

Web Crawling & Information Extraction

Gianni Pantaleo

Dipartimento di Ingegneria dell'Informazione, DINFO

Università degli Studi di Firenze

Via S. Marta 3, 50139, Firenze, Italy

Tel: +39-055-2758517

DISIT Lab

<http://www.disit.org>
gianni.pantaleo@unifi.it

MASTER: Big Data Analytics And Technologies For Management – MABIDA

Sabato 18 Novembre 2017

Outline

- 1. Sistemi di Web Crawling
 - 1.1 Introduzione
 - 1.2 Strategie di Crawling
 - 1.3 Robot Exclusion Protocol
 - 1.4 Concorrenza
- 2. Tecniche di Parsing ed Estrazione di Informazioni
 - 2.1 Introduzione
 - 2.2 NLP: Cenni Storici
 - 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
 - 2.4 NLP Tools
- 3. Applicazioni



Outline

- 1. Sistemi di Web Crawling

- 1.1 Introduzione 
- 1.2 Strategie di Crawling
- 1.3 Robot Exclusion Protocol
- 1.4 Concorrenza

- 2. Tecniche di Parsing ed Estrazione di Informazioni

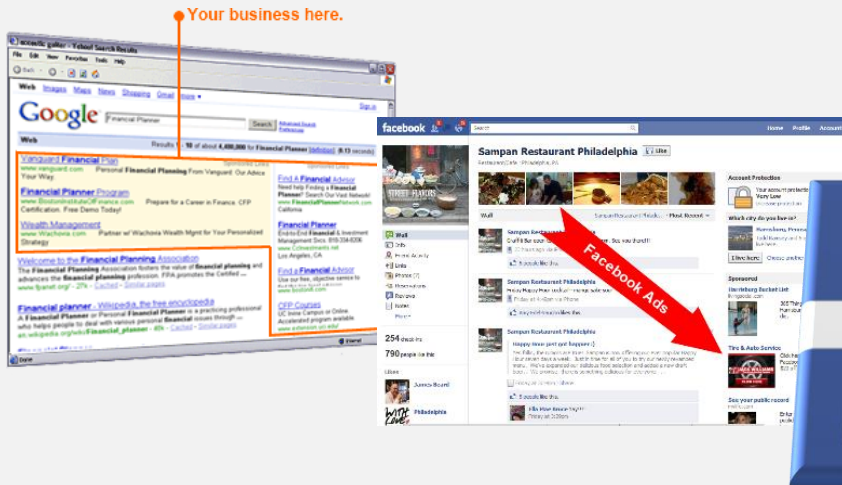
- 2.1 Introduzione
- 2.2 NLP: Cenni Storici
- 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
- 2.4 NLP Tools

- 3. Applicazioni



1.1 - Introduzione: Dati e Informazione sul Web

- **The Zettabyte Era (1 ZB = 10^{12} GB):** Secondo studi e stime di **CISCO**, la quantità di dati prodotti ed archiviati nel cloud data center globale sarà pari nel **2020**, a **1.8 Zettabytes** (1.8 mila miliardi di Gigabytes), equivalente alla quantità di spazio archiviabile in circa **45 milioni di DVD all'ora**, con una crescita di 5 volte rispetto al 2015 (*).



- La quantità di traffico dati nel web globale (*global data center traffic*), arriverà a **15.3 Zettabyte nel 2020 (**)**.

* [Cisco Global Cloud Index: Forecast and Methodology, 2015–2020 White Paper](#) (Cisco Public Knowledge)

** <http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html>
(Source: Global Cloud Index Infographic [GCI 2016](#))

1.1 - Introduzione: Generalità di un Web Crawler

Definizione di **Web Crawler**

- Un **Crawler Web** è un software usato per collezionare ed archiviare il contenuto di pagine web, documenti presenti in una rete o in un database



- Definito anche **Web Spider**, **Web Robot**, o semplicemente **Bot**

1.1 - Introduzione: Aree Applicative

Un **Web Crawler** può essere usato per:

