

Sistemi Collaborativi e di Protezione (SCP)

Corso di Laurea in Ingegneria

Data Mining, Natural Language Processing, OSIM

Eng. Ph.D. Michela Paolucci

Eng. Ph.D. Gianni Pantaleo

Department of Information Engineering

University of Florence

Via S. Marta 3, 50139, Firenze, Italy

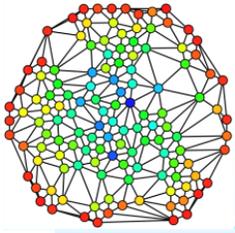
tel: +39-055-4796523, fax: +39-055-4796363

Lab: DISIT, Sistemi Distribuiti e Tecnologie Internet

`michela.paolucci@unifi.it`

`gianni.pantaleo@unifi.it`

<http://www.disit.dinfo.unifi.it/>



Outline

● Sistemi di Web Crawling

- ♣ Introduzione
- ♣ Strategie di Crawling
- ♣ Robot Exclusion Protocol
- ♣ Concorrenza

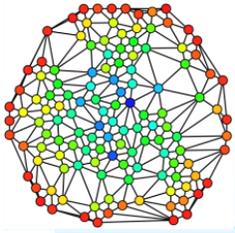
● NLP: Natural Language Processing

- ♣ Introduzione
- ♣ Stato dell'Arte
- ♣ Fasi dell'Elaborazione in Linguaggio Naturale
- ♣ NLP Tools: GATE

● Progetto OSIM

- ♣ Scenario
- ♣ Architettura del Sistema
- ♣ Funzionalità del Sistema
- ♣ Modello Ontologico
- ♣ Interrogare la Conoscenza – Query Wizard

● Sistemi di Question-Answering in NLP: Aqualog



Outline

- **Sistemi di Web Crawling**

- ♣ Introduzione
- ♣ Strategie di Crawling
- ♣ Robot Exclusion Protocol
- ♣ Concorrenza

- **NLP: Natural Language Processing**

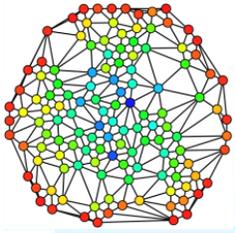
- ♣ Introduzione
- ♣ Stato dell'Arte
- ♣ Fasi dell'Elaborazione in Linguaggio Naturale
- ♣ NLP Tools: GATE

- **Progetto OSIM**

- ♣ Scenario
- ♣ Architettura del Sistema
- ♣ Funzionalità del Sistema
- ♣ Modello Ontologico
- ♣ Interrogare la Conoscenza – Query Wizard

- **Sistemi di Question-Answering in NLP: Aqualog**





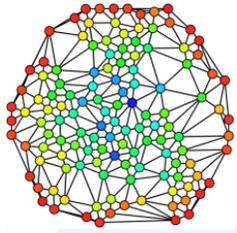
Sistemi di Web Crawling

Introduzione

Strategie di Crawling

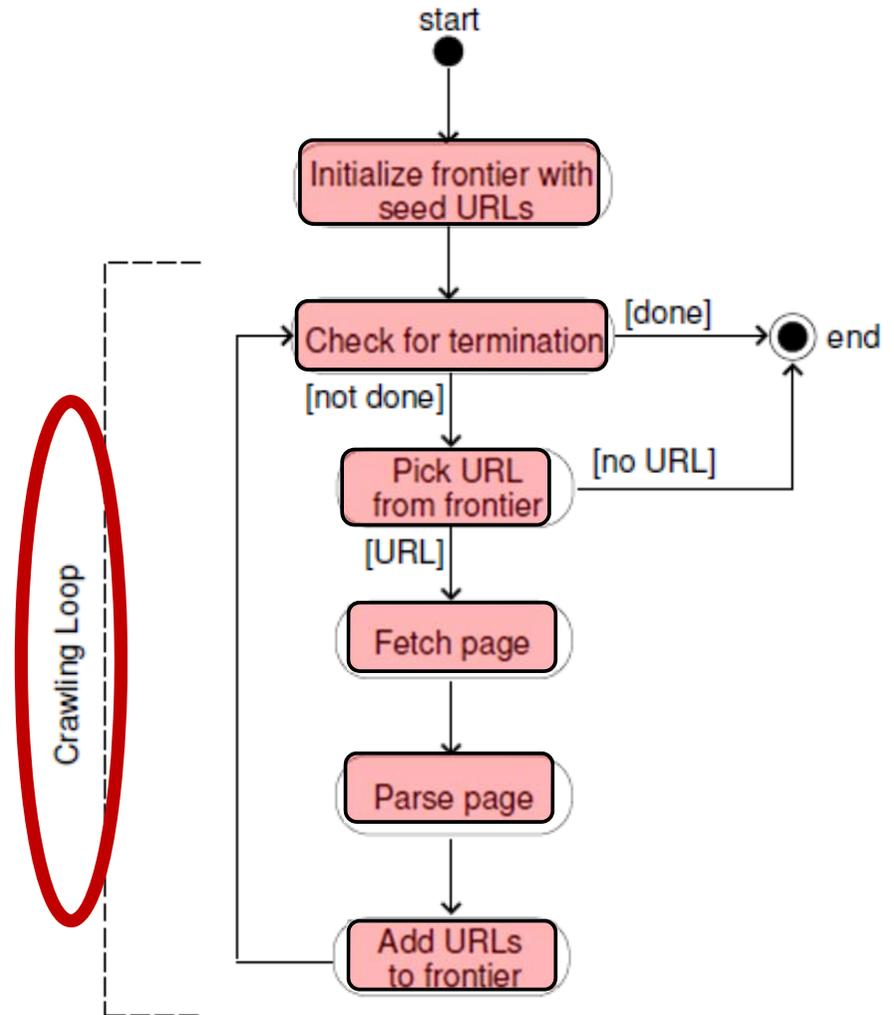
Robot Exclusion Protocol

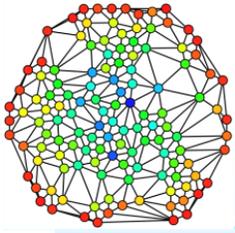
Concorrenza



Web Crawler - Generalità

- Un web crawler è un programma per estrarre automaticamente l'informazione contenuta nel web
- E' utilizzato per creare una copia locale di tutte le pagine visitate per una loro successiva elaborazione (estrazione di informazioni, indicizzazione, ecc..)
- Il web è visto come un grafo: i nodi sono le pagine web e gli archi sono gli hyperlinks



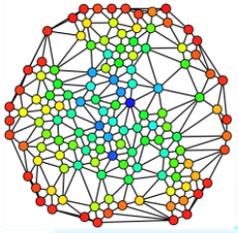


Web Crawler - Frontiera

- contiene gli URLs non ancora visitati.
- può essere implementata in memoria o su disco
- può riempirsi molto velocemente rispetto alle pagine web via via crawlate
- con una media (stimata) di n links per pagina, la velocità di popolazione della frontiera è lineare, ma cresce di circa n volte rispetto al numero delle pagine già crawlate
- occorre limitare la sua dimensione con un valore massimo



Occorre stabilire un meccanismo per decidere quale URL ignorare in caso di limite raggiunto



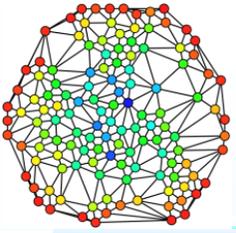
Sistemi di Web Crawling

Introduzione

Strategie di Crawling

Robot Exclusion Protocol

Concorrenza

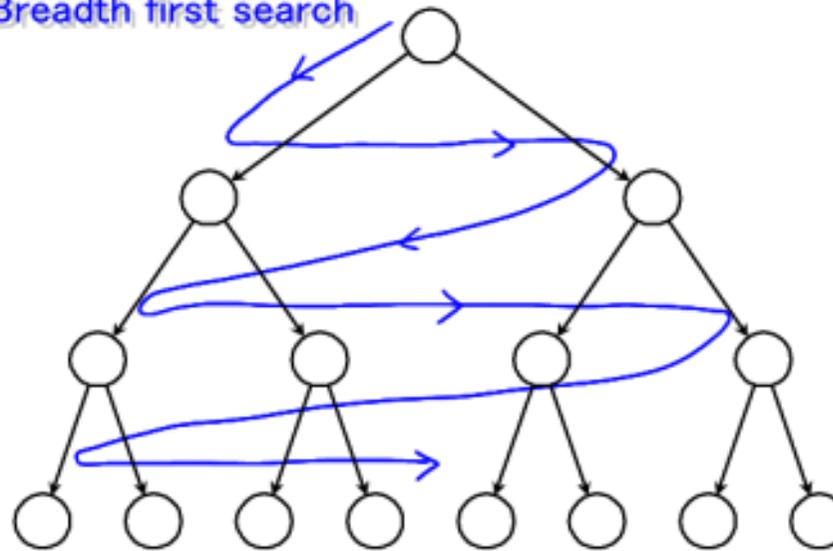


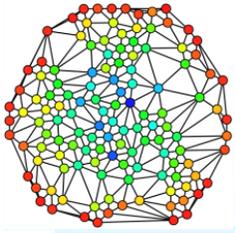
Strategie di crawling (1)

- **Breadth First Search**
- la frontiera è implementata come una coda
- Struttura dati lineare con politica **FIFO**

Attenzione a non inserire URL doppi

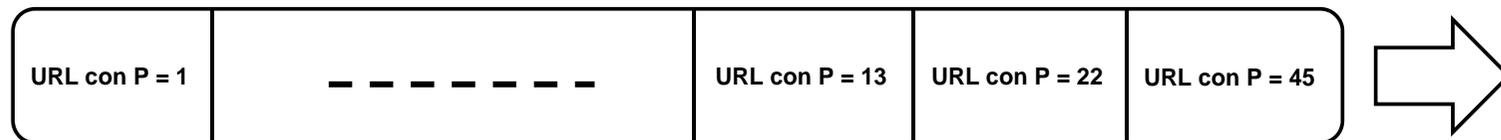
Breadth first search

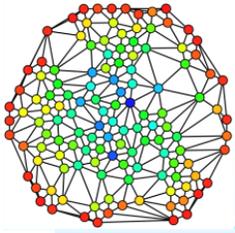




Strategie di crawling (3)

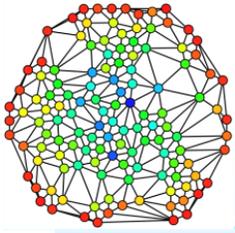
- **Priority Search**
- Struttura dati lineare: l'URL con maggiore priorità viene scelto
- Array dinamico ordinato in base allo score attribuito ad ogni URL
- Gli url sono aggiunti nella frontiera in maniera tale da preservarne l'ordine in base allo score





Considerazioni implementative

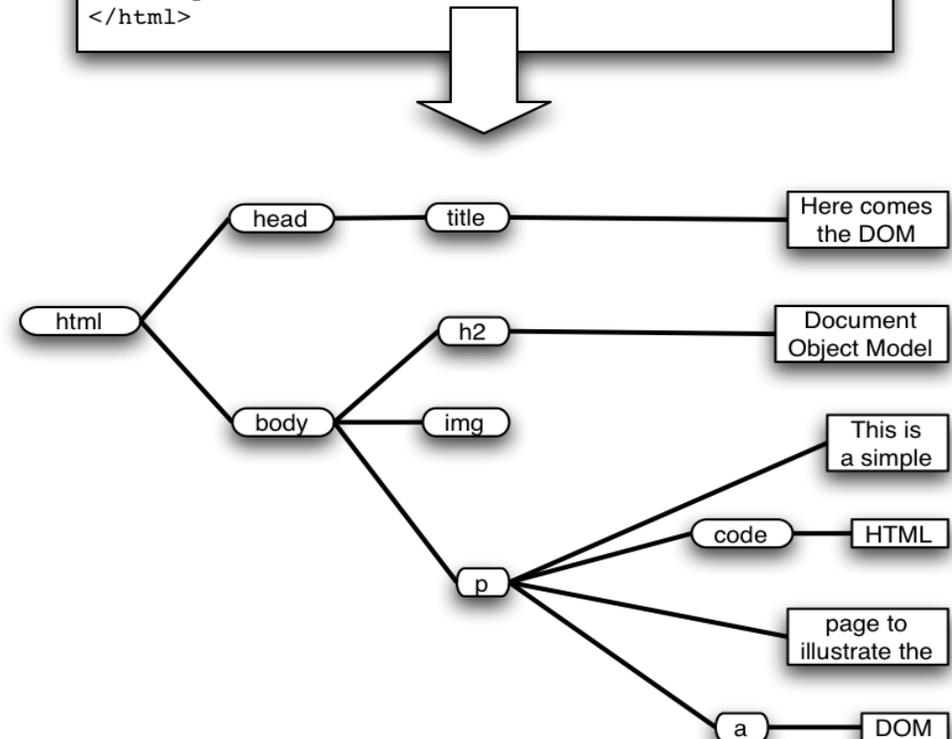
- Evitare il fetch della stessa pagina:
 - ♣ Tenere in memoria una hash delle pagine visitate
- La dimensione della frontiera cresce velocemente
 - ♣ Può essere necessaria una politica di priorità sugli URLs
- Fetcher deve essere robusto
 - ♣ Evitare crash in caso di fallimenti nei downloads
 - ♣ Prevedere meccanismi di timeout
- Determinazione dei files non desiderati
 - ♣ Esaminare estensioni
 - ♣ Content-Type (MIME) headers

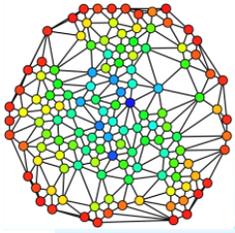


Considerazioni implementative: Parsing

- I documenti HTML hanno una struttura ad albero - DOM (Document Object Model)
- Spesso i documenti HTML non rispettano gli standard di sintassi
- Occorre trattare le entità HTML e gli unicode nei testi
- Vi sono molti formati diversi di files:
 - ♣ Flash, SVG, RSS, AJAX...

```
<html>
  <head>
    <title>Here comes the DOM</title>
  </head>
  <body>
    <h2>Document Object Model</h2>
    
    <p>
      This is a simple
      <code>HTML</code>
      page to illustrate the
      <a href="http://www.w3.org/DOM/">DOM</a>
    </p>
  </body>
</html>
```





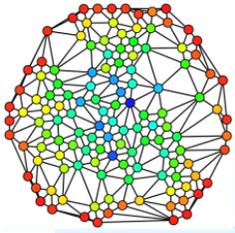
Sistemi di Web Crawling

Introduzione

Strategie di Crawling

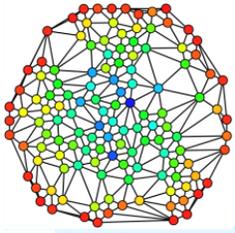
Robot Exclusion Protocol

Concorrenza



Robot Exclusion Protocol (1)

- ♣ Un server può specificare a quale parte dell'albero dei propri documenti può essere accessibile a un crawler (robot) esterno
- ♣ Questa informazione è nel file 'robots.txt' posto nell' HTTP root directory
- ♣ Un crawler dovrebbe sempre verificare la presenza di questo file prima di inviare richieste al server



Robot Exclusion Protocol (2)

www.apple.com/robots.txt

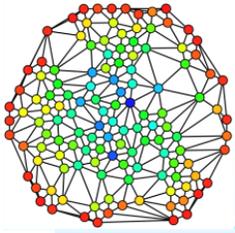
```
# robots.txt for http://www.apple.com/
```

```
User-agent: *
```

```
Disallow:
```

Tutti i crawlers ...

...possono analizzare
qualsiasi pagina!



Robot Exclusion Protocol (3)

www.microsoft.com/robots.txt

```
# Robots.txt file for http://www.microsoft.com
```

```
User-agent: *
```

```
Disallow: /canada/Library/mnp/2.aspx/
```

```
Disallow: /communities/bin.aspx
```

```
Disallow: /communities/eventdetails.aspx
```

```
Disallow: /communities/blogs/PortalResults.aspx
```

```
Disallow: /communities/rss.aspx
```

```
Disallow: /downloads/Browse.aspx
```

```
Disallow: /downloads/info.aspx
```

```
Disallow: /france/formation/centres/planning.asp
```

```
Disallow: /france/mnp_utility.aspx
```

```
Disallow: /germany/library/images/mnp/
```

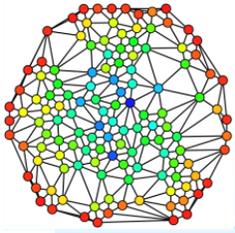
```
Disallow: /germany/mnp_utility.aspx
```

```
Disallow: /info/customerror.htm
```

```
#etc...
```

Tutti i crawlers ...

... non hanno
accesso a questi
paths



Robot Exclusion Protocol (4)

Robots.txt for <http://www.springer.com> (fragment)

User-agent: Googlebot
Disallow: /ch/*
Disallow: /uk/*
Disallow: /italy/*
Disallow: /france/*

Google crawler può analizzare tutte le pagine eccetto queste

User-agent: slurp
Disallow:
Crawl-delay: 2

Yahoo e MSN/Windows Live possono analizzare tutte le pagine ma il processo deve essere "lento"

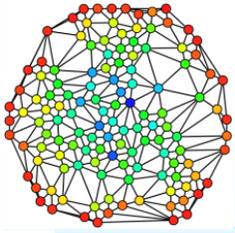
User-agent: MSNBot
Disallow:
Crawl-delay: 2

User-agent: scooter
Disallow:

AltaVista non ha limiti

all others
User-agent: *
Disallow: /

A tutti gli altri crawlers non è permesso niente



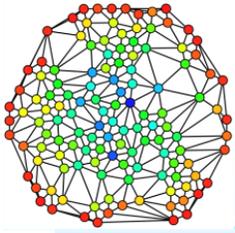
Sistemi di Web Crawling

Introduzione

Strategie di Crawling

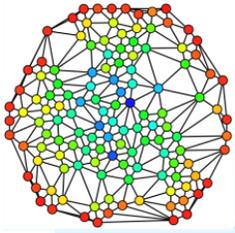
Robot Exclusion Protocol

Concorrenza



Concorrenza (1)

- Le operazioni effettuate dal crawler possono richiedere diverso tempo:
 - ♣ Risoluzione indirizzi IP tramite DNS
 - ♣ Connessioni al server e invio risposte
 - ♣ Ricezione pagina di risposta
- Soluzione: Overlap dei suddetti tempi eseguendo il fetch di più pagine in maniera concorrente

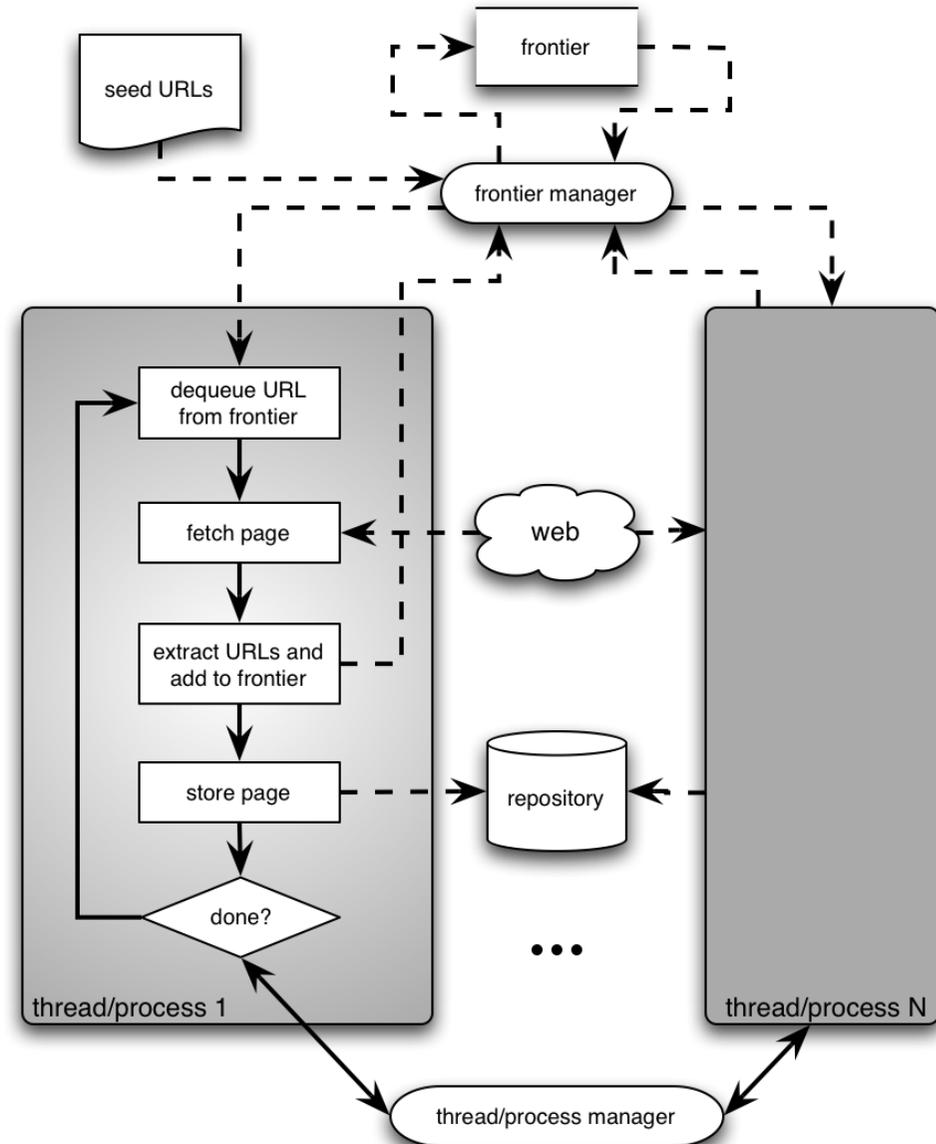


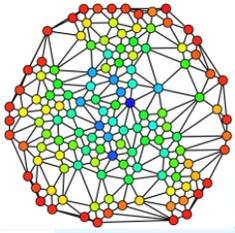
Concorrenza (2)

Ogni thread lavora come un crawler sequenziale e condivide le strutture dati: frontiera e repository

Le strutture dati condivise devono essere sincronizzate (concorrenza in scrittura)

Fattore di speedup 5-10





Outline

● Sistemi di Web Crawling

- ♣ Introduzione
- ♣ Strategie di Crawling
- ♣ Robot Exclusion Protocol
- ♣ Concorrenza

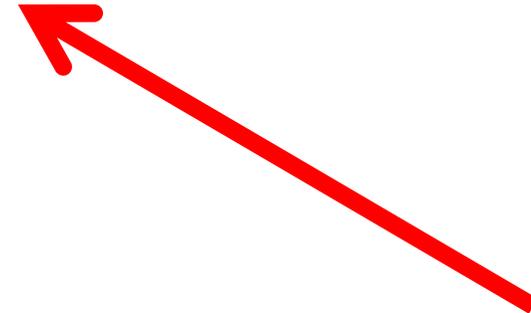
● NLP: Natural Language Processing

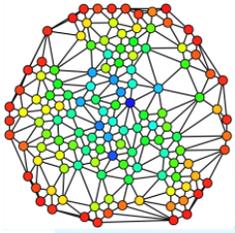
- ♣ Introduzione
- ♣ Stato dell'Arte
- ♣ Fasi dell'Elaborazione in Linguaggio Naturale
- ♣ NLP Tools: GATE

● Progetto OSIM

- ♣ Scenario
- ♣ Architettura del Sistema
- ♣ Funzionalità del Sistema
- ♣ Modello Ontologico
- ♣ Interrogare la Conoscenza – Query Wizard

● Sistemi di Question-Answering in NLP: Aqualog





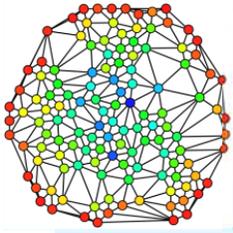
Sistemi di Web Crawling

Introduzione

Stato dell'Arte

Fasi dell'Elaborazione in Linguaggio Naturale

NLP Tools: GATE



Natural Language Processing NLP (1)

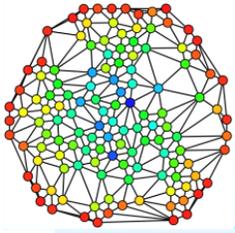
● Scenario / Requisiti

- ❖ Dotare l'IA delle abilità linguistica proprie dell'essere umano
- ❖ Comprensione e generazione del testo
- ❖ Contesto multi-language: differenti regole e strutture a seconda della lingua

● Applicazioni

- ❖ Generalizzazione delle query nei motori di ricerca
 - *"Chi si occupa di sistemi distribuiti nell'Università di Firenze ?"*
- ❖ Supporto automatizzato per Help-Desk
- ❖ Tutoring assistito (e-tutoring, e-teaching...)
- ❖ Summarization: creare compendi da una collezione eterogenea di documenti
- ❖ Machine translation: tradurre testi in lingue diverse



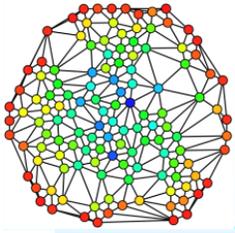


Ambiguità dei linguaggi naturali (1)

● Scenario / Requisiti

- ♣ I linguaggi sono purtroppo ambigui.
- ♣ Le ambiguità si possono avere a 4 livelli:
 - ✓ Ambiguità lessicale: «*attacco*» (verbo, sostantivo)
 - ✓ Ambiguità strutturale: «*leri ho visto l'uomo col telescopio*»
«*Una vecchia legge la regola*»
 - ✓ Ambiguità semantica: «*acuto*» (persona intelligente, tipo di suono)
 - ✓ Ambiguità pragmatica: «*se Buffon non gioca contro la Spagna, l'Italia perderà*»
 - interpretazione emotiva: *l'assenza di Buffon è psicologicamente fondamentale per i tifosi*
 - Interpretazione referenziale: *l'Italia senza Buffon è più debole*

Ciò rende il processo di elaborazione del linguaggio naturale molto complicato



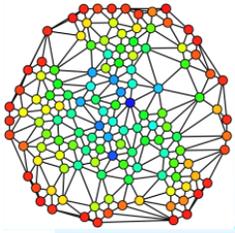
Sistemi di Web Crawling

Introduzione

Stato dell'Arte

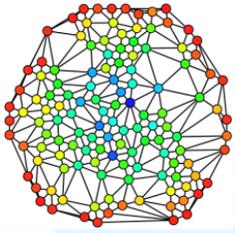
Fasi dell'Elaborazione in Linguaggio Naturale

NLP Tools: GATE



NLP: Stato dell'Arte (1)

- **Primi teorie e sviluppi applicativi: *fine anni '60 – inizio anni '70***
 - ♣ I primi studi erano basati su un ***approccio linguistico***
 - ♣ L'aumento del numero e dimensione dei corpora testuali, insieme all'elevata specializzazione del linguaggio in relazione al dominio specifico di conoscenza, ha reso necessario approcci differenti: ***machine learning, metodi statistici, pattern matching ecc...***

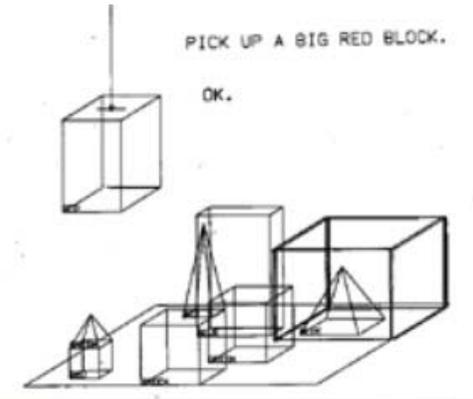


NLP: Stato dell'Arte (2)

- **SHRDLU** [Winograd – 1971]

- ♣ Interfaccia per automa preposto al semplice movimento di blocchi 3D
- ♣ Dominio limitato, query semplici

```
Person: PICK UP A BIG RED BLOCK.  
Computer: OK. (does it)  
Person: GRASP THE PYRAMID.  
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.  
Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.  
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.  
Computer: OK. (does it)  
Person: WHAT DOES THE BOX CONTAIN?  
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK. [...]
```



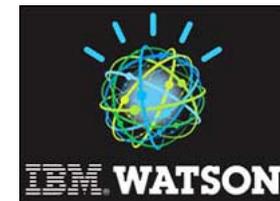
- **IOS Siri** [IOS Siri – 2010]

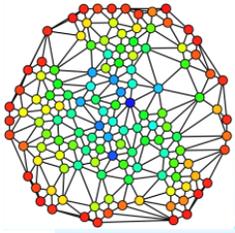
- ♣ Intelligent personal assistant – Knowledge navigator – User recommendation system.



- **IBM Watson** [IBM Watson – 2012]

- ♣ Nato come sistema Question & Answer per il quiz televisivo americano *Jeopardy!*, successivamente è stato sviluppato come *cognitive system* completo con meccanismi di auto-apprendimento.

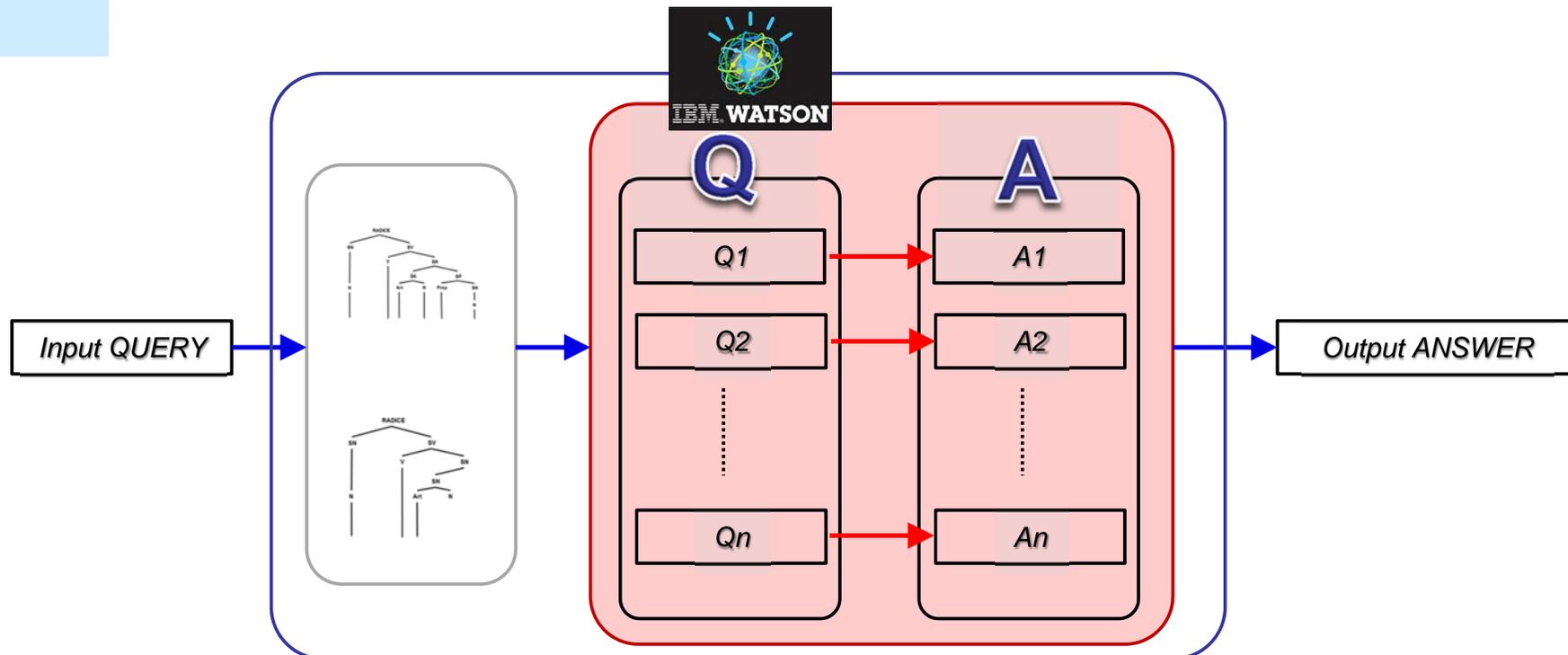
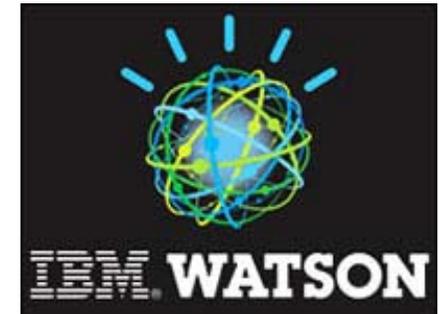


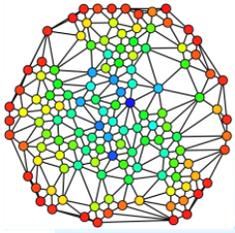


NLP: Stato dell'Arte (3)

● IBM Watson

- ♣ Database built-in in cui sono già presenti le domande relazionate con le rispettive risposte corrette.
- ♣ Il sistema si limita ad analizzare la query in input, attraverso l'elaborazione del grafo sintattico, cercando un match con una delle domande presenti nel set.





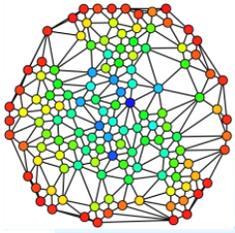
Sistemi di Web Crawling

Introduzione

Stato dell'Arte

Fasi dell'Elaborazione in Linguaggio Naturale

NLP Tools: GATE



Fasi dell'elaborazione in Linguaggio Naturale (1)

Morphological Analysis: le parole vengono analizzate (distinzione dei morfemi che le compongono) ed i simboli (punteggiature) vengono separati dalle parole .

Syntactic Analysis: Le sequenze di parole sono trasformate in strutture che mostrano come le parole sono in relazione l'una con l'altra.

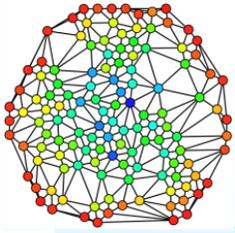
Semantic Analysis: Viene assegnato un significato alle strutture sintattiche trovate.

Discourse integration: il significato di una frase spesso dipende dalla frase che la precede e può influenzare quello della frase che la segue.

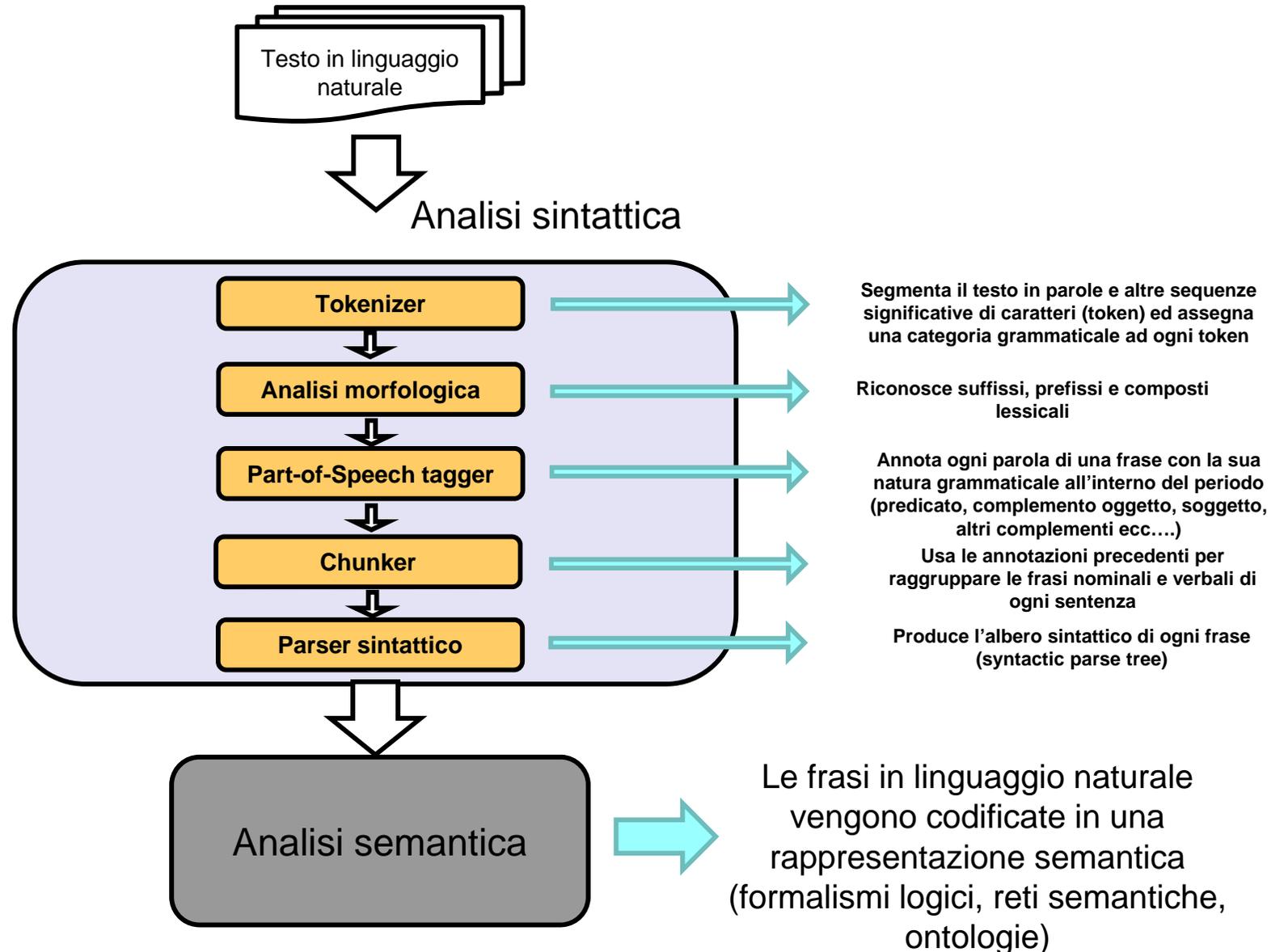
Pragmatic Analysis: la frase è reinterpretata per determinare il significato specifico della frase stessa.

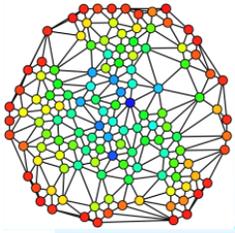
“la porta è aperta” necessita di conoscere quale è stata l'intenzione dell'interlocutore:

- Si è creata una corrente d'aria...
- Invito ad entrare liberamente...
- Richiesta affinché qualcuno chiuda la porta...



Fasi dell'elaborazione in Linguaggio Naturale (2)



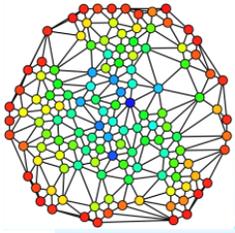


Un Esempio

Paolo Nesi è il docente di «Sistemi distribuiti»

Tokenizer + analizzatore morfologico

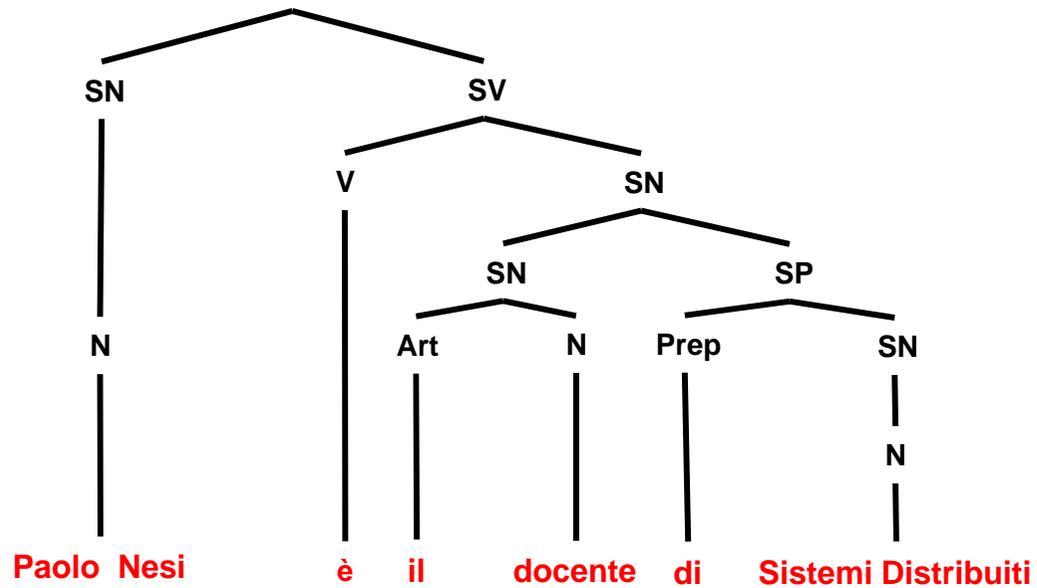
(Paolo Nesi):	<i>nome proprio</i>
(è):	<i>verbo essere terza persona singolare indicativo presente</i>
(il):	<i>articolo determinativo maschile singolare</i>
(docente):	<i>sostantivo maschile singolare</i>
(di):	<i>preposizione semplice</i>
(Sistemi distribuiti):	<i>nome</i>



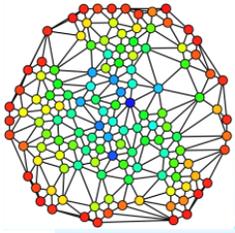
(Paolo Nesi): *nome proprio*
(è): *verbo essere terza persona singolare indicativo presente*
(il): *articolo determinativo maschile singolare*
(docente): *sostantivo maschile singolare*
(di): *preposizione semplice*
(Sistemi distribuiti): *nome*

Part-of-Speech + chunker + parser
sintattico

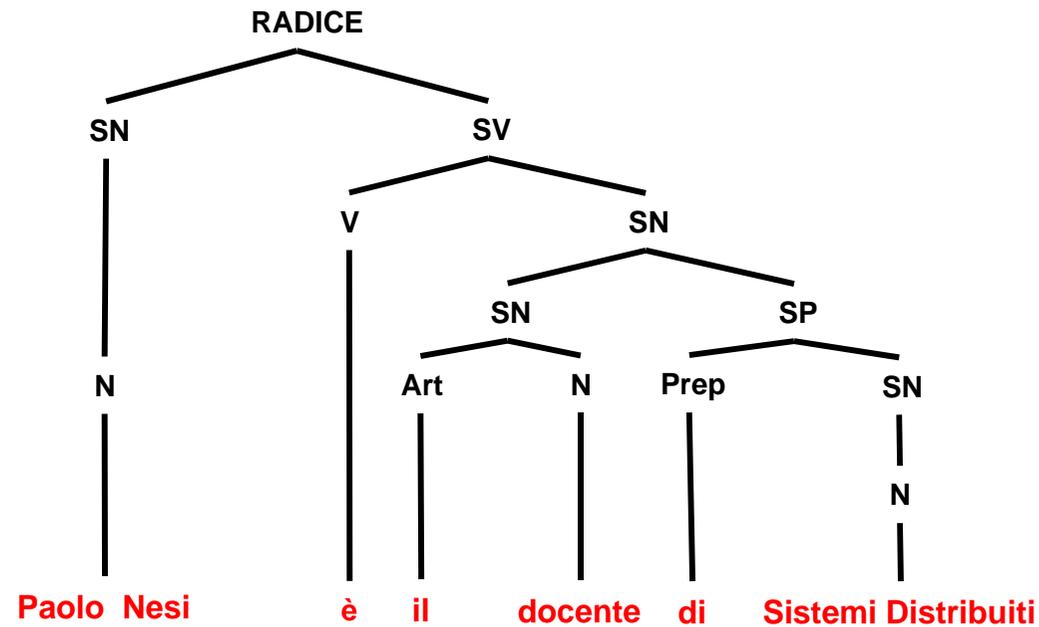
RADICE



SN: sentenza nominale
SV: sentenze verbale
SP: sentenza preposizionale
N: nome
V: verbo
Art: articolo
Prep: preposizione



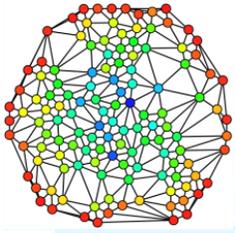
Un Esempio



Analisi semantica

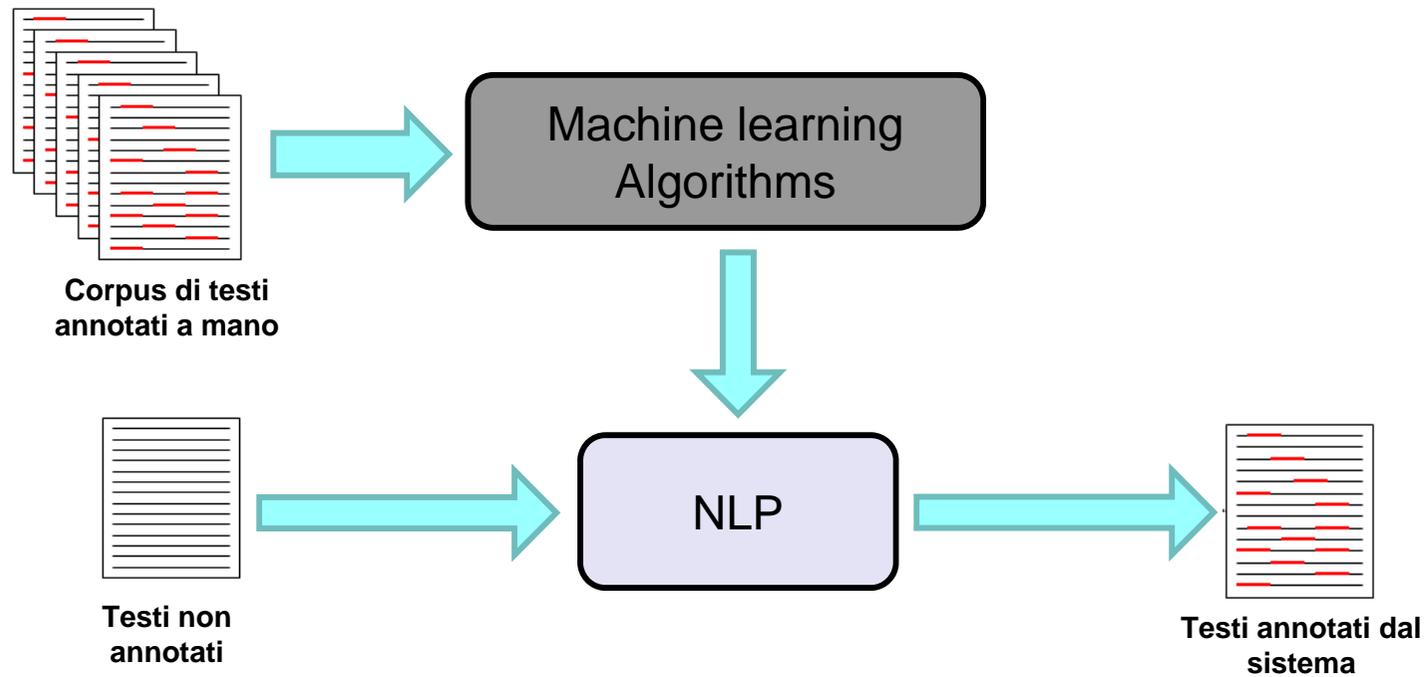


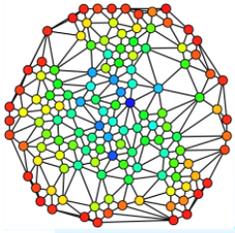
E' (Paolo Nesi($\exists y$ (**DOCENTE**(y Sistemi Distribuiti))))



Machine Learning

I sistemi di NLP usano principalmente algoritmi di machine learning addestrati su grandi corpus di testi annotati a mano





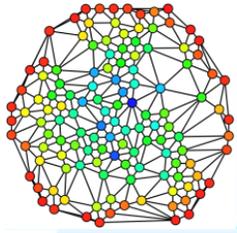
Sistemi di Web Crawling

Introduzione

Stato dell'Arte

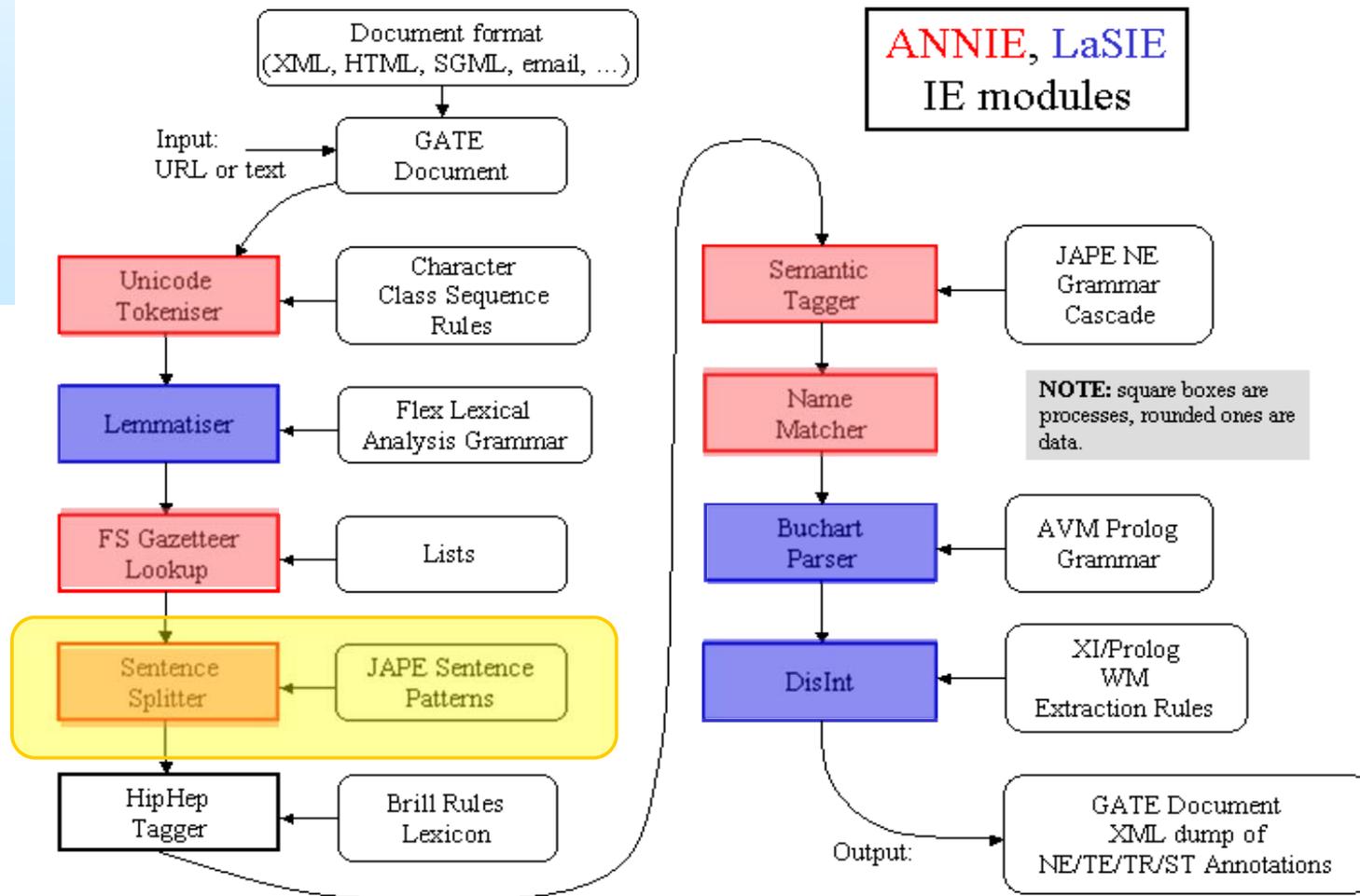
Fasi dell'Elaborazione in Linguaggio Naturale

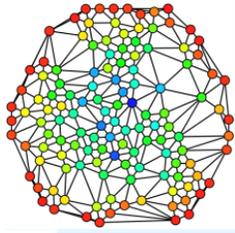
NLP Tools: GATE



NLP Tools: GATE (1)

- **GATE – General Architecture for Text Engineering**
 - ♣ Supporta documenti plain text, XML, RTF, HTML, SGML





NLP Tools: GATE (2)

Principali Moduli & Plug-in di GATE

- **ANNIE (A Nearly New IE System)**

- ✦ Utilizza regole per il **Part of Speech (POS) Tagging**. Tali regole sono predefinite (tag html) o customizzate dall'utente

- ✦ Una regola ha un left hand side (LHS) e un right hand side (RHS)

- **LHS**: espressione regolare da riscontrare nel testo in input

- **RHS**: descrive le annotazioni che devono essere aggiunte all'*AnnotationSet*

- Utilizzano regole predefinite (tag html) o customizzate dall'utente

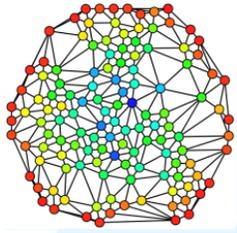
- ✦ Sintassi:

`{LHS} > {Annotation type}; {attribute1}={value1};...;{attribute n}={value n}`

- ✦ Es.:

`"UPPERCASE_LETTER" "LOWERCASE_LETTER"* > Token; orth=upperInitial; kind=word.`

- ✦ Tipi di Token previsti: Word, Number, Symbol, Punctuation, Space Token



NLP Tools: GATE (2)

Principali Moduli & Plug-in di GATE

- **JAPE (Java Annotations Pattern Engine)**

- ♣ Permette di ricercare espressioni regolari nel testo in input
- ♣ Regole composte da **LHS** e **RHS**

- ♣ Es. di sintassi:

```
Phase: findVerb
```

```
Input: Token
```

```
Options: control = appelt
```

← Headers della regola

```
Rule: FindITVerb
```

```
Priority: 20
```

← Nome della regola e priorità

```
(  
  ({Token.category == "VER:pres"}) |  
  ({Token.category == "VER:pper"}) |  
  ({Token.category == "VER:infi"})  
)
```

← LHS

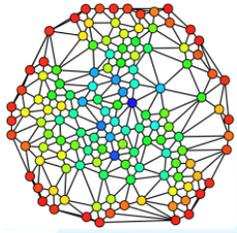
```
:competence
```

← Label

```
-->
```

← RHS

```
:competence.Verb= {kind = "verb" rule = "FindITVerb" lang = "it" lemma =  
:competence.Token.lemma}
```



NLP Tools: GATE (3)

Principali Moduli & Plug-in di GATE

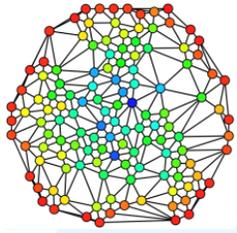
The screenshot shows the website of the University of Florence (unifi.it). The navigation menu on the left includes sections like 'Conoscere l'università', 'STUDENTI', 'Offerta formativa', 'Immatricolazioni', 'Informazioni generali', 'orario di ricevimento', 'Insegnamenti', 'Biografia', 'Curriculum', 'Interessi', and 'Pubblicazioni'. The main content area displays the profile of Paolo Nesi, including a list of publications with a legend for document types like 'Contributo su rivista', 'Articolo su libro', etc.

Legenda

- Contributo su rivista
- Articolo su libro
- Libro
- Contributo in atti di convegno (proceeding)
- Brevetto
- Curatela
- Altro
- Tesi di Dottorato

Pubblicazioni

- Pierfrancesco Bellini, Ivan Bruno, Paolo Nesi, Daniele Cenni (2012). Micro Grids for Scalable Media Computing and Intelligence in Distributed Scenarios. IEEE MULTIMEDIA, vol. 19, pp. 69-79, ISSN:1070-986X
- Paolo Nesi, Raffaella Santucci (a cura di) (2012). ECLAP 2012 Conference on Information Technologies for Performing Arts, Media Access and Entertainment. FUP - FIRENZE UNIVERSITY PRESS.
- D. Cenni, P. Nesi, M. Paolucci, (2012). Functionalities and Flow Analyses of Knowledge Oriented Web Portals. In: Proceeding of the 18th International Conference on Distributed Multimedia Systems, DMS 2012, Eden Roc Renaissance Miami Beach, USA, SEKE software engineering and knowledge engineering institute, 9-11 August, 2012., pp. 1-6.
- Pierfrancesco Bellini, D. Cenni, P. Nesi, (2012). Indexing and Searching Cross Media Content in a Social Network. In: Proceeding of the ECLAP 2012 International Conference on Information Technologies for Performing Arts, Media Access and Entertainment", Florence University Press, May 2012,, pp. 52-57, ISBN:9788966551256
- Pierfrancesco Bellini, Sergio Boncinelli, Francesco Grossi, Marco Mangini, Luca Malandrino, Paolo Nesi, (2012). Mobile Emergency: shortening communications in hospital. In: Media Tablets & Apps for Medicine, Health, and Home-Care, R&D Workshops & Demos, 1st-2nd March 2012, Politecnico di Milano, Milano., pp. 1-6.
- Pierfrancesco Bellini, Ivan Bruno, Daniele Cenni, Alice Fuzier, Nicola Mitolo, Paolo Nesi, Michela Paolucci (2012). Mobile Medicine: Smart Applications and Content for Medical Training and Support. In: Media Tablets & Apps for Medicine, Health, and Home-Care, R&D Workshops & Demos, 1st-2nd March 2012, Politecnico di Milano, Milano., pp. 1-6.
- P. Bellini, D. Cenni, P. Nesi, (2012). On the Effectiveness and Optimization of Information Retrieval for Cross Media Content. In: Proceeding of the KDIR 2012 is part of IC3K 2012, International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 4-7 October 2012, Barcelona, Spain., pp. 1-8.
- F. Argenti, P. Nesi, G. Pantaleo (2011). Automatic Transcription of Polyphonic Music Based on the Constant-Q Bispectral Analysis. IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, vol. 19 (#6), pp. 1610-1630, ISSN:1558-7916 URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5640655
- F. Argenti, P. Nesi, G. Pantaleo (2011). Automatic music transcription: from monophonic to polyphonic. In: J. Solis, K. Ng. Musical Robots and Interactive Multimodal Systems, pp. 27-46, Heidelberg: Springer. DOI: 10.1007/978-3-642-22291-7
- Pierfrancesco Bellini, Ivan Bruno, Daniele Cenni, Alice Fuzier, Paolo Nesi, Michela Paolucci (2011). Mobile Medicine: semantic computing management for health care applications on desktop and mobile devices. MULTIMEDIA TOOLS AND APPLICATIONS, vol. 53, pp. 1-39, ISSN:1380-7501 URL: <http://www.springerlink.com/content/q8512555u0j00584/fulltext.pdf>

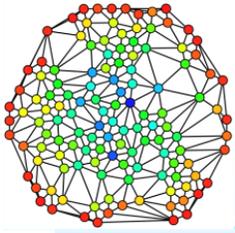


NLP Tools: GATE (4)

Principali Moduli & Plug-in di GATE

The screenshot displays the GATE Developer 5.2.1 interface. The main window shows a text document with several annotations. A list of processing resources is visible on the left, including ANNIE, Language Resources, and Processing Resources. The right side shows a list of annotation sets, with 'PersonType' and 'Location' selected. The bottom part of the window shows a table of annotations.

Type	Set	Start	End	Id	Features
Location		26	33	28775	{locType=city, rule1=Location1, rule2=LocFinal}
Location		73	80	28776	{locType=city, rule1=Location1, rule2=LocFinal}
FirstPerson		91	96	27758	{gender=male, rule=FirstName}
Location		108	115	28778	{locType=city, rule1=Location1, rule2=LocFinal}



Outline

● Sistemi di Web Crawling

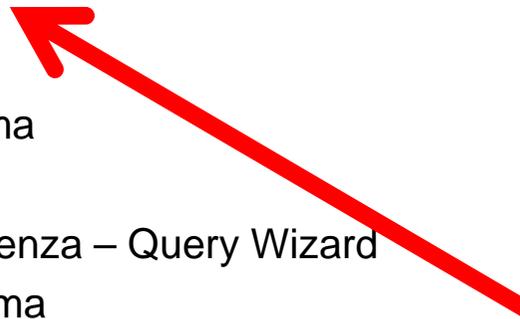
- ♣ Introduzione
- ♣ Strategie di Crawling
- ♣ Robot Exclusion Protocol
- ♣ Concorrenza

● NLP: Natural Language Processing

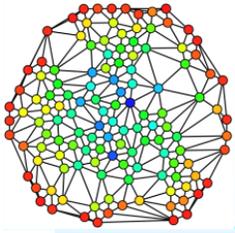
- ♣ Introduzione
- ♣ Stato dell'Arte
- ♣ Fasi dell'Elaborazione in Linguaggio Naturale
- ♣ NLP Tools: GATE

● Progetto OSIM

- ♣ Scenario
- ♣ Architettura del Sistema
- ♣ Modello Ontologico
- ♣ Interrogare la Conoscenza – Query Wizard
- ♣ Funzionalità del Sistema



● Sistemi di Question-Answering in NLP: Aqualog



Progetto OSIM

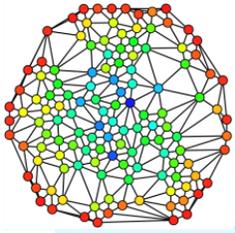
Scenario

Architettura del Sistema

Modello Ontologico

Interrogare la Conoscenza – Query Wizard

Funzionalità del Sistema



OSIM – Scenario (1)

- I motori di ricerca attuali, ad esempio Google, sono keyword-based

Vengono restituiti i documenti che contengono esattamente le parole specificate dall'utente nella ricerca, senza tener conto della semantica di quanto espresso dall'utente stesso

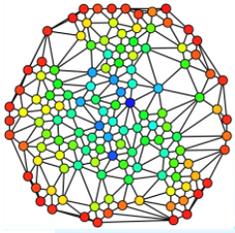


Se per esempio un utente chiede una specifica competenza non è desiderabile avere come risposta tutte le pagine che contengono quella competenza, ma ci si aspetta di conoscere quali Professori hanno quella competenza, in quali corsi universitari può essere acquisita e quali gruppi di ricerca hanno maggior esperienze in quel settore.

La risposte trovate dovrebbero essere ordinate secondo una certa rilevanza



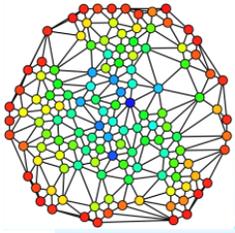
In riferimento ad una competenza specifica, per esempio, un professore può avere più rilevanza, se possiede una competenza più specifica rispetto a quella cercata, di un professore che possiede una competenza più generale rispetto a quella cercata



OSIM – Scenario (2)

- Le ricerche effettuate con Marsilius restituiscono risposte non contestualizzate:
 - ♣ Se si cerca un determinato argomento è desiderabile trovare risposte distinte per pubblicazioni, corsi, persone, dipartimenti, centri, etc.
 - ♣ Le risposte trovate dovrebbero essere ordinate secondo una certa rilevanza
- Si dovrebbero ottenere risposte piu' intelligenti:
 - ♣ Per contesto: Dipartimenti, strutture che, Corsi che, Persone che
 - ♣ Per relazioni di collaborazione: chi ha lavorato con ...
 - ♣ Per relazioni verbali: chi progetta / sviluppa / mantiene software...





Il Progetto O.S.I.M.

- **Open Space Innovative Mind** (<http://openmind.disit.org>) è un progetto per la realizzazione di un portale nel quale industrie, ricercatori, studenti possono effettuare ricerche per individuare, all'interno dell'Ateneo Fiorentino:

- ♣ le competenze possedute dai gruppi di ricerca, dai laboratori e dal personale universitario
- ♣ le competenze offerte da corsi specifici nell'ambito dei vari corsi di laurea.
- ♣ Le relazioni esistenti tra competenze diverse
- ♣ Le pubblicazioni scientifiche
- ♣ Le relazioni di conoscenza tra docenti
- ♣ . . .

Nell'architettura di OSIM sono implementate tutte le tecnologie che sono state presentate in queste lezioni:

- ❖ **Data Mining**
- ❖ **Web Crawling**
- ❖ **NLP**
- ❖ **Produzione di Ontologie e SKOS (OSIM Keywords & Ontology Manager)**
- ❖ **Interrogazione della conoscenza e Reasoning(OSIM Query Wizard)**

OSIM Open Space Innovative Mind BETA Università degli Studi di Firenze

Home Documentation Search Managing Knowledge Browsing People & Publications Contact DISIT

Question Answer Query Wizard

Wizard

Quale Persona o Struttura ha questa competenza ? try

Quale Persona (all) ha questa competenza ? try

Elenca le competenze relazionate con try

Elenca le competenze della Persona Aaron Thomas Fargion try

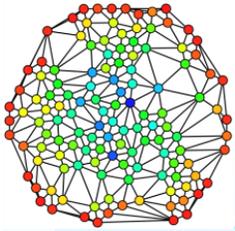
Elenca le Persone del Dipartimento di Agronomia e produzioni erbacee try

Elenca le pubblicazioni della Persona Aaron Thomas Fargion

dal 1955 al 2012 try

Elenca le Persone con area CUN Scienze matematiche e informatiche e settore scientifico disciplinare (SSD) INFORMATICA

dal 1955 al 2012 try



MARSILIUS: Conoscenza relativa a «Sistemi Distribuiti»

Informatica e Applicazioni

Referente: Rosario Pugliese

Obiettivi:

Scopo del Dottorato è la formazione di ricercatori che a verso gli aspetti applicativi. Questo garantirà la possibilità (che debbono fare i conti con frequenti cambiamenti tecnologici) soprattutto ad allargare la base culturale ed all'eccellenza lavorative non collegate alla ricerca.

Il corso punterà a completare la formazione su:

- Algoritmi per sistemi distribuiti: progettazione ed
- Elaborazione delle immagini.
- Metodi formali per la specifica e la verifica di sistemi
- Progettazione di algoritmi di approssimazione, programmazione
- Progettazione ed analisi di sistemi affidabili.
- Sicurezza di sistemi distribuiti.
- Strumenti formali per l'analisi degli algoritmi.
- Trattamento numerico e modellizzazione di sistemi

Visto l'esiguo numero di dottorandi non si prevede l'articolazione degli interessi e che garantisca gli obiettivi formativi generali

www.unifi.it/dist/cmpro-v-p-6.htm

Dottorato in Informatica, ... Sono di interesse i sistemi di software, i sistemi distribuiti, i www.unifi.it/dist/cmpro-v-p-16.htm

Dipartimento di Energetica ... F., Morin F., Miglioramento della Manutenzione ... della catena di www.ibis.unifi.it/cmpro-l-s-4.htm

Primo risultato

DIPARTIMENTO DI ENERGETICA - LABORATORIO IBIS

[home ateneo](#) | [home polo](#) | [home dipartimento](#) | [home laboratorio](#)

Lingua - Language



Menù

- ▶ Home
- ▶ Aree di ricerca
- ▶ Collaborazioni
- ▶ Progetti
- ▶ Prodotti
- ▶ Strumenti
- ▶ Persone
- ▶ Contatti
- ▶ Dove siamo

Utilità

- ▶ Mappa
- ▶ Statistiche
- ▶ Redazione

Aree di ricerca

Nei primi 20 risultati non si trova la pagina del corso.

Al 14-esimo posto si trova la pagina del laboratorio DISIT

Condizioni di lavoro & Condition Based Maintenance

Progettazione di sistemi per l'acquisizione dei dati di campo e per il monitoraggio remoto delle condizioni di impianti industriali, macchinari, flotte, ecc., sviluppo sistemi e modelli per la manutenzione predittiva

Reliability analysis & Expert Systems

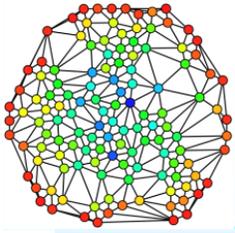
Modellazione e analisi di affidabilità e di disponibilità di sistemi complessi, sviluppo di modelli diagnostici e previsione vita utile residua del bene).

Service Management and Engineering

Analisi del valore, sviluppo service concept, ingegnerizzazione servizi, due diligence tecnologica, progettazione e gestione logistica di supporto.

Di seguito si illustrano le competenze acquisite in riferimento alle tematiche di cui al presente documento. I membri del comitato scientifico del laboratorio.

Secondo risultato



OSIM: Conoscenza relativa a «Sistemi Distribuiti»

OSIM Open Space Innovative Mind **BETA** Università degli Studi di Firenze 

Home Documentation Search Managing Knowledge Browsing People & Publications Contact DISIT

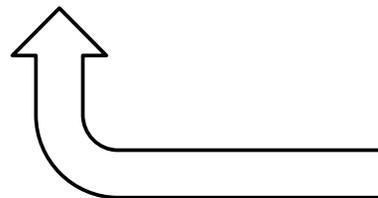
Question Answer [Query Wizard](#)

sistemi distribuiti 

Results Displayed / Found: 1 - 3 / 3 in 21235 millisecc

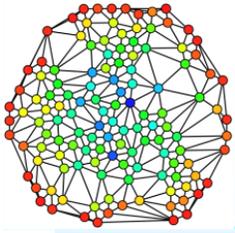
sistemi distribuiti (course)	Freqs: 1	score: 5.14	Paolo Nesi (full professor)	Freqs: 0	score: 4.91
sistemi distribuiti (skill)	Freqs: 0	score: 4.32			

Previous 1 Next



I risultati non sono i documenti che contengono la parola “sistemi distribuiti”

I risultati sono “il corso di Sistemi distribuiti” e la competenza “Sistemi distribuiti”



Progetto OSIM

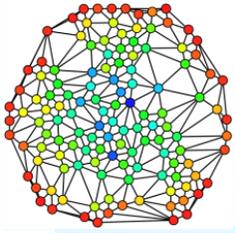
Scenario

Architettura del Sistema

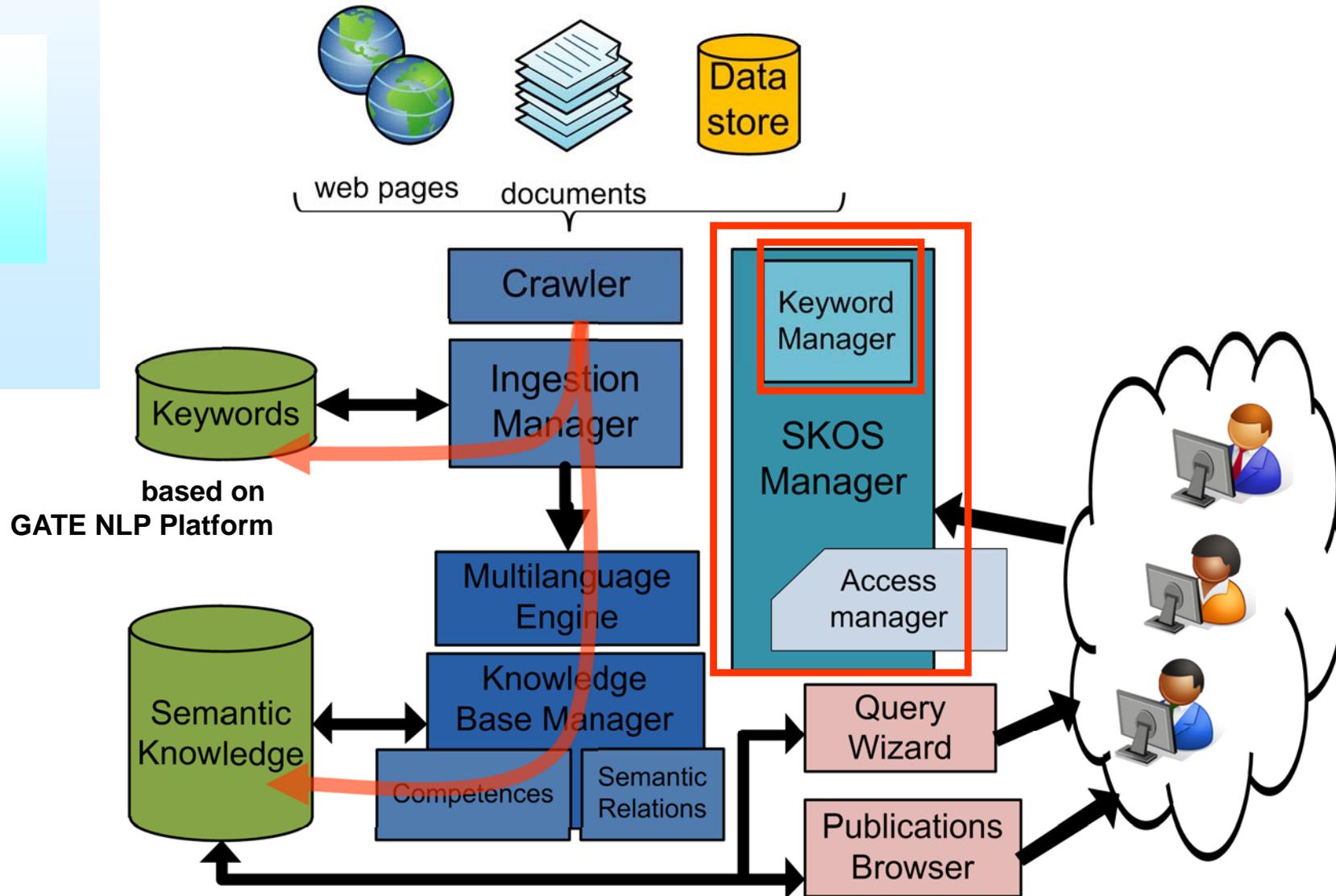
Modello Ontologico

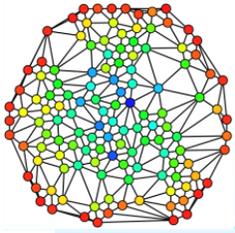
Interrogare la Conoscenza – Query Wizard

Funzionalità del Sistema



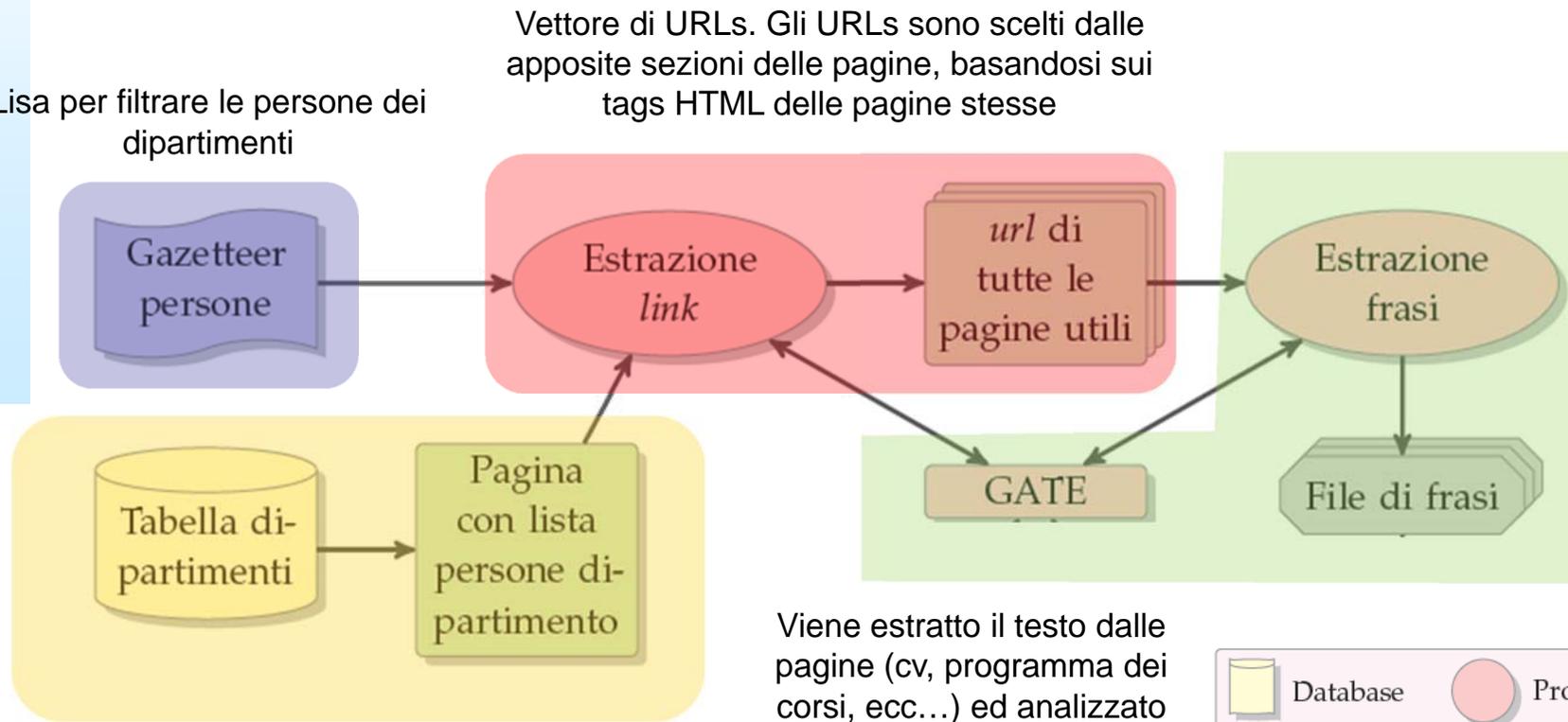
OSIM: Architettura





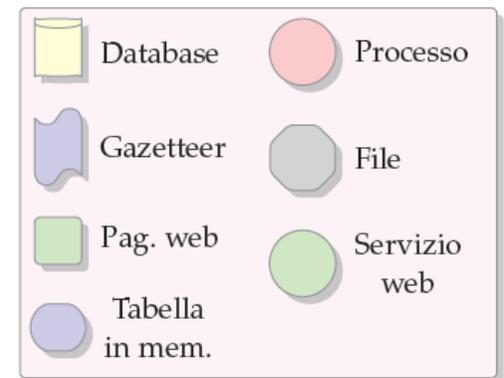
Data Ingestion: *OSIM Web Crawler*

Lisa per filtrare le persone dei dipartimenti



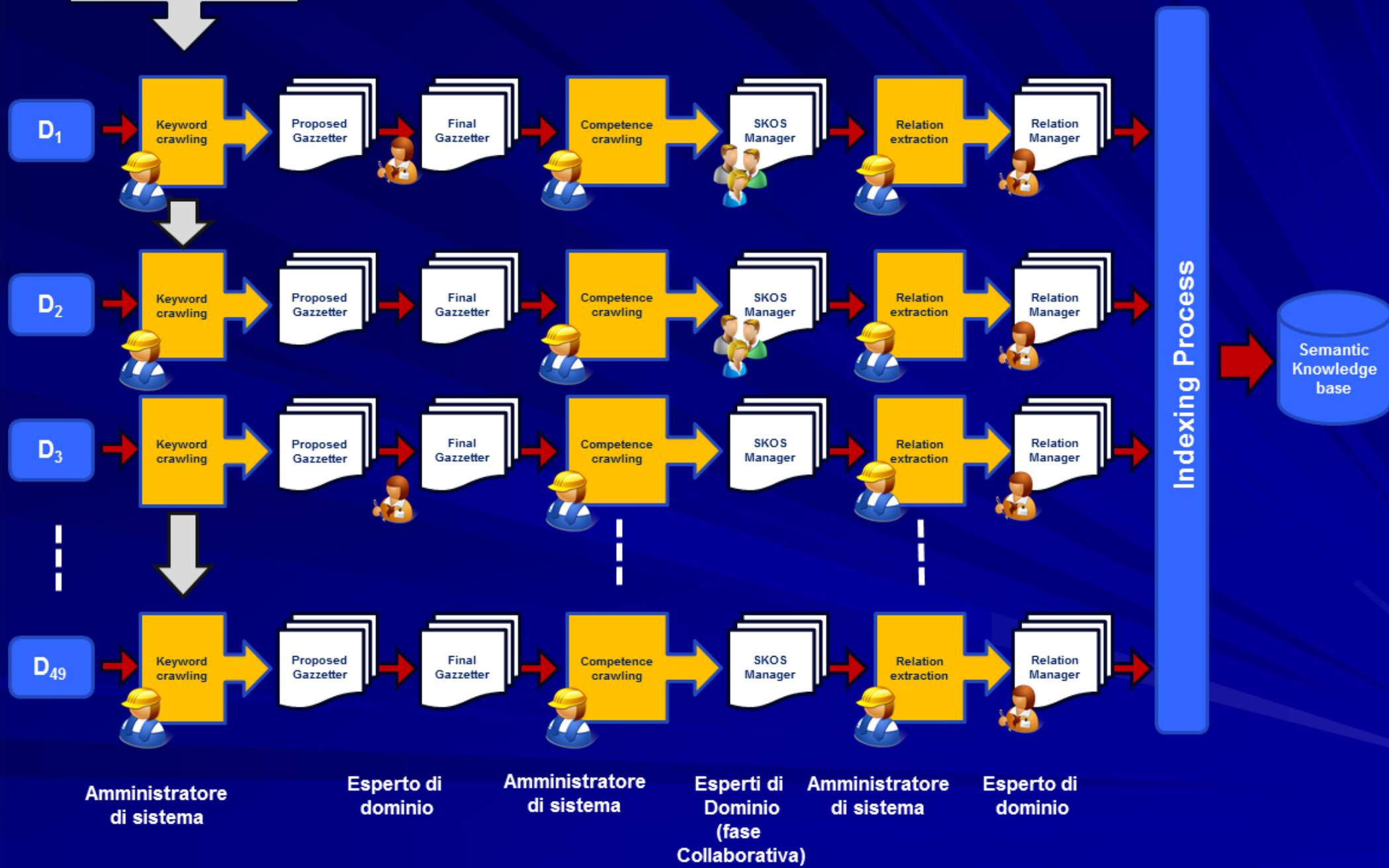
La seed list del crawler sono gli URLs delle pagine dipartimentali del cercachi

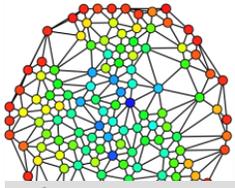
Viene estratto il testo dalle pagine (cv, programma dei corsi, ecc...) ed analizzato da GATE per il **Natural Language Processing**



Knowledge Base Creation Process

Web pages, documents, publications, etc...





Keyword management

Welcome root [Logout](#)

[OSIM Managing Knowledge HOME](#)

dipartimento di matematica per le decisioni - crawler is running

english

ONTOLOGY MANAGER

KEYS SELECTION

RELATIONS MANAGER



all

20/113 (#2258)

id	value	translated values	occurrences	gazetteer	black list	no action	lang	Proposed
9390	algebra	algebra	9	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	en	0
1201	complementary	complementare	9	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	en	0
9139	changes	variazioni	9	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	en	0
9143	horizon	orizzonte	9	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	en	0
7611	decomposition	decomposizione	9	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	en	0
9148	infinite	infiniti	9	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	en	0
7365	laboratory	laboratorio	9	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	en	0
8646	angles	angoli	9	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	en	0
8397	embrechts	embrechts	9	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	en	0
9427	bond	obbligazione	9	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	en	0
9431	fields	campi	9	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	en	0
9689	edition	edizione	9	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	en	0
8673	min	min	9	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	en	0
482	decomposition	scomposizione	9	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	en	0
8426	year-old	anni	9	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	en	0

Trova: [Successivo](#) [Precedente](#) [Evidenzia](#) [Mauscole/minuscole](#)

ONTOLOGY MANAGER

KEYS SELECTION

RELATIONS MANAGER



ISTANCES

filtered by black list 5

- Concepts Repository
 - 2
 - A
 - able (18)
 - academic (25)
 - access (20)
 - access methods (5)
 - acm (21)
 - acm multimedia (9)
 - acquired (54)
 - acquired skills (48)
 - acquisition (6)
 - actions (8)
 - addresses (5)
 - addressing (5)
 - agreement (7)
 - allocation (26)
 - analyze (27)
 - and phase margin pulse crossing (5)
 - applications (105)
 - applied (6)

SKOS TREE

with frequencies

- Concept Schema
 - architectural (2)
 - area of software engineering (1)
 - artificial intelligence (2)
 - automated control (0)
 - computer science (0)
 - algorithm (95)
 - application (10)
 - code (8)
 - binary (4)
 - information (220)
 - notation (11)
 - xml (0)
 - database (0)
 - distributed systems (4)
 - life cycle (0)
 - programming (0)
 - condition (0)
 - e-commerce (0)
 - e-learning (2)
 - event (0)

LOG

1. skos tree node is re-loaded
2. skos tree node is re-loaded
3. **[INFO]: LOOKUP FOR acquisition (6)**
4. Related Subject:
5. http://www.unifi.it/off_form/insegnamenticc.php?cmd=2&cds=B070&cur=GEN&esa=B010480-FIRENZE&fac=200006<s=INGEGNERIA&AA=2009&codice=4480&bol=&coqnome=&nome=&f=s
6. http://www.unifi.it/off_form/insegnamenticc.php?cmd=2&cds=B070&cur=GEN&esa=B010480-FIRENZE&fac=200006<s=INGEGNERIA&AA=2010&codice=4480&bol=&coqnome=&nome=&f=s
7. Related Person:
8. Carlo Colombo (6)

CoSKOSAM

ONTOLOGY MANAGER KEYS SELECTION RELATIONS MANAGER

INSTANZE

filtra per lista nera 10

Concepts Repository

- A
- B
- C
- D
- E
- F
 - famiglia (12)
 - finito (13)
 - fisica (12)
 - flusso massimo (17)
 - fondamenti (35)
 - fondamenti di programmazione (11)
 - fondazioni (35)
 - forma (33)
 - forme (22)
 - fornire (58)
 - fornire strumenti (11)
 - frequenza obbligatoria (150)
 - funzionamento (14)
 - funzioni (55)
- G

ALBERO SKOS

con traduzione

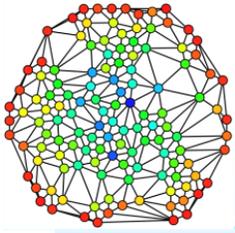
Concept Schema

- L'architettura multi-tier (EN: multi-tier architecture)
- algoritmi di ricerca (EN: search algorithms)
- architeturali (EN: architectural)
- area dell'ingegneria del software (EN: area of software)
- condizione (EN: condition)
- controllo automatizzato (EN: automated control)
- e-commerce (EN: e-commerce)
- e-learning (EN: e-learning)
- evento (EN: event)
- evento (EN: time concept)
- gestione (EN: management)
- grafico (EN: graphic)
- informatica (EN: computer science)
- intelligenza artificiale (EN: artificial intelligence)
- interazione (EN: interaction)
- matematica (EN: math)
- media (EN: media)
- metriche (EN: metrics)
- middleware (EN: middleware)
- modello (EN: model)

LOG

3. [INFO]: LOOKUP FOR fornire strumenti (11)
4. Related Subject
5. http://www.unifi.it/off_form/insegnamenticc.php?cmd=2&cds=B086&cur=B38&esa=B001635-&fac=200049<s=PSICOLOGIA&AA=2009&codice=4563&bol=&coqnome=&nome=&f=s
6. http://www.unifi.it/off_form/insegnamenticc.php?cmd=2&cds=B064&cur=D02&esa=B010314-&fac=200006<s=INGEGNERIA&AA=2009&codice=138&bol=&coqnome=&nome=&f=s
7. http://www.unifi.it/off_form/insegnamenticc.php?cmd=2&cds=B086&cur=C39&esa=B001635-&fac=200049<s=PSICOLOGIA&AA=2009&codice=139&bol=&coqnome=&nome=&f=s
8. http://www.unifi.it/off_form/insegnamenticc.php?cmd=2&cds=B064&cur=D02&esa=B010314-&fac=200006<s=INGEGNERIA&AA=2009&codice=3460&bol=&coqnome=&nome=&f=s

CoSKOSAM



Progetto OSIM

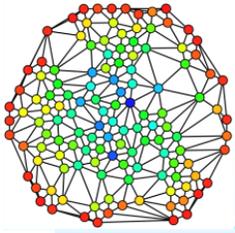
Scenario

Architettura del Sistema

Modello Ontologico

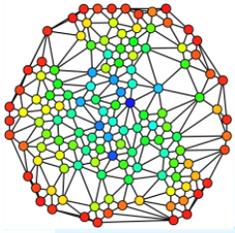
Interrogare la Conoscenza – Query Wizard

Funzionalità del Sistema



OSIM Ontology

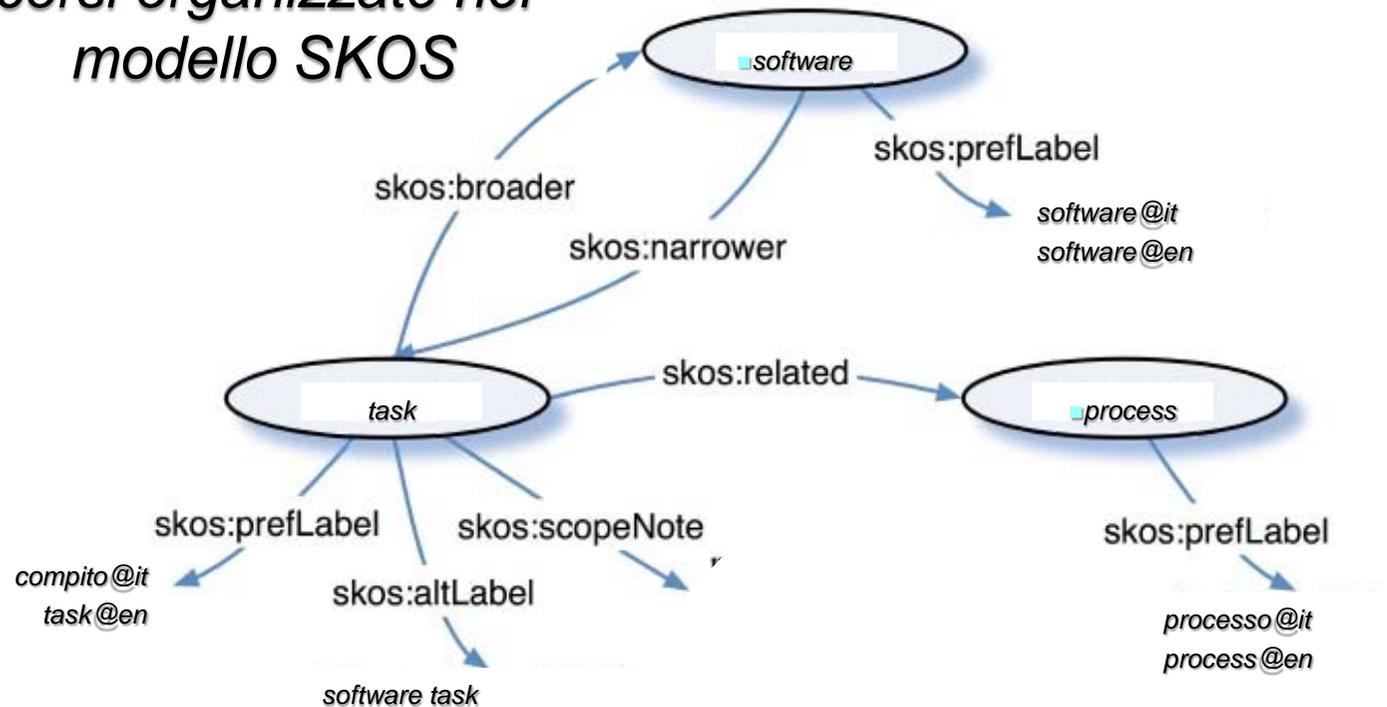
- **L'ontologia di dominio di OSIM è composta da 4 ontologie diverse**
 - ♣ *Academy Life Ontology (Unifi)* modella l'ateneo fiorentino in termini di docenti, corsi, strutture di affiliazione, facoltà, gruppi di ricerca, laboratori, ecc...
 - ♣ *Friend of a Friend (FOAF)* modella le persone in termini di professori, ricercatori, phd e relazioni tipo nome, indirizzo, e-mail, settore scientifico, relazioni di conoscenza, di co-autore di pubblicazioni, ecc
 - ♣ *Simple Knowledge Organization System (SKOS)* che modella ed organizza semanticamente le competenze delle persone e dei corsi.
 - ♣ *Time Ontology (TIME)* che modella i concetti di intervalli ed istanti temporali per quantificare temporalmente i fatti asseriti nell'ontologia.



SKOS di OSIM

SKOS Example

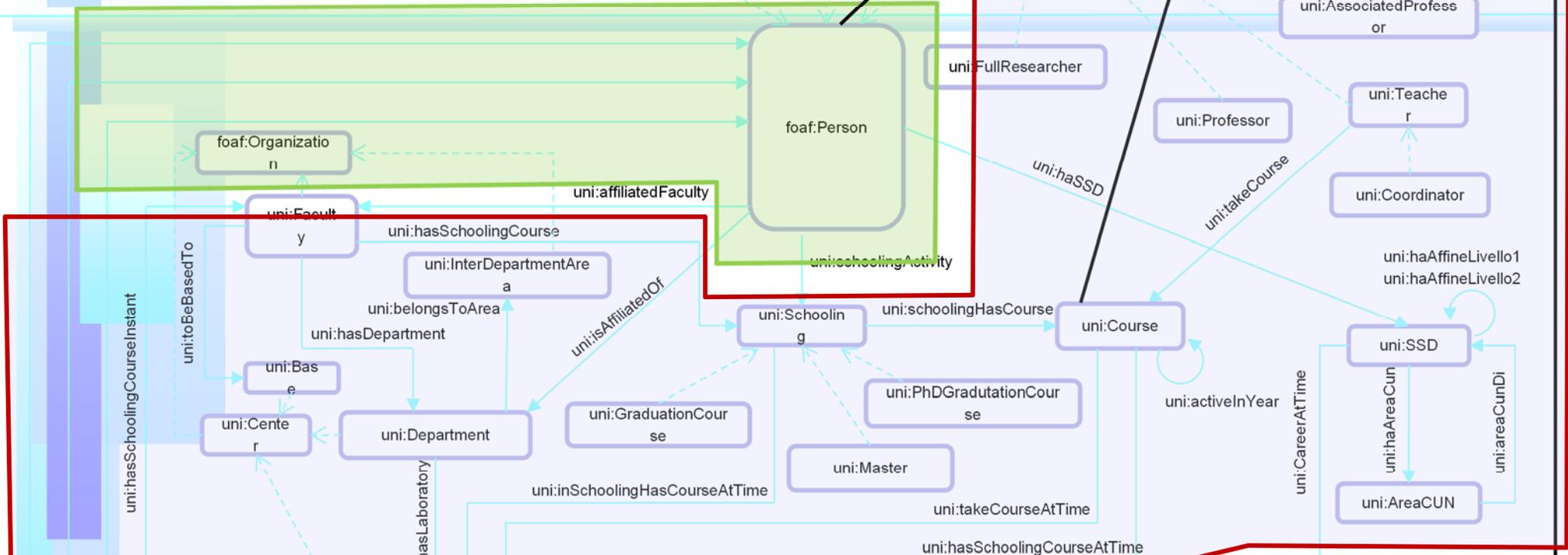
Competenze di docenti e corsi organizzate nel modello SKOS



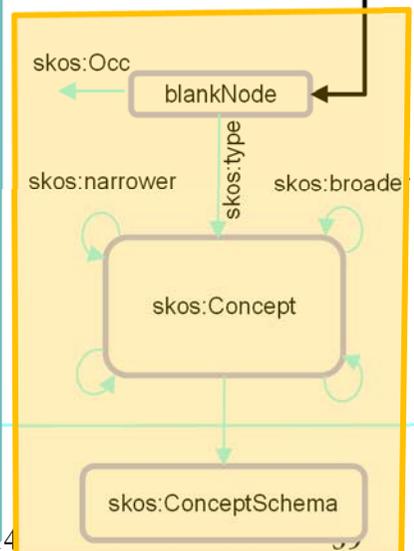
OSIM – Frammento Ontologia di dominio
Frammento di unifi ontology

Relazioni gerarchiche
 Relazioni semantiche

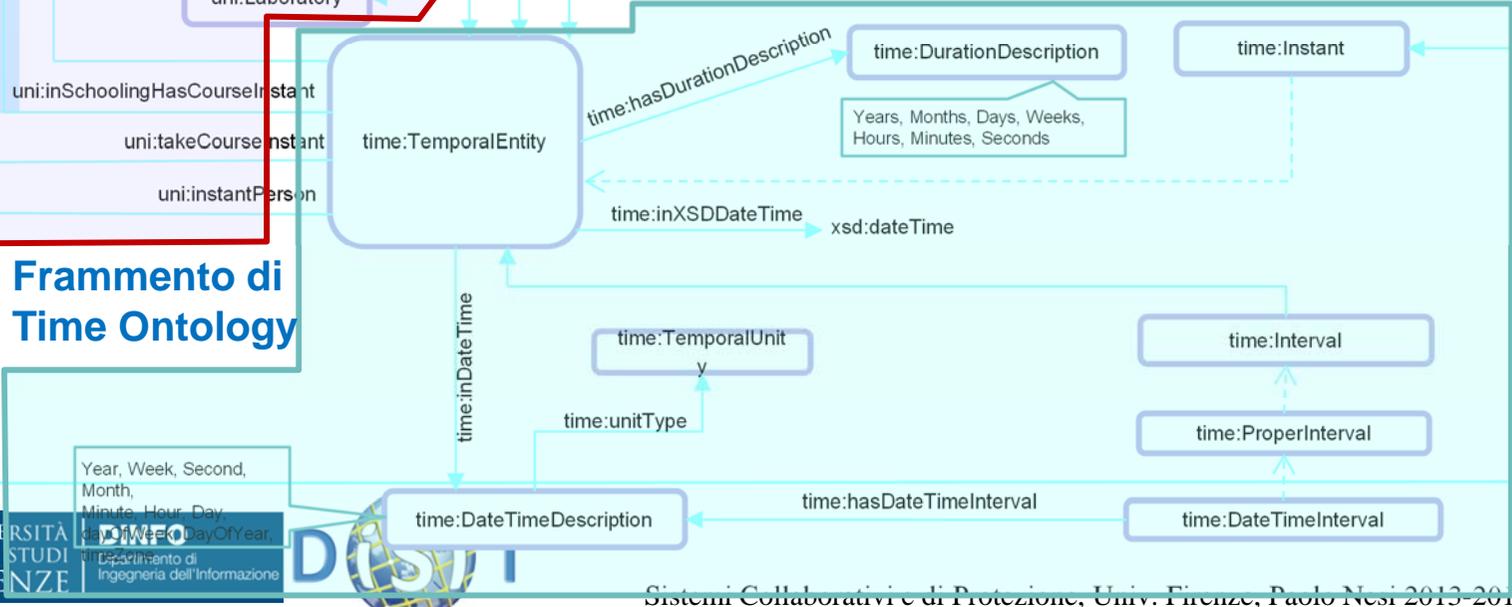
Frammento di FOAF



Frammento di SKOS



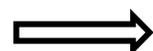
Frammento di Time Ontology



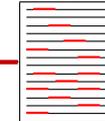
Crawler



text



NLP System



Annotated text



Semantic Repository

Concepts

- FullProfessor
- Department
- Course
- ConceptSkill
- Instant
- Interval
-

Relazioni tra entità nominali

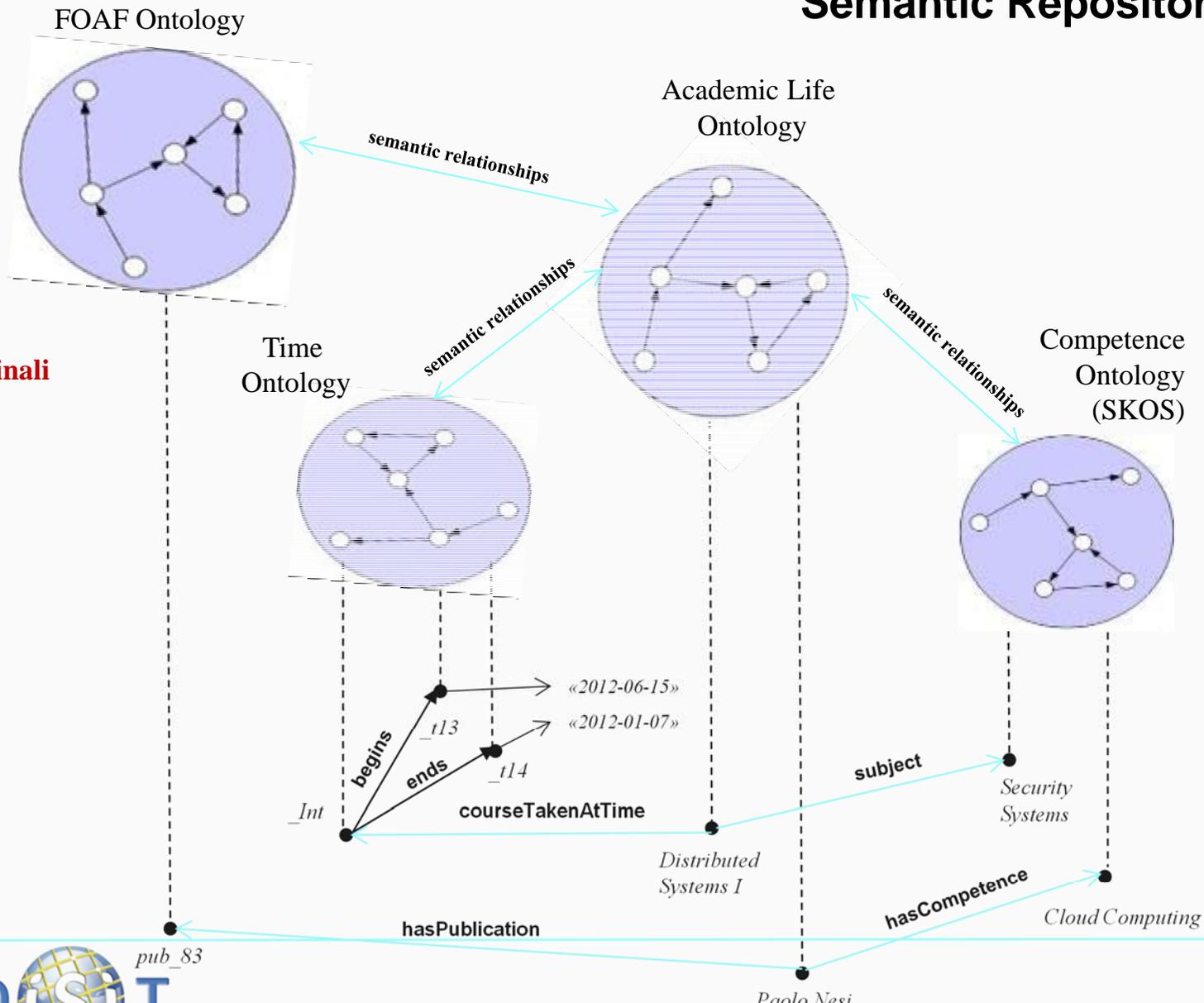
Roles

- Begins
- Ends
- courseTakenAtTime
- Subject
- hasCompetence
- hasPublication
-

Entità nominali annotate

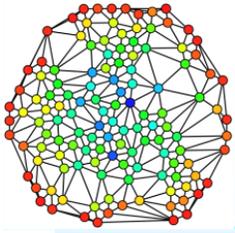
Individuals

- Paolo Nesi
- Distributed Systems I
- Security Systems
- Cloud Computing
- t13
- t14



TBOXOSIM

ABOXOSIM



Progetto OSIM

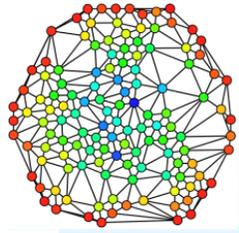
Scenario

Architettura del Sistema

Modello Ontologico

Interrogare la Conoscenza – Query Wizard

Funzionalità del Sistema



OSIM Question & Answer



Ricerca Full Text
con Logica Fuzzy

Question Answer **Query Wizard**

Ricerca
Assistita

Wizard

Quale Persona o Struttura ha questa competenza ?

Quale ha questa competenza ?

Elenca le competenze relazionate con

Elenca le competenze

Elenca le Persone del di

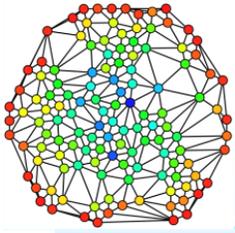
Elenca le Pubblicazioni

Dal Al

Elenca con Area CUN e settore scientifico disciplinare (SSD) dal al

Quale persona o corso presenta aspetti legati a ?

Quali competenze legate all'attività di ?



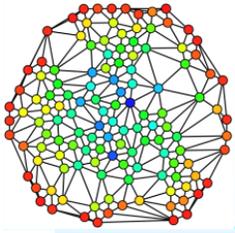
Validazione (1)

- OSIM è stato validato e confrontato con Marsilius
 - ♣ La performance dei due sistemi, su un sottoinsieme di 8 dipartimenti, è stata misurata e confrontata
- Per sistemi di IR viene tipicamente utilizzato lo standard di validazione TREC (Text REtrieval Conference)

$$Precision = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})}$$

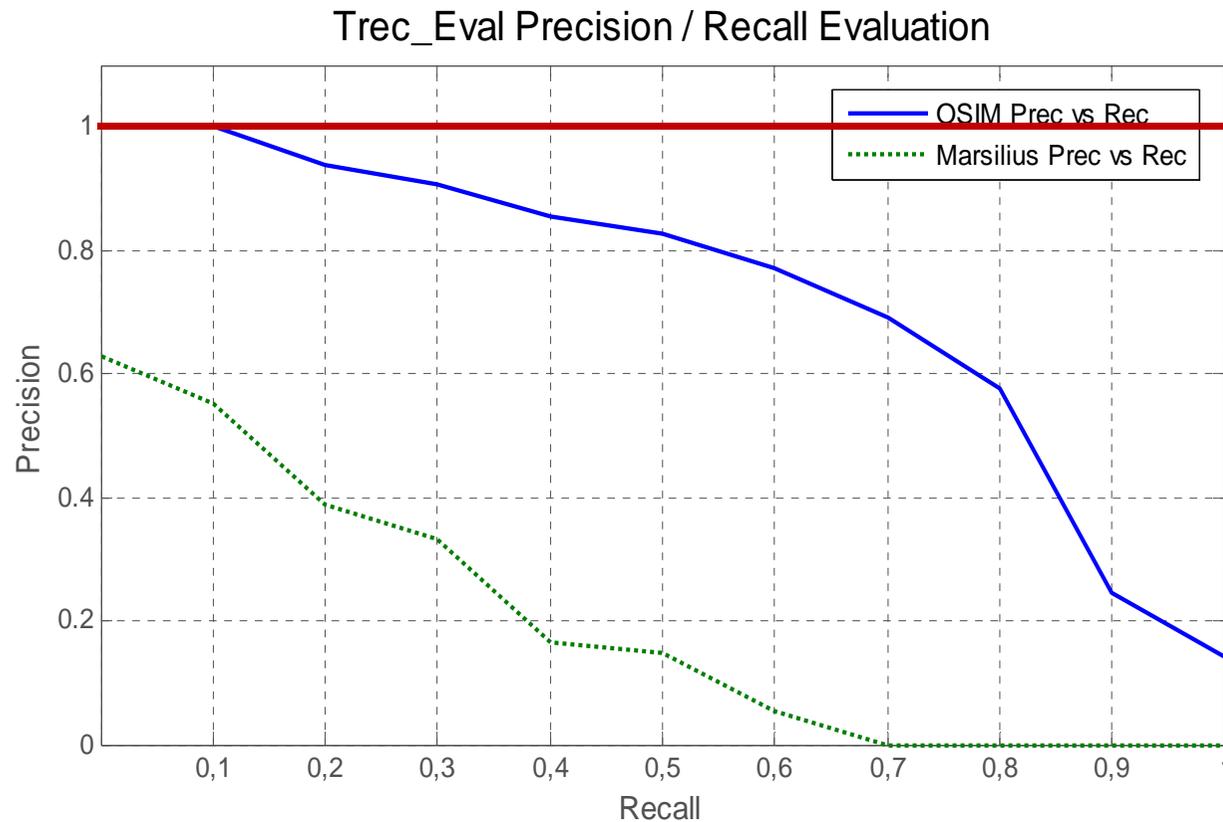
$$Recall = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})}$$

- I documenti / risultati sono considerati rilevanti se soddisfano il tipo di informazione richiesta, non solo perché contengono tutte le keywords immesse nella ricerca.
- Query di validazione eseguite da esperti dei vari domini di conoscenza analizzati

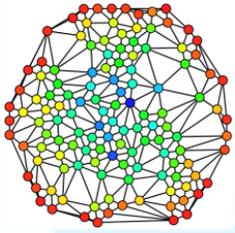


Validazione (2)

- Set di query su un sottoinsieme di 4 dipartimenti
- Profondità dei risultati fissata $N = 20$
- Curva *Precision – Recall* ottenuta con il software standard *Trec_Eval*



Ottimo ideale



Progetto OSIM

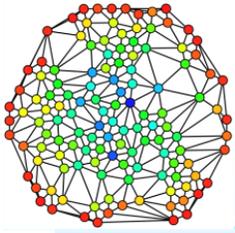
Scenario

Architettura del Sistema

Modello Ontologico

Interrogare la Conoscenza – Query Wizard

Funzionalità del Sistema



MARSILIUS: Conoscenza relativa a «Sistemi cooperativi e di protezione»



ha cercato nell'area di INGEGNERIA

Cerca in [Ricerca avanzata](#)
[Come usare il motore di ricerca](#)

Ricerca base

*Forse cercavi: sistemi **cooperative** di protezione*

Paolo Nesi's Home page

... del programma viene svolta nel Corso di **Sistemi Collaborativi e di protezione** della Laurea ... **Sistemi Cooperativi 3D: esempi di Second Life** ...
www.dsi.unifi.it/~nesi/didactical.html - 63k

Paolo Nesi's Home page

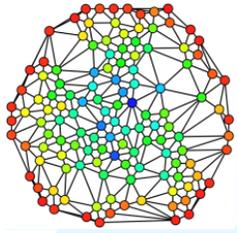
... del programma viene svolta nel Corso di **Distemi Collaborativi e di protezione** della Laurea ... **Sistemi Cooperativi 3D: esempi di Second Life** ...
www.disit.dsi.unifi.it/didactical.html - 62k

DISIT: Distributed Systems and Internet Tech. Lab. Home ...

... **sistemi di protezione** per la distribuzione di contenuti ... controllo e supervisione; **sistemi di certificazione** e ... **modelli cooperativi** per l'educazione in ...
www.disit.dsi.unifi.it/tesi.html - 60k

DISIT: Distributed Systems and Internet Tech. Lab. Home ...

... **sistemi di protezione** per la distribuzione di contenuti ... controllo e supervisione; **sistemi di certificazione** e ... **modelli cooperativi** per l'educazione in ...
www.dsi.unifi.it/~nesi/tesi.html - 60k



OSIM: Conoscenza relativa a «Sistemi cooperativi e di protezione»

OSIM Open Space Innovative Mind **BETA** TESTING *Università degli Studi di Firenze* **STUDIOR**

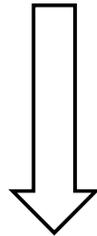
Home Documentation Search Managing Knowledge Browsing People & Publications Contact DISIT

Question Answer [Query Wizard](#)

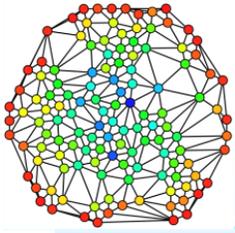
sistemi collaborativi e di protezione

Results Displayed / Found: 1 - 3 / 3 in 368641 millisec

sistemi cooperativi e di protezione (course)	score: 3.14 Freqs: 1	Paolo Nesi (full professor)	Freqs: 0 score: 2.53
		systems and security solutions (skill)	score: 2.24 Freqs: 1



“navighiamo” la conoscenza relativa al corso



OSIM: Conoscenza relativa alla competenza «Sistemi cooperativi e di protezione»

sistemi cooperativi e di protezione [a.a. 2011/2012](#), [a.a. 2012/2013](#)

Insegnante

[Paolo Nesi](#)

Argomenti

[DRM](#)

[accessibile](#)

[accessibilità](#)

[accesso](#)

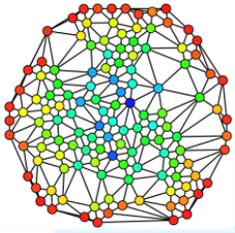
[adattamento](#)

[aggregazione](#)

[aggregazione contenuti](#)

[algoritmi di clustering confronto](#)

[algoritmi e architetture](#)

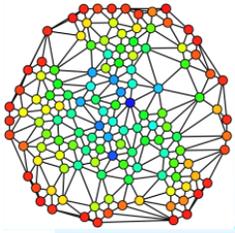


OSIM: quali sono le competenze che si acquisiscono sostenendo l'esame di «Sistemi cooperativi e di protezione» ?

Results Displayed / Found: 31 - 60 / 556 in 31 millisec

profiles (skill)	Freqs: 2	score: 1.0	causality (skill)	Freqs: 6	score: 1.0
interfaces (skill)	Freqs: 11	score: 1.0	comunicazione (skill)	Freqs: 41	score: 1.0
time (skill)	Freqs: 44	score: 1.0	calendar (skill)	Freqs: 353	score: 1.0
study (skill)	Freqs: 513	score: 1.0	laboratorio (skill)	Freqs: 69	score: 1.0
synchronization (skill)	Freqs: 8	score: 1.0	examination (skill)	Freqs: 147	score: 1.0
programming (skill)	Freqs: 170	score: 1.0	development (skill)	Freqs: 24	score: 1.0
assessment (skill)	Freqs: 352	score: 1.0	reconstruction (skill)	Freqs: 331	score: 1.0
html (skill)	Freqs: 85	score: 1.0	edition (skill)	Freqs: 30	score: 1.0
tutorial (skill)	Freqs: 101	score: 1.0	letter (skill)	Freqs: 309	score: 1.0
items (skill)	Freqs: 78	score: 1.0	license (skill)	Freqs: 305	score: 1.0
process (skill)	Freqs: 34	score: 1.0	procedures (skill)	Freqs: 9	score: 1.0
caratteristiche (skill)	Freqs: 81	score: 1.0	phd (skill)	Freqs: 305	score: 1.0
delays (skill)	Freqs: 2	score: 1.0	conversion (skill)	Freqs: 5	score: 1.0
scheduling (skill)	Freqs: 33	score: 1.0	search (skill)	Freqs: 581	score: 1.0
information (skill)	Freqs: 914	score: 1.0	xml (skill)	Freqs: 4	score: 1.0

Previous 1 **2** 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 Next



OSIM: Conoscenza relativa alla competenza «programming»

programmazione (skill)

Competenze collegate

Concetti broader

[informatica](#)

Concetti narrower

[esperienza nella programmazione](#)

[espressioni regolari](#)

[fondamenti di programmazione](#)

[funzioni](#)

[lingua](#)

[linguaggio di programmazione](#)

[linguaggio macchina](#)

[livello di programmazione](#)

[modello di programmazione](#)

Concetti related

Nessun concetto presente

Persone

Persone con questa competenza

[Michele Boreale](#) (4/170)

[Michele Loreti](#) (1/170)

[Paola Cappanera](#) (1/170)

[Paolo Frasconi](#) (9/170)

[Paolo Nesi](#) (12/170)

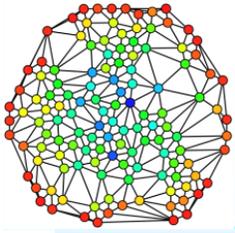
[Pierluigi Crescenzi](#) (7/170)

[Pietro Pala](#) (1/170)

[Rosario Pugliese](#) (9/170)

[Simone Marinai](#) (2/170)

[Stefano Berretti](#) (13/170)



OSIM: Conoscenza relativa alla persona

«Paolo Nesi»

Paolo Nesi (full professor)

Informazioni

Facoltà

[ingegneria](#)

Corsi di Laurea

[ingegneria elettronica e delle telecomunicazioni](#)

[ingegneria informatica](#)

Affiliazioni

[Sistemi E Informatica](#)

Corsi

[sistemi cooperativi e di protezione](#) (cod. b010488)

[sistemi distribuiti](#) (cod. b003734)

Links

[Penelope web page](#)

[Information page](#)

Competenze

[DRM](#) (10/10)

[Wedelmusic](#) (9/9)

[Windows](#) (1/22)

[accessibilità](#) (5/302)

[adattamento](#) (3/11)

[aggregazione](#) (1/2)

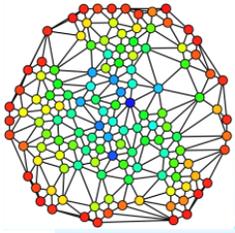
[algoritmo](#) (1/108)

[analisi](#) (11/253)

[applicazione](#) (1/31)

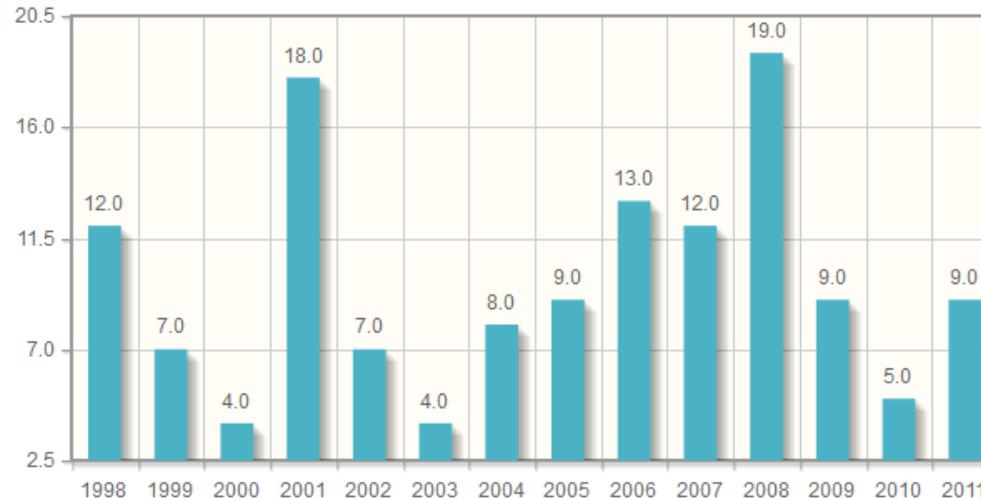
Pubblicazioni

[Link alle pubblicazioni](#)



OSIM: Conoscenza relativa alle pubblicazioni e ai coautori di «Paolo Nesi»

Totale pubblicazioni: 167



Anno:

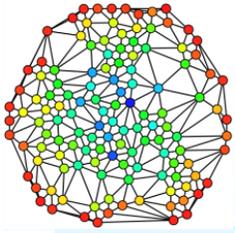
[1998](#) (12) [1999](#) (7) [2000](#) (4) [2001](#) (18) [2002](#) (7) [2003](#) (4) [2004](#) (8) [2005](#) (9) [2006](#) (13) [2007](#) (12) [2008](#) (19) [2009](#) (9) [2010](#) (5) [2011](#) (9)

[Elenco di tutte le pubblicazioni](#) (167)

Autori che hanno lavorato con questa persona:

[ARGENTI FABRIZIO](#) (Registrato CINECA)
[BALDASSARRE ANTONIO](#) (Registrato CINECA)
[BELLINI PIERFRANCESCO](#) (Registrato CINECA)
[BRUNO IVAN](#) (Registrato CINECA)
[BUCCI GIACOMO](#) (Registrato CINECA)
[CENNI D.](#) (Non Registrato CINECA)
[CENNI DANIELE](#) (Registrato CINECA)
[DE LUCIA M.](#) (Non Registrato CINECA)
[DE LUCIA MAURIZIO](#) (Registrato CINECA)
[EVANGELISTI ATTILIO](#) (Registrato CINECA)

[Visualizza le pubblicazioni in comune](#) (2)
[Visualizza le pubblicazioni in comune](#) (1)
[Visualizza le pubblicazioni in comune](#) (41)
[Visualizza le pubblicazioni in comune](#) (22)
[Visualizza le pubblicazioni in comune](#) (5)
[Visualizza le pubblicazioni in comune](#) (3)
[Visualizza le pubblicazioni in comune](#) (5)
[Visualizza le pubblicazioni in comune](#) (1)
[Visualizza le pubblicazioni in comune](#) (1)
[Visualizza le pubblicazioni in comune](#) (1)



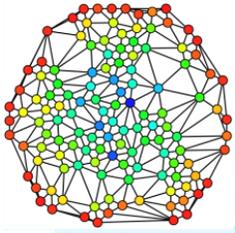
OSIM: Conoscenza relativa alle pubblicazioni di «Paolo Nesi»

Publications count: 167

Legenda

📄 1a - Articolo su rivista ISI | 📄 1b - Articolo su rivista NON ISI | 📖 2a - Art/Cap/Saggio libro scient/tech | 📖 3f - Libro scientifico/tecnico | 📄 4a - Articolo/Proceedings in atti di congresso | 📄 6a - Brevetto attivo | 📖 7d - Curatela di libro scientifico/tecnico | 📄 A Altro

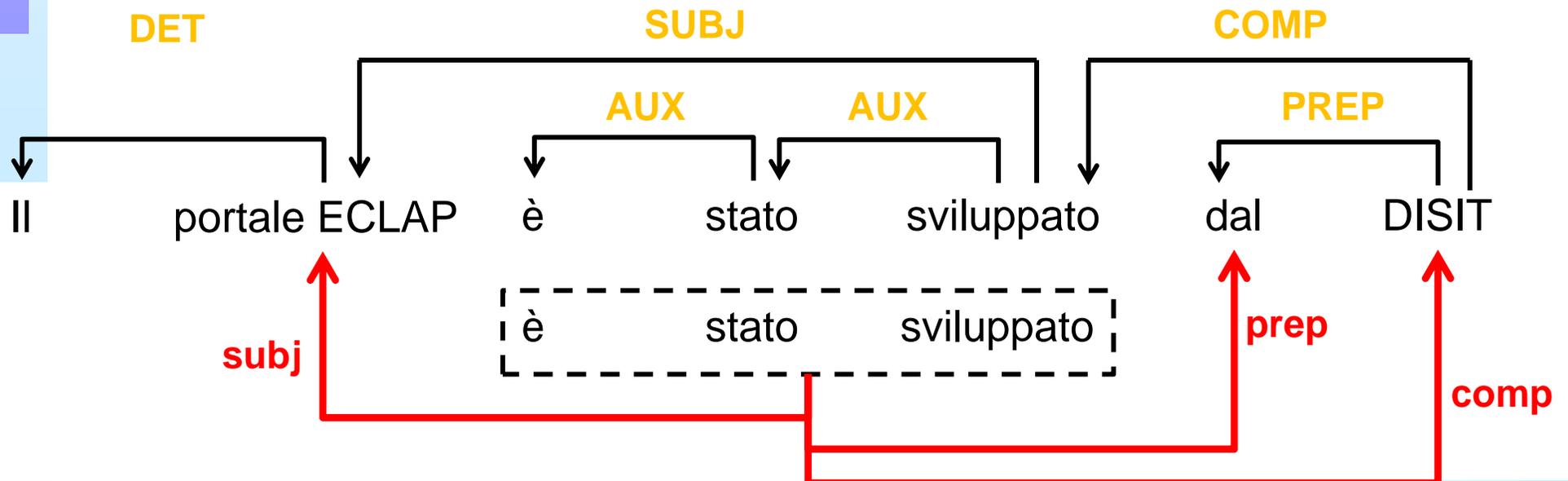
1. 📄 A **Paolo Nesi, Raffaella Santucci**, *ECLAP 2012 Conference on Information Technologies for Performing Arts, Media Access and Entertainment*, FUP - FIRENZE UNIVERSITY PRESS, 2012. [Info](#).
2. 📄 A **Pierfrancesco Bellini, Ivan Bruno, Paolo Nesi, Daniele Cenni**, *Micro Grids for Scalable Media Computing and Intelligence in Distributed Scenarios*, *IEEE MULTIMEDIA*, , Vol. 19, n. 2, Month, pp. 69-79, 2012. [Info](#).
3. 📄 D. Cenni, P. Nesi, M. Paolucci, *Functionalities and Flow Analyses of Knowledge Oriented Web Portals*, Proceeding of the 18th International Conference on Distributed Multimedia Systems, DMS 2012, Eden Roc Renaissance Miami Beach, USA, SEKE software engineering and knowledge engineering institute, 9-11 August, 2012., pp. 1-6, 2012. [Info](#).
4. 📄 P. Bellini, D. Cenni, P. Nesi, *On the Effectiveness and Optimization of Information Retrieval for Cross Media Content*, Proceeding of the KDIR 2012 is part of IC3K 2012, International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 4-7 October 2012, Barcelona, Spain., pp. 1-8, 2012. [Info](#).
5. 📄 Pierfrancesco Bellini, Sergio Boncinelli, Francesco Grossi, Marco Mangini, Luca Malandrino, Paolo Nesi, *Mobile Emergency: shortening communications in hospital*, Media Tablets & Apps for Medicine, Health, and Home-Care, R&D Workshops & Demos, 1st-2nd March 2012, Politecnico di Milano, Milano., pp. 1-6, 2012. [Info](#).
6. 📄 Pierfrancesco Bellini, Ivan Bruno, Daniele Cenni, Alice Fuzier, Nicola Mitolo, Paolo Nesi, Michela Paolucci, *Mobile Medicine: Smart Applications and Content for Medical Training and Support*, Media Tablets & Apps for Medicine, Health, and Home-Care, R&D Workshops & Demos, 1st-2nd March 2012, Politecnico di Milano, Milano., pp. 1-6, 2012. [Info](#).

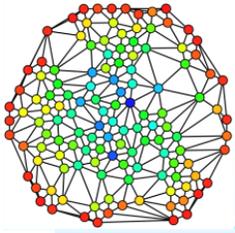


Relazioni verbali (1)

- Il sistema permette di estrarre relazioni semantiche non tassonomiche dalle pagine dell'Università di Firenze che riguardano i corsi e le persone
- Per estrarre le relazioni semantiche dai testi, il sistema analizza le frasi, genera un grafo a partire dall'albero delle dipendenze di ogni frase e individua eventuali **PATTERN PREDEFINITI** presenti sul grafo

Esempio: Il portale ECLAP è stato sviluppato dal DISIT



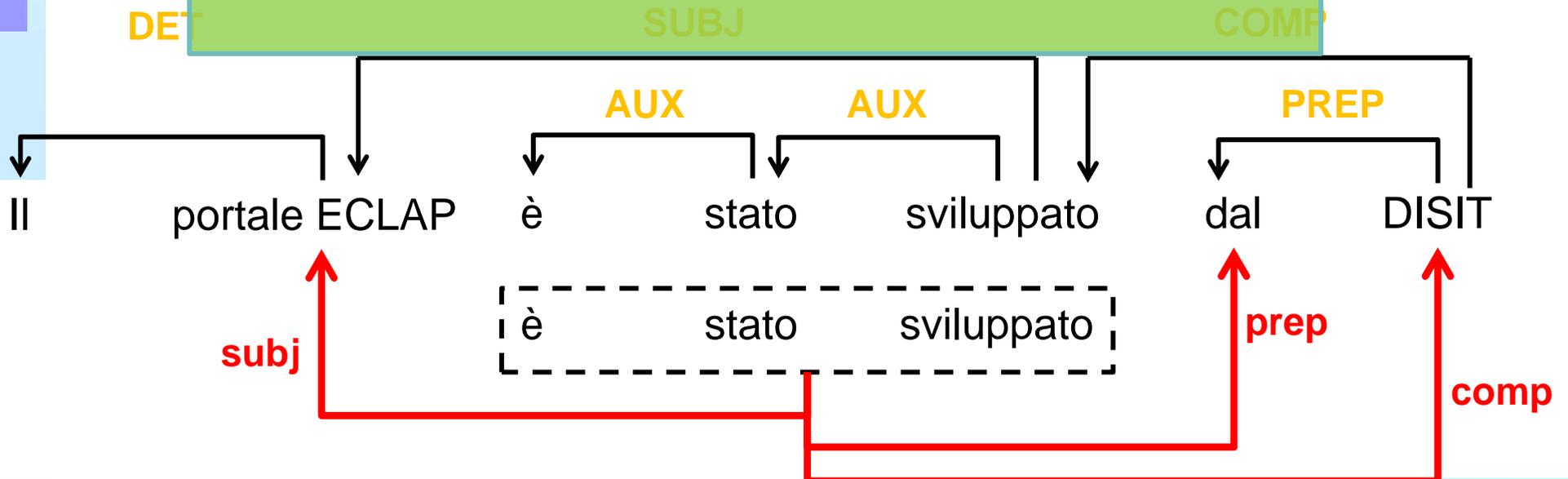


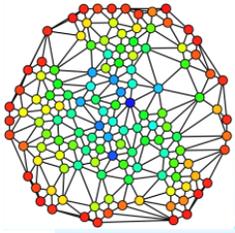
Relazioni verbali (1)

- Il sistema permette di estrarre relazioni semantiche non tassonomiche dalle pagine dell'Universit di Firenze che riguardano i corsi e le persone
- Per estrarre le relazioni semantiche dai testi, il sistema analizza le frasi, genera un grafo a partire dall'albero delle dipendenze di ogni frase e individua eventuali **PATTERN PREDEFINITI** presenti sul grafo

DISIT sviluppa portale ECLAP

Esempio: Il portale ECLAP è stato sviluppato dal DISIT

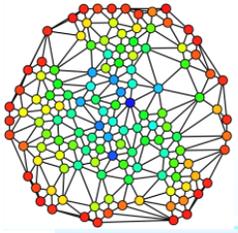




Relazioni verbali (2)

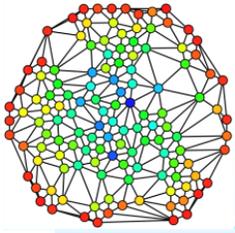
I **PATTERN PREDEFINITI** sono di 5 tipi

- Pattern basati sui **verbi sostantivizzati**: hanno lo scopo di individuare espressioni come «Analisi di tecniche di Information Retrieval». Generano relazioni del tipo «soggetto implicito - analizza - tecniche di Information Retrieval».
- Pattern basati su **verbi coniugati alla prima persona**: hanno lo scopo di individuare frasi del tipo «Abbiamo realizzato un progetto». Generano relazioni del tipo «soggetto implicito - realizza – progetto».
- Pattern basati su **verbi coniugati alla terza persona singolare**. Questi pattern si possono suddividere ulteriormente in due tipi:
 - ❖ il soggetto è specificato nella frase ed è individuato dal pattern: hanno lo scopo di individuare frasi del tipo «Tizio sviluppa in C++». Generano relazioni del tipo «Tizio - sviluppa - in - C++».
 - ❖ il soggetto non è specificato nella frase: hanno lo scopo di individuare frasi del tipo «analizza tecniche di Information Retrieval». Generano relazioni del tipo «soggetto implicito - analizza - tecniche di Information Retrieval»



Relazioni verbali (3)

- Pattern basati sui **verbi all'infinito**: hanno lo scopo di individuare frasi come «implementare algoritmi». Generano relazioni del tipo «soggetto implicito - implementa – algoritmi»
- Pattern basati sui **verbi passivi**: hanno lo scopo di individuare frasi come «l'algoritmo implementato da Tizio». Generano relazioni del tipo «Gianni - implementa – algoritmo»



Relazioni verbali - Manager

Welcome root [Logout](#) [OSIM Managing Knowledge HOME](#)

dipartimento di matematica per le decisioni - crawler is running

english

ONTOLOGY MANAGER

KEYS SELECTION

RELATIONS MANAGER



L'amministratore avvia la fase di estrazione delle relazioni verbali

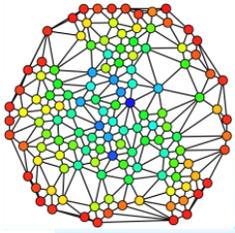


18/40 (#392)

Subject	Verb	Object	Frequency	Sentences	Enable/Disable	Delete
Luigi Vannucci	impiega	simulazione	4		✗	
Domenico Menicucci	attua	scienze	3		✗	
Domenico Menicucci	calcola	numeri	3		✗	
Augusto Bellieri Dei Belliera	insegna	scienze	3		✗	
Augusto Bellieri Dei Belliera	modella	analisi economica	3		✗	
Augusto Bellieri Dei Belliera	assicura	finanziarie	3		✓	
Andrey Sarychev	modella	dinamica	3		✗	
Andrey Sarychev	opera	insiemi	3		✗	
Andrey Sarychev	risolve	intervallo	3		✓	
Andrey Sarychev	insegna	cognomi	3		✓	

RELATION DETAILS VIEWER of Augusto Bellieri Dei Belliera

Infinite form	Third person form	Competence	Sentence
modellare	modella	analisi economica	un modello per l'analisi economica, finanziaria e gestionale di operazioni di leasing.



Relazioni verbali – Query Wizard

Question Answer [Query Wizard](#)

Wizard

Quale Persona o Struttura ha questa competenza ?

Quale ha questa competenza ?

Elenca le competenze relazionate con

Elenca le competenze

Elenca le Persone del di

Elenca le pubblicazioni

dal al

Elenca con area CUN

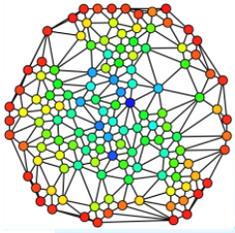
e settore scientifico disciplinare (SSD)

dal al

Chi ?

Quale competenza ?

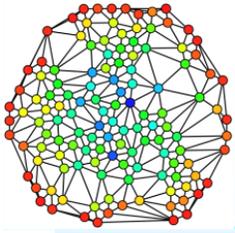




Stato (1)

- Publications ingested:
 - ♣ all those which are present on CINECA data base
 - ♣ for about 6000 registered people of the UNIFI
- 49 Departments:

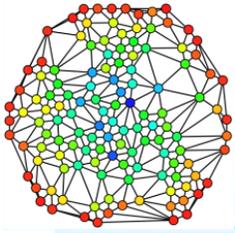
Dipartimento	n° Keywords	Gazetteer	n° Doc.	Fasi Successive
Area Critica Medico Chirurgica	10891	1367 (manuale)	721	Competenze
Anatomia, Istologia e Medicina Legale	5205	693 (manuale)	400	Competenze
Architettura - Disegno, Storia, Progetto	6534	406 (man. + prop.)	245	Competenze
Biotechnologie Agrarie	8114	707 (man. + prop.)	293	Competenze
Biologia Evoluzionistica Leo Pardi	7789	386 (man. + prop.)	238	Competenze
Chimica "Ugo Schiff"	12354	761 (manuale)	434	Competenze, Relazioni
Costruzioni e Restauro	4421	370 (man. + prop.)	178	
Economia, Ingegneria, Scienze e Tecnologie Agrarie e Forestali (D.E.I.S.T.A.F.)	5967	466 (man. + prop.)	240	Competenze
Scienze delle Produzioni Vegetali, del Suolo e dell'Ambiente Agroforestale (DI.P.S.A.)	5448	321 (man. + prop.)	196	Competenze
Diritto Comparato e Penale	2227	250 (man. + prop.)	87	Competenze
Diritto Privato e Processuale	2421	332 (man. + prop.)	128	Competenze
Diritto Pubblico "Andrea Orsi Battaglini"	2424	366 (man. + prop.)	237	Competenze
Diritto dell'Economia	2122	343 (manuale)	195	Competenze, Relazioni
Scienze Economiche	5115	950 (manuale)	285	Competenze, Relazioni
Elettronica e Telecomunicazioni	6660	430 (manuale)	255	Competenze
Energetica "Sergio Stecco"	4868	413 (man. + prop.)	42	Competenze



Stato (2)

Dipartimento	n° Keywords	Gazetteer	n° Doc.	Fasi Successive
Fisica e Astronomia	7944	621 (proposed)	357	Solo Keywords
Fisiopatologia Clinica	928	147 (proposed)	467	Solo Keywords
Filosofia	3948	342 (proposed)	180	Solo Keywords
Farmacologia Preclinica e Clinica	7119	625 (man. + prop.)	243	Competenze
Ingegneria Civile e Ambientale	6502	461 (man. + prop.)	215	Competenze
Italianistica	5735	521 (proposed)	252	Solo Keywords
Lingue, Letterature e Culture Compare	5507	361 (proposed)	467	Solo Keywords
Matematica per le Decisioni	2258	292 (proposed)	122	Competenze, Relazioni
Meccanica	2958	523 (manuale)	92	Competenze, Relazioni
Matematica "Ulisse Dini"	522	77 (proposed)	274	Solo Keywords
Scienze Farmaceutiche	4622	479 (man. + prop.)	166	Competenze
Scienze Aziendali	3584	693 (man. + prop.)	280	Competenze, Relazioni
Sistemi e Informatica	7803	1741 (manuale)	335	Competenze, Relazioni
Teoria e Storia del Diritto	6573	560 (man. + prop.)	43	Competenze
Urbanistica e Pianificazione del Territorio	4176	294 (man. + prop.)	167	Competenze

- Navigazione tra profili utente (contenenti strutture di affiliazione, corsi, competenze ingerite, relazioni semantiche estratte ecc.) e relazioni di collaborazione, pubblicazioni
- Query Wizard semantico
 - ♣ Ricerca full text / assistita con tecnologia fuzzy e indicizzazione semantica



Outline

● Sistemi di Web Crawling

- ♣ Introduzione
- ♣ Strategie di Crawling
- ♣ Robot Exclusion Protocol
- ♣ Concorrenza

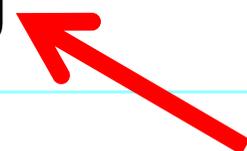
● NLP: Natural Language Processing

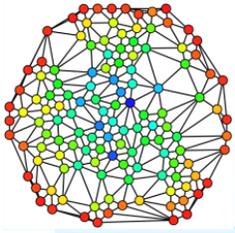
- ♣ Introduzione
- ♣ Stato dell'Arte
- ♣ Fasi dell'Elaborazione in Linguaggio Naturale
- ♣ NLP Tools: GATE

● Progetto OSIM

- ♣ Scenario
- ♣ Architettura del Sistema
- ♣ Modello Ontologico
- ♣ Interrogare la Conoscenza – Query Wizard
- ♣ Funzionalità del Sistema

● Sistemi di Question-Answering in NLP: Aqualog





Q&A in Natural Language: AQUALOG (2)

- **AQUALOG: Question & Answer System**

The screenshot shows a web browser window displaying the UniFI CercaChi website. The browser's address bar shows the URL: `www.unifi.it/cercachi/scheda.php?f=p&codice=4190&bol=AND&cognome=marinai&nome=simone`. The website header includes the UniFI logo and the text "Università degli Studi di Firenze". Below the header, there is a navigation menu with options like "home", "Ateneo", "Studenti", "Relazioni internazionali", "Ricerca", "Personale", and "Biblioteche". The main content area is titled "CercaChi Database anagrafico del personale" and displays the profile of Simone Marinai, a professor associated with the Department of Systems and Informatics. The profile includes contact information such as address, phone, fax, and email. A calendar for November 2012 is visible on the left side of the page. The footer contains accessibility information, copyright details, and logos for MD-Pro, MySQL, and PHP.

UniFI - CercaChi - Scheda personale - Simone Marinai - Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Strumenti Aiuto

UniFI - CercaChi - Scheda personale - Si... +

www.unifi.it/cercachi/scheda.php?f=p&codice=4190&bol=AND&cognome=marinai&nome=simone

Università degli Studi di Firenze

Università degli Studi di Firenze - P.zza S.Marco, 4 - 50121 Firenze - Centralino +39 055 27571 - E-mail: urp@unifi.it - Posta certificata: ufficio.urp@pec.unifi.it - P.IVA/Cod.Fis. 01279680480

home

Ateneo

Studenti

Students from abroad

Accoglienza studenti con disabilità

Relazioni internazionali

Ricerca

Personale

Biblioteche

Agenda eventi

Novembre 2012

D	L	M	M	G	V	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

CercaChi Database anagrafico del personale

Informazioni su:

Simone **Marinai** Via S. Marta, 3 - 50139 Firenze
Tel. 055 4796452
Professore associato Fax 055 4796363
E-mail simone.marinai@unifi.it - simone@dsi.unifi.it

Afferente a Dipartimento di Sistemi e Informatica

Curriculum, offerta formativa e pubblicazioni

Home > CercaChi > Cerca persona > Risultato > Simone Marinai

urp

cercachi

cercadove

mappa

web@

RSS 2.0

RSS info

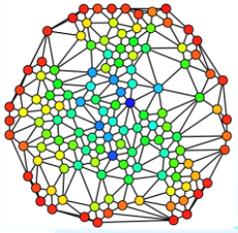
Accessibilità

© Progettazione e realizzazione: CSIAF-Ufficio Siti web | Ufficio Stampa e Redazione sito web - Redazione -

Sito ottimizzato per IE 6, Netscape 7, Opera 8, Firefox 2, Google Chrome e loro versioni superiori.

POWERED BY MD-Pro MySQL PHP

Tutti i loghi e marchi in questo sito sono di proprietà dei rispettivi proprietari.
Questo sito è stato creato con MD-Pro. MD-Pro è un software libero rilasciato sotto la licenza GNU/GPL.



Q&A in Natural Language: AQUALOG (3)

- **AQUALOG: Question & Answer System**

- ♣ **Question:** Query in Linguaggio Naturale (per ora supportata solo la lingua inglese)
- ♣ **Answer.** Processo basato su ontologie predefinite o create dall'utente (indipendente dall'ontologia)

- Ontologie espresse da triple:

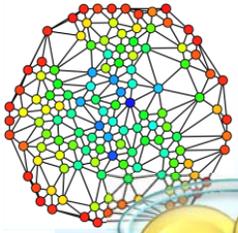
<oggetto> <predicato> <oggetto>

- Il sistema di Q&A deve convertire le query **Q** (in NL) in una forma compatibile **Q*** con l'ontologia.

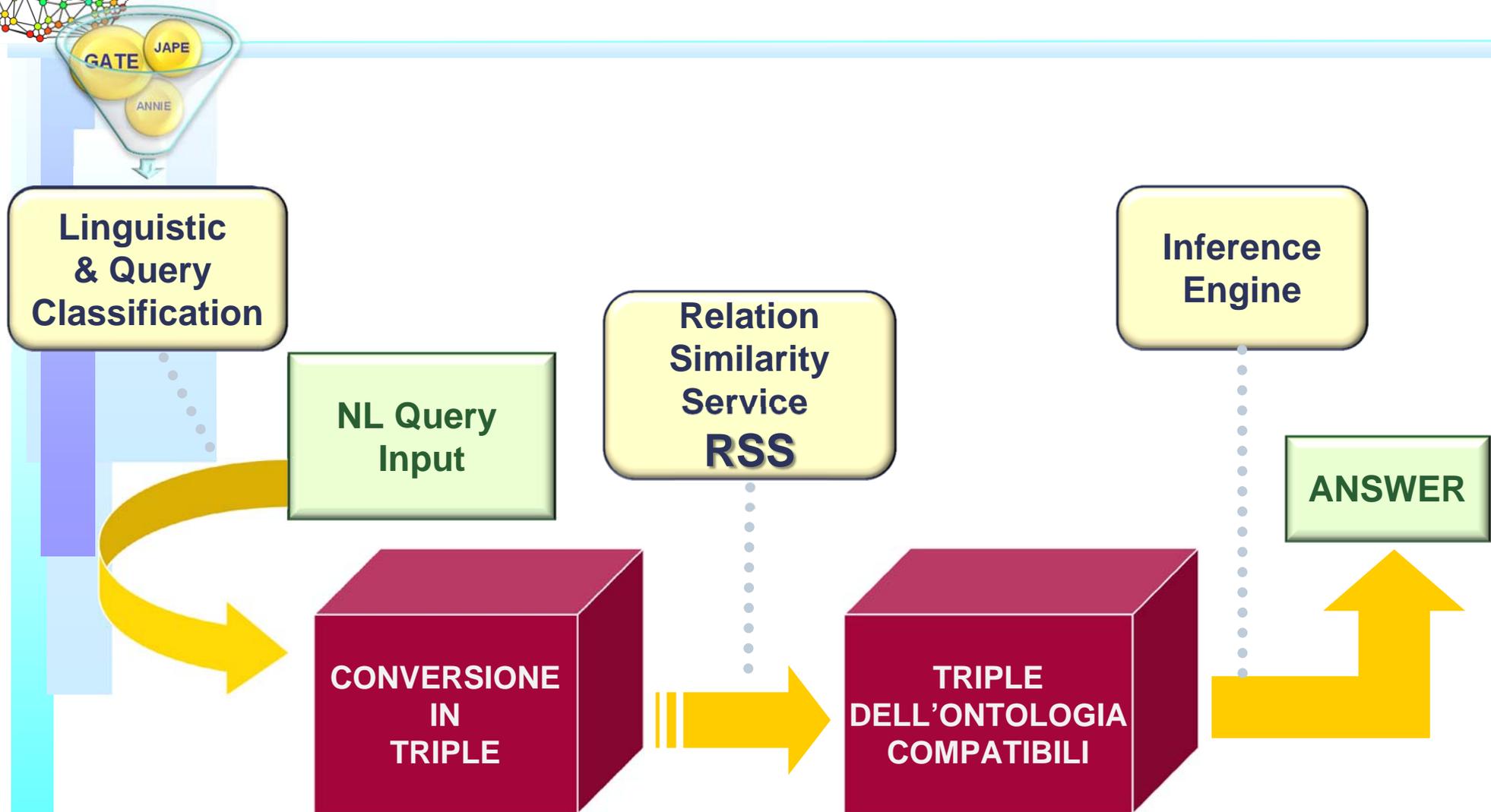
Q : “What is the homepage of Luca who has an interest on Semantic Web?”

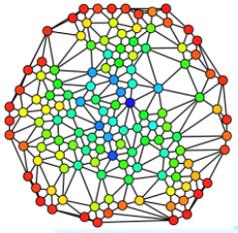
Q* : 1) <what? , has-web-page , Luca Rossi>

2) <person? , has-research-interest , Semantic Web>



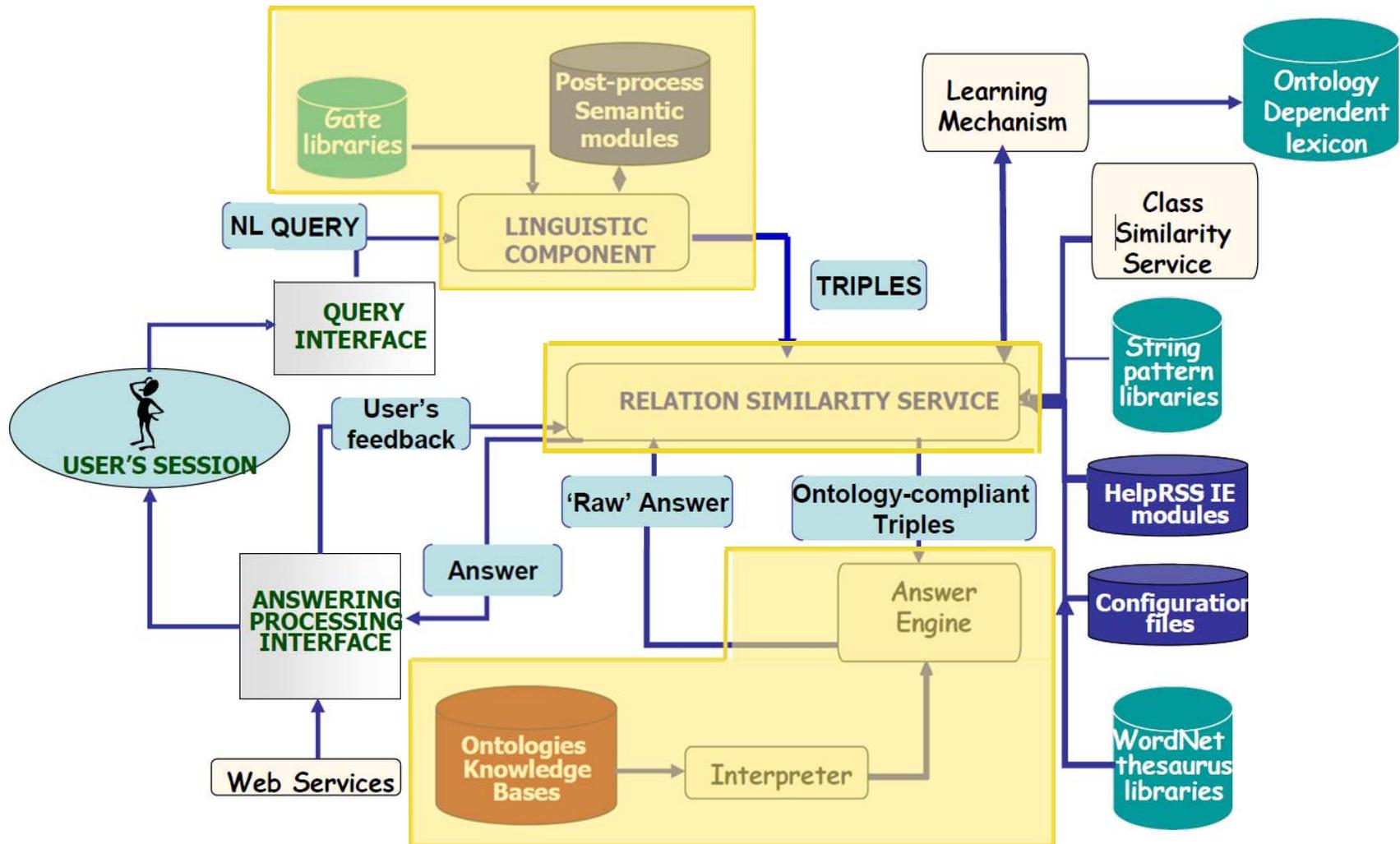
Q&A in Natural Language: AQUALOG (4)

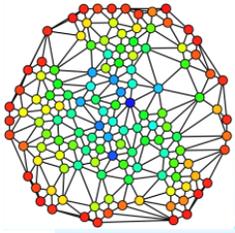




Q&A in Natural Language: AQUALOG (5)

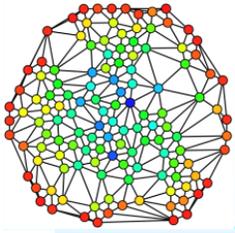
● AQUALOG: Architettura





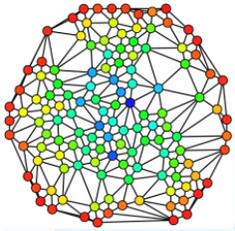
Q&A in Natural Language: AQUALOG (6)

- **Il Componente Linguistico**, attraverso le grammatiche JAPE, definisce 14 differenti pattern per identificare diverse tipologie di domande:
 - **WH-GENERIC TERM - term relation term - :**
Es.: *In which projects is Paolo Nesi working on?*
Tripla : *projects – working – Paolo Nesi*
 - **WH- UNKNOWN TERM - value relation term -:**
Es.: *Name all teachers in Disit departement*
Tripla : *value?- teachers – Disit departement*
 - **WH- UNKNOWN RELATION: term - ? term**
Es.: *Are there any projects about semantic web?*
Tripla : *project - ? - semantic web*
 - **DESCRIPTION: term**
Es.: *Who is Paolo Nesi?*
Termine : *Paolo Nesi*
 - **AFFIRMATIVE/NEGATIVE: term - relation - term**
Es.: *Is Enrico Vicario a resaercher?*
Tripla : *Enrico Vicario – is a – reseaarcher*
 - **WH- 3 TERMS: term - relation - term - term**
Es.: *Does anybody works on Semantic Web in Disit?*
2 Triple : *person/organizazion – works – Semantic Web – Disit*
 - **WH- 3 TERMS (in first clause): term - term - relation - term**
Es.: *Which courses about algebra are taken by Michele Basso?*
2 Triple : *courses – algebra – taken – Michele Basso*
 - **WH- 3 TERMS UNKNOWN RELATION: term - ? - term – term**
Es.: *Is there any projects about Semantic Web in Disit?*
2 Triple : *project - ? - Semantic Web – Disit*
 - **WH- 3 TERMS UNKNOWN TERM WITH CLAUSE: value - relation - term - term**
Es.: *Show me all publications about Semantic Web in Disit*
Tripla : value? - publication – Semantic Web – Disit
 - **WH- COMBINATION AND:**
Es.: *Does anybody works in Disit and in Micc?*
2 Triple : *person/organization – works – Disit*
person/organization – works – Micc
 - **WH- COMBINATION OR:**
Es.: *Who is interesting in algebra or in systems?*
2 Triple : *persono/organization – interesting – algebra*
person/organization – interesting – systems
 - **WH- COMBINATION CONDITIONED:**
Es.: *What are contact details for researchers working on Eclap project?*
2 Triple : *Which is – contact details – resaerchers*
Which is – working – Eclape project
 - **WH- GENERIC WITH WH-CLAUSE:**
Es.: *Which researcher , that take coruse of Sistemi Distribuiti, are working in Disit?*
Tripla : *researcher – takeCourse – Sistemi Distribuiti*
researcher – working – Disit
 - **2- PATTERNS:**
Es.: *Who are the researcher in micc , who have competence in language?*
2 Triple *di categoria [Wh-generic term]*



Q&A in Natural Language: AQUALOG (7)

- **Un sistema Q&A in linguaggio naturale** è un sistema molto complesso, soggetto a errori. Una possibile classificazione dei fallimenti:
 - ♣ **FALLIMENTO CONCETTUALE** : ontologia non correttamente popolata.
 - ♣ **FALLIMENTO DI SERVIZIO**: mancanza di appropriati servizi, definiti sull'ontologia.
 - ♣ **FALLIMENTO LINGUISTICO**: il componente NLP non è in grado di generare la rappresentazione intermedia (N-triple, XML, RDF...). Tuttavia, la domanda può essere riformulata, ed è possibile che il sistema produca successivamente una risposta.
 - ♣ **FALLIMENTO DEL MODELLO DEI DATI**: Si verifica quando la query in NL è troppo complessa per la rappresentazione intermedia.
 - ♣ **FALLIMENTO DELL'RSS**: il Relation Similarity Service non è in grado di mappare una rappresentazione intermedia nella corretta espressione logica (non vengono trovate relazioni adeguate).



Q&A in Natural Language: AQUALOG (8)

- **Fallimento dell’RSS**

Question
Answering

You are logged in as **DISIT** | [Log out](#) | Upload an ontology from a **local file** or an **URL**

The active ontology is **DumpTest2** [View](#) [Download](#) [Select other](#)

Ask a query

[Examples](#)

Make Use of Learning Mechanism for relations Make Use of WordNet

Query Validated ... Category WH_3TERM

Logical Representation ... Query Term - Relation - Second Term - Third Term.

Linguistic Triple:	courses	- taken	- Paolo Frasconi	- artificial intelligence
Ontology Triple:	Course	- takeCourse	- http://www.unifi.it/cercachi/scheda.php?f=s&codice=3976&fonte=dipartimento%20di%20sistemi%20e%20informatica	- artificial%20intelligence
	Course	-	- artificial%20intelligence	-

Concepts for courses	Relations for taken	Instances/classes for Paolo Frasconi
<u>Course</u>	<u>takeCourse</u> <input type="radio"/> Or <input type="radio"/> And <input checked="" type="checkbox"/> Learn	scheda.php?f=s&codice=3976&fonte=dipartimento%20di%20sistemi%20e%20informatica
Concepts for the clause	Relations	Instances for artificial intelligence
<u>Course</u>	Not relations found in the ontology between these types. <input type="radio"/> Or <input type="radio"/> And <input checked="" type="checkbox"/> Learn	artificial%20intelligence

TRY AGAIN

Link utili e Riferimenti

- [IBM Watson – 2012]

<https://www-03.ibm.com/innovation/us/watson/>

- [IOS Siri – 2010]

<http://www.apple.com/ios/siri/>

- [Winograd – 1971] Winograd, Terry (1971),
*Procedures as a Representation for Data in a Computer Program
for Understanding Natural Language*,
MAC-TR-84, MIT Project MAC, 1971.

- GATE: <http://gate.ac.uk/>

- <http://technologies.kmi.open.ac.uk/aqualog/>