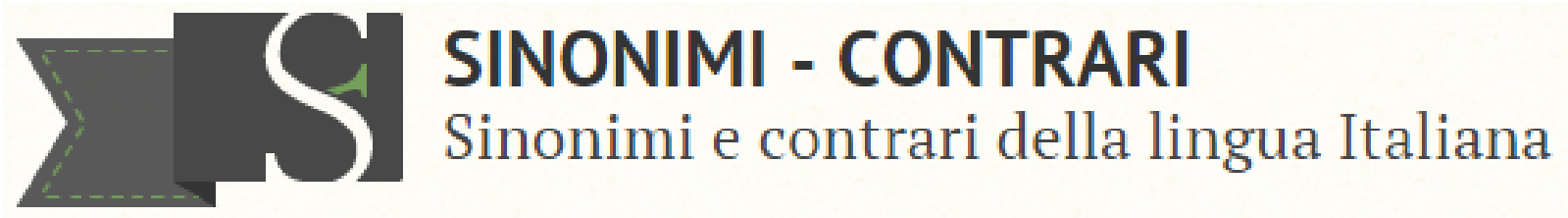
- **Curl** per:
  - Estrazione html
  - Generazione xhtml
- **Pentaho Data Integration (Kettle)** per ottenere le xpath
- **Exist** per testare le xpath

Vediamo dunque le xpath di scraping per ciascuna fonte.

Sinonimi e contrari

Per la fonte *http://www.sinonimi-contrari.it* le xpath utilizzate sono:

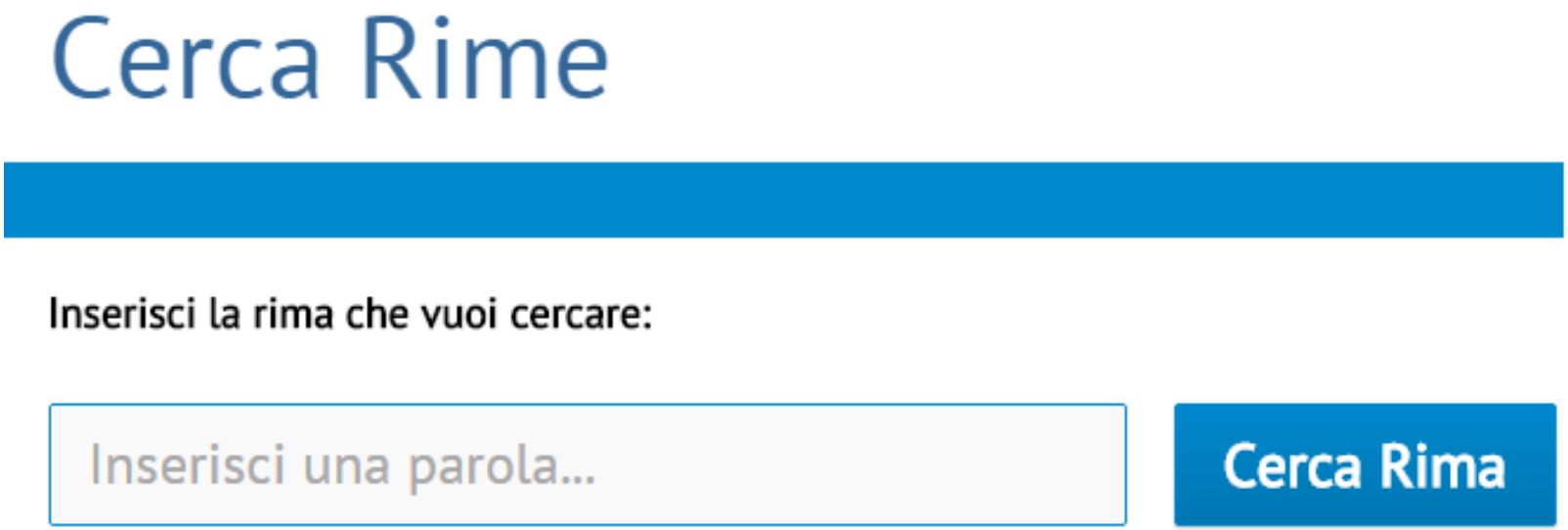
- /*[name()='html']/*[name()='body']/*[name()='div']/*[name()='div']/*[name()='div']/*[name()='div'][1]/*[name()='ol']/*[name()='li']/*[name()='span']/*[name()='a']/text()
- /*[name()='html']/*[name()='body']/*[name()='div']/*[name()='div']/*[name()='div']/*[name()='div'][3]/*[name()='ol']/*[name()='li']/*[name()='span']/*[name()='a']/text()



Rime

Per la fonte *http://www.cercarime.it/?rima* le xpath utilizzate sono:

- VECCHIO /*[name()='html']/*[name()='body']/*[name()='div']/*[name()='div']/*[name()='div']/*[name()='div'] /*[name()='ul']/*[name()='li']/*[name()='p']/*[name()='i']/*[name()='a']
- NUOVO //*[name()='div']/*[name()='div'][1]/*[name()='ul'][1]/*[name()='li']/*[name()='p']/*[name()='i'] /*[name()='a']/text()



Contesti

Per la fonte *http://web-ngram.research.microsoft.com/similarity/* è stato sufficiente utilizzare la seguente sintassi *php*:

- *valori* = *explode*(”” ,a);
  valori[0] = *substr*(valori[0], 3);

Clustering words based on context similarity

Dialetti

Per la fonte *http://www.dialettando.com/dizionario/* è stata utilizzata la seguente espressione xpath:

- /*[name()='html']/*[name()='body']/*[name()='center']/*[name()='table']/*[name()='tr'][5]
  /*[name()='td'][2]/*[name()='table']/*[name()='tr'][1]/*[name()='td'][2]/*[name()='center']/*[name()='table']
  /*[name()='tr']/*[name()='td']/text()

Google translate

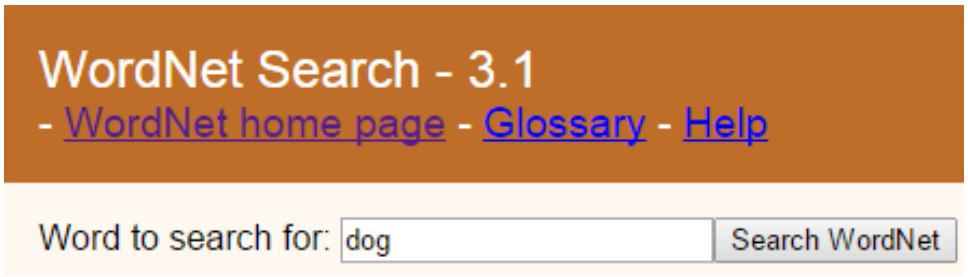
Per la traduzione sono state utilizzate delle API che interagiscono con Google translate:

- $obj = newGoogleTranslate(\$da, \$a);$
  $obj->translate(\$parola);$

Wordnet

Per la fonte *wordnet.princeton.edu/* sono state utilizzate delle API:

- "/usr/local/bin/wn ". $parola ." -over > scraping.html";
  "/usr/local/bin/wn ". $parola ." -synsn > scraping.html";



Wikipedia

Per la fonte *www.wikipedia.com/* sono state utilizzate delle API che estraggono il contenuto dell'output di wikipedia e lo ripuliscono restituendo una stringa:

- *html* = *file_get_contents*(url);
  info = *get_text*(html);
  info = *strip_tags*(info);
  info = *clean_input*(info);

Modi di dire

Per la fonte *http://dizionari.corriere.it/dizionario-modi-di-dire/* è stata utilizzata la seguente xpath:

- /*[name()='html']/*[name()='body']/*[name()='div'][9]/*[name()='div']/*[name()='div']/*[name()='div'][2]/*[name()='div'][1]/*[name()='div'][2]/*[name()='div']/*[name()='a']/text()

Outline

Schema globale

Lo schema globale del sistema di interrogazione può essere riscritto come segue:

- SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione,Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo) :-
  SinonimiContrari(Parola, Sinonimo, Contrario) $\vee$ Cercarime(Parola, Rima) $\vee$ Contesti(Parola, Contesto) $\vee$ Dialetti(Parola, Dialetto, Regione) $\vee$ GoogleTranslator(Parola, Da, A, Traduzione) $\vee$ WordNet(Parola, Overview, Synonims) $\vee$ Wikipedia(Parola, Descrizione) $\vee$ Modi(Parola, Modo)

Query

Riportiamo un **sottoinsieme di query** effettuabili sul sistema:

1. Ricerca di un sinonimo e di una rima della parola "'gatto"'
   q(Sinonimo, Rima) = SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione, Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo) $\wedge$ Parola = 'gatto' $\wedge$ Da='it' $\wedge$ A='en'
2. Ricerca tutte le informazioni della parola italiana "'bello"'
   q(Sinonimo, Contrario, Rima, Dialetto, Modo, Wikipedia) = SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione, Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo) $\wedge$ Parola = 'gatto' $\wedge$ Da='it' $\wedge$ A='en'
3. Ricerca informazioni su Informazioni Wikipedia e la traduzione della parola "'dog"'
   q(Wikipedia, Traduzione) = SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione, Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo) $\wedge$ Parola = 'dog' $\wedge$ Da='en' $\wedge$ A='it'
4. Prendere i dialetti e le rime della parola "'cavallo"'
   q(Dialetto, Rima) = SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione, Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo) $\wedge$ Parola = 'cavallo' $\wedge$ Da='it' $\wedge$ A='en'

GAV

Di seguito la riformulazione GAV delle interrogazioni:

1. q'(Sinonimo, Rima) = SinonimiContrari(Parola, Sinonimo, Contrario) $\vee$ Cercarime(Parola, Rima) $\wedge$ Parola = "gatto" $\wedge$ Da = "it" $\wedge$ A = "en"
2. q'(Sinonimo, Contrario, Rima, Dialetto, Regione, Modo, Wikipedia) = SinonimiContrari(Parola, Sinonimo, Contrario) $\vee$ CercaRima(Parola, Rima) $\vee$ Dialetti(Parola, Dialetto, Regione) $\vee$ Modi(Parola, Modo) $\vee$ Wikipedia(Parola, Descrizione) $\wedge$ Parola = "bello" $\wedge$ Da = "it" $\wedge$ A = "en"
3. q'(Wikipedia, Traudzione) = Wikipedia(Parola, Descrizione) $\vee$ GoogleTranslator(Parola, Da, A) $\wedge$Parola = "dog" $\wedge$ Da = "en" $\wedge$ A = "it"
4. q'(Dialetto, Rima) = CercaRime(Parola, Rima) $\vee$ Dialetti(Parola, Dialetto, Regione) $\wedge$ Parola = "cavallo" $\wedge$ Da = "it" $\wedge$ A = "en"

LAV

- SinonimiContrari(Parola, Sinonimo, Contrario):- SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione,Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo)
- Cercarime(Parola, Rima):- SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione,Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo)
- Contesti(Parola, Contesto):- SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione,Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo)
- Dialetti(Parola, Dialetto, Regione):- SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione,Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo)
- GoogleTranslator(Parola, Da, A, Traduzione):- SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione,Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo)
- WordNet(Parola, Overview, Synonims):- SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione,Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo)
- Wikipedia(Parola, Descrizione):- SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione,Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo)
- Modi(Parola, Modo):- SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione,Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo)

LAV - 2

1. q(Sinonimo, Rima) :- SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione,Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo) $\wedge$ Parola = "gatto" $\wedge$ Da = "it" $\wedge$ A = "en"
2. q(Sinonimo, Contrario, Rima, Dialetto, Regione, Modo, Wikipedia) :- SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione,Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo) $\wedge$ Parola = "bello" $\wedge$ Da = "it" $\wedge$ A = "en"
3. q(Wikipedia, Traduzione):- SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione,Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo) $\wedge$Parola = "dog" $\wedge$ Da = "en" $\wedge$ A = "it"
4. q(Dialetto, Rima):- SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione,Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo) $\wedge$ Parola = "cavallo" $\wedge$ Da = "it" $\wedge$ A = "en"

Bucket example

- q(Sinonimo, Rima) :- SchemaGlobale(Parola, Sinonimo, Contrario, Rima, Dialetto, Regione, Traduzione,Da, A, Overview, Synonyms, Descrizione, Wikipedia, Modo) $\wedge$ Parola = "gatto" $\wedge$ Da = "it" $\wedge$ A = "en"
  Bucket=SchemaGlobale{V1('gatto', Sinonimo, Contrario), V2('gatto',Rima), V3('gatto', Contesto), V4('gatto', Dialetto, Regione), V5('gatto', Da, A, Traduzione), V6('gatto', Overview, Synonims), V7('gatto', Descrizione), V8('gatto', Modo)}

Outline

Mapping con il codice e tecnologie utilizzate

Per realizzare l'applicazione web descritte abbiamo utilizzato:

- **php** come linguaggio di programmazione
- database **sql** per memorizzare le informazioni informazioni statiche
- **xml**, **xpath** ed **xslt** per estrarre i dati in real time dal database e mostrarli a video

End

Grazie per l'attenzione