- Creare un indice generale (**general Web search**)
- Creare indici di topics o argomenti specifici (**vertical Web search**)
- Archiviare il contenuto di pagine web e documenti (**Web archival**)
- Analizzare il contenuto di pagine e siti web per produrre statistiche e analisi aggregate (**Web characterization**)
- Tenere copie o repliche di siti web (**Web mirroring**)
- Analisi di siti web (**Web site analysis**)



1.1 - Introduzione: Processo di Crawling



1.1 - Introduzione: Principali Search Bot

Bot Name	% of Sites Crawled	Bot Type
Googlebot	96%	Search Bot
Baidu Spider	89%	Search Bot
MSN Bot/BingBot	82%	Search Bot
Yandex Bot	73%	Search Bot
Soso Spider	61%	Search Bot
ExaBot	35%	Search Bot
Sogou Spider	31%	Search Bot
Google Plus Share	24%	Crawler
Facebook External Hit	24%	Crawler
Google Feedfetcher	22%	Feed Fetcher

Fonte: <https://www.incapsula.com/blog/know-your-top-10-bots.html>

1.1 - Introduzione: Principali Crawler Open Source



<https://webarchive.jira.com/wiki/display/Heritrix>



<http://nutch.apache.org/>

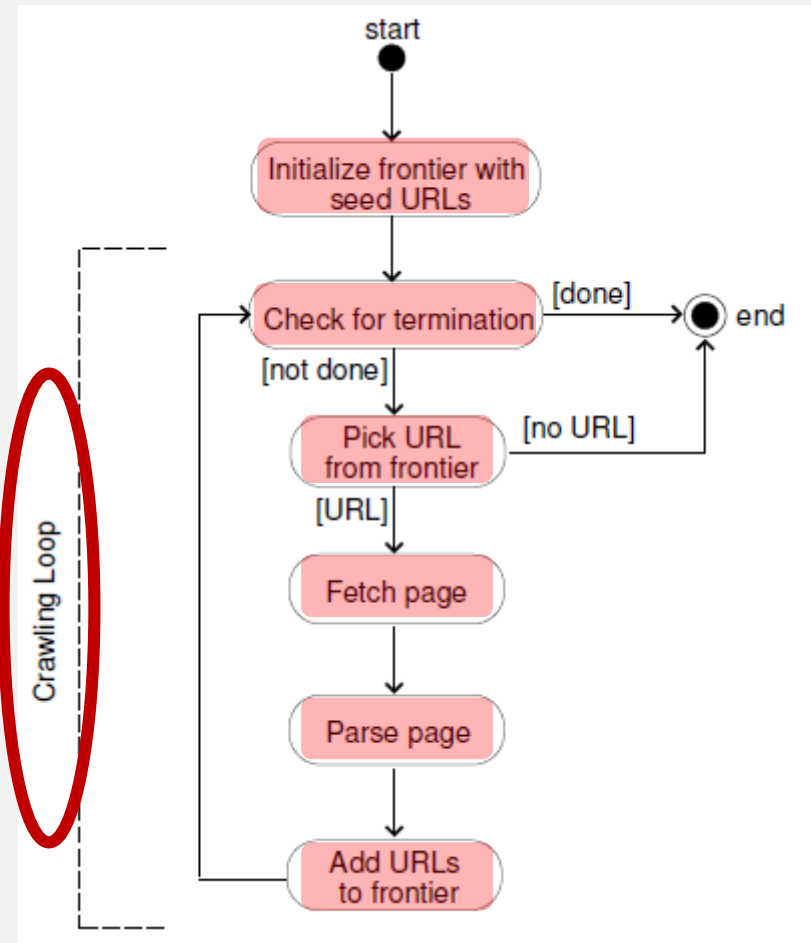


<https://www.cs.cmu.edu/~rcm/websphinx/>

1.1 - Introduzione: Principi di Funzionamento



- Il web è visto come un grafo: i nodi sono le pagine web e gli archi sono i link (*hyperlinks*)
- E' utilizzato per archiviare le pagine web visitate per una loro successiva elaborazione (estrazione di informazioni, indicizzazione, ecc..)
- Analizza le pagine partendo da una lista di indirizzi web (**URL**) iniziali. Gli URL sono contenuti in una struttura dati che si chiama **frontiera**




1.1 - Introduzione: La Frontiera di un Crawler

- E' una struttura dati dinamica che contiene URL non ancora visitati
- 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

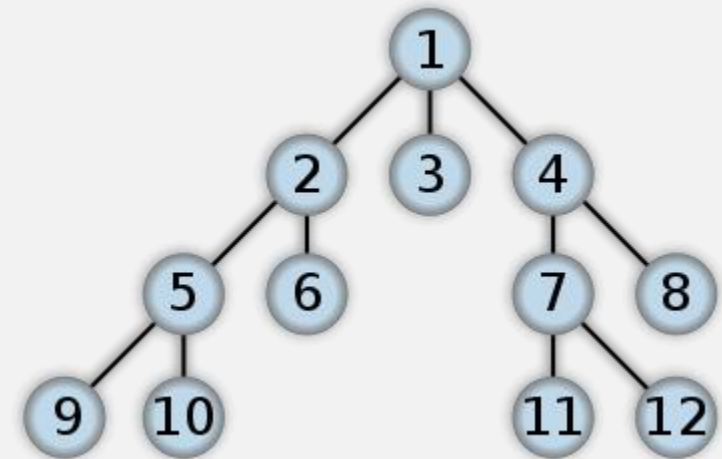
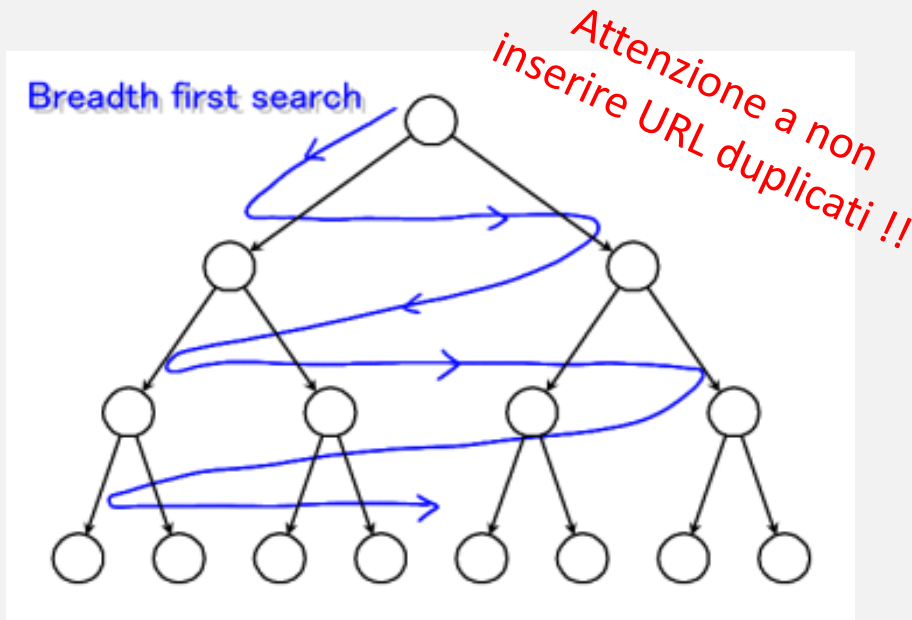
Outline

- 1. Sistemi di Web Crawling
 - 1.1 Introduzione
 - 1.2 Strategie di Crawling 
 - 1.3 Robot Exclusion Protocol
 - 1.4 Concorrenza
- 2. Tecniche di Parsing ed Estrazione di Informazioni
 - 2.1 Introduzione
 - 2.2 NLP: Cenni Storici
 - 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
 - 2.4 NLP Tools
- 3. Applicazioni

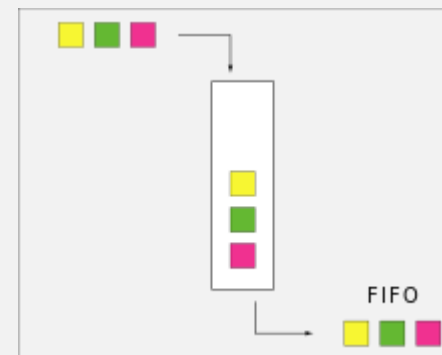


1.2 - Strategie di Crawling: Breadth First Search

➤ Breadth First Search

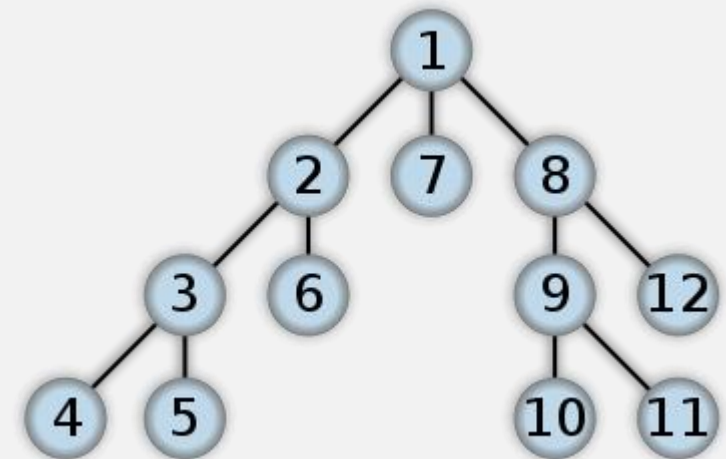
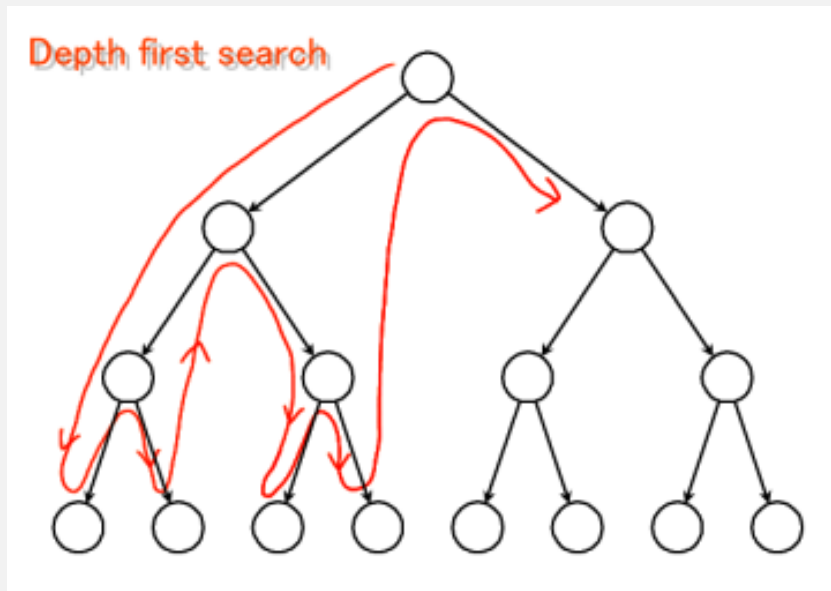


- Struttura dati lineare con politica
FIFO : First In First Out (CODA)



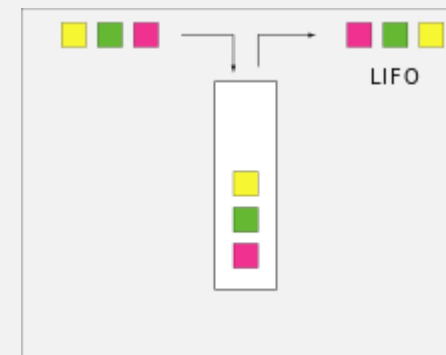
1.2 - Strategie di Crawling: Depth First Search

➤ Depth First Search



- Struttura dati lineare
con politica

LIFO: Last In First Out (PILA o STACK)



1.2 - Strategie di Crawling

➤ 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



1.2 - Strategie di Crawling

- Evitare il fetch della stessa pagina:
 - Tenere in memoria una lista delle pagine visitate

- La dimensione della frontiera cresce velocemente
 - Può essere necessaria una politica di priorità sugli URLs

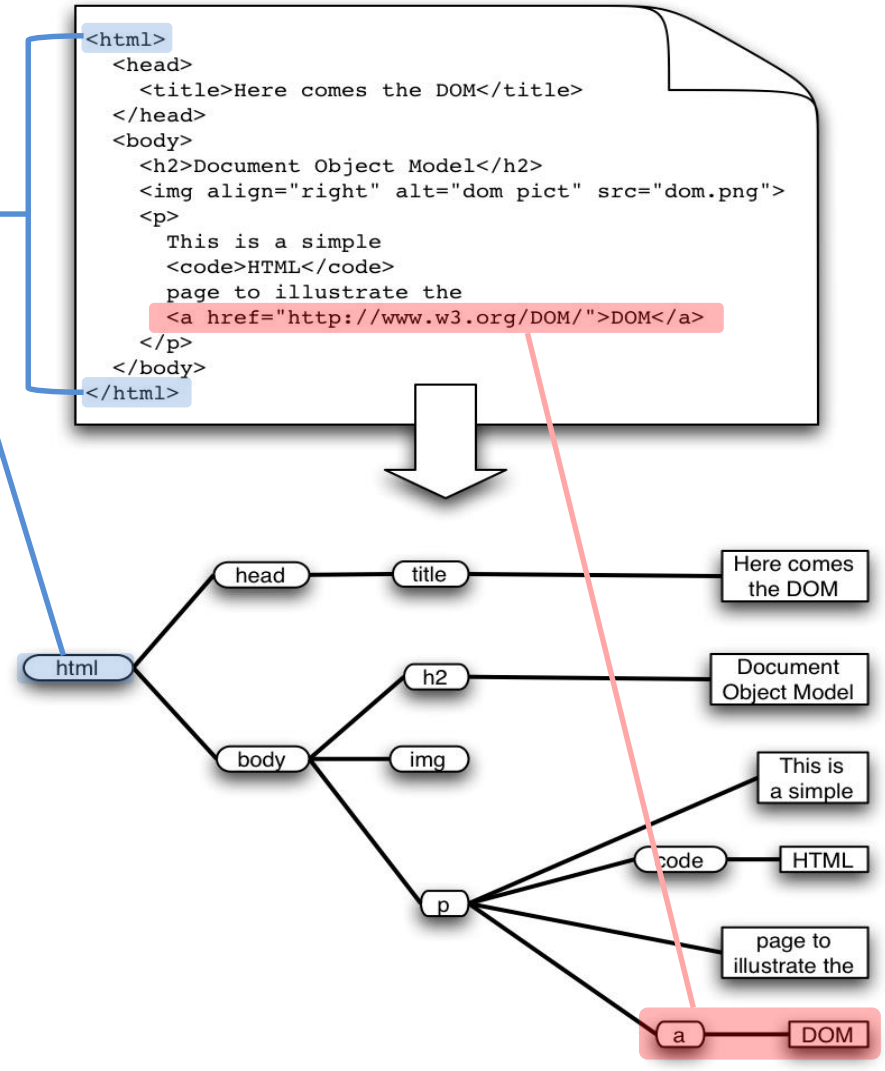
- Fetcher deve essere robusto
 - Evitare crash in caso di fallimenti nei downloads
 - Prevedere meccanismi di timeout

- Determinazione dei file non desiderati
 - Esaminare estensioni dei file (filtrando formati come multimedia, immagini, video ecc.)
 - Content-Type (MIME) headers



1.2 - Strategie di Crawling

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



Outline

- 1. Sistemi di Web Crawling

- 1.1 Introduzione
- 1.2 Strategie di Crawling
- 1.3 Robot Exclusion Protocol
- 1.4 Concorrenza



- 2. Tecniche di Parsing ed Estrazione di Informazioni

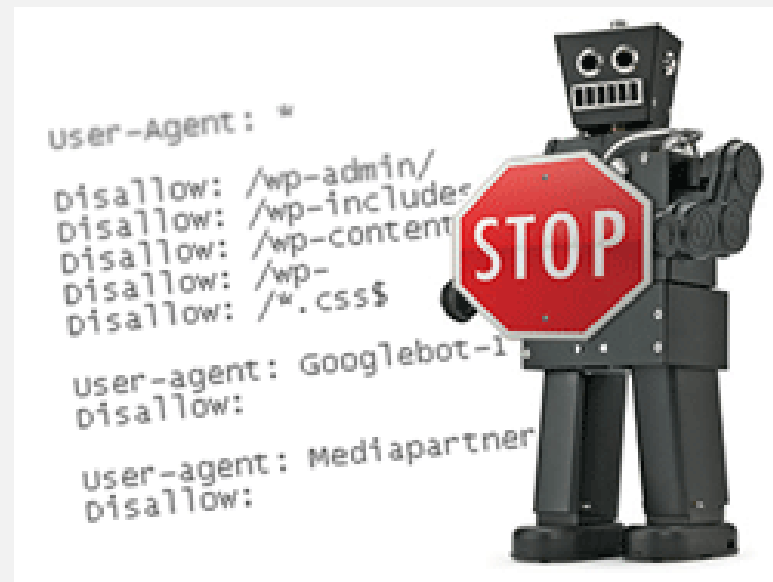
- 2.1 Introduzione
- 2.2 NLP: Cenni Storici
- 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
- 2.4 NLP Tools

- 3. Applicazioni



1.3 Robot Exclusion Protocol

- 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



1.3 Robot Exclusion Protocol

www.apple.com/robots.txt

robots.txt for <http://www.apple.com/>

User-agent: *

Disallow:

Tutti i crawlers ...

...possono analizzare qualsiasi
pagina!

1.3 Robot Exclusion Protocol

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 percorsi del
grafo del sito web

1.3 Robot Exclusion Protocol

www.springer.com/robots.txt

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

User-agent: Googlebot

Disallow: /chl/*

Disallow: /uk/*

Disallow: /italy/*

Disallow: /france/*

Google crawler può analizzare tutte le pagine eccetto queste

User-agent: slurp

Disallow:

Crawl-delay: 2

User-agent: MSNBot

Disallow:

Crawl-delay: 2

Yahoo e MSN/Windows Live possono analizzare tutte le pagine ma il processo deve essere "lento" (2 secondi)

User-agent: scooter

Disallow:

AltaVista non ha limiti


all others

User-agent: *

Disallow: /

A tutti gli altri crawlers non è permesso niente

Outline

- 1. Sistemi di Web Crawling
 - 1.1 Introduzione
 - 1.2 Strategie di Crawling
 - 1.3 Robot Exclusion Protocol
 - 1.4 Concorrenza 
- 2. Tecniche di Parsing ed Estrazione di Informazioni
 - 2.1 Introduzione
 - 2.2 NLP: Cenni Storici
 - 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
 - 2.4 NLP Tools
- 3. Applicazioni



1.4 - Concorrenza

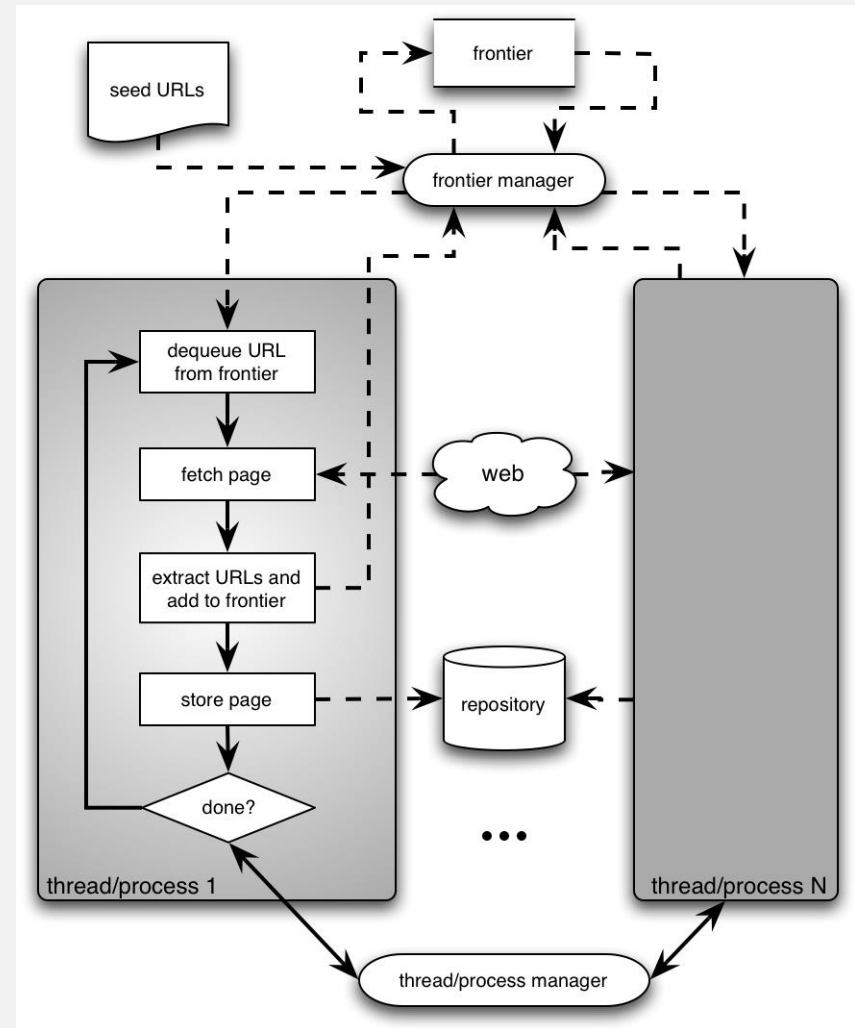
- Le operazioni effettuate dal crawler possono richiedere diverso tempo:
 - Risoluzione indirizzi IP
 - Connessioni al server e invio risposte
 - Ricezione pagina di risposta

- Soluzione: Ridurre i suddetti tempi eseguendo la scansione di più pagine in maniera concorrente




1.4 - Concorrenza

- Ogni thread lavora come un crawler sequenziale e condivide le strutture dati: frontiera e repository (concorrenza in lettura)
- Le strutture dati condivise devono essere sincronizzate (concorrenza in scrittura)



Outline

- 1. Sistemi di Web Crawling
 - 1.1 Introduzione
 - 1.2 Strategie di Crawling
 - 1.3 Robot Exclusion Protocol
 - 1.4 Concorrenza
- 2. Tecniche di Parsing ed Estrazione di Informazioni
 - 2.1 Introduzione 
 - 2.2 NLP: Cenni Storici
 - 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
 - 2.4 NLP Tools

3. Applicazioni



2.1 - Tecniche di Parsing: Introduzione

➤ Natural Language Processing

➤ Scenario / Requisiti

- ❖ Dotare l'IA delle abilità linguistiche proprie dell'essere umano
- ❖ Comprensione e generazione di testo non strutturato (*linguaggio naturale*)
- ❖ 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



2.1 - Tecniche di Parsing: Introduzione


➤ Scenario / Requisiti

- ❖ I linguaggi sono purtroppo ambigui.
- ❖ Le ambiguità si possono avere a 4 livelli:
 - ✓ Ambiguità lessicale: «*attacco*» (verbo, sostantivo)
 - ✓ Ambiguità strutturale: «*Ieri 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 automatica del
linguaggio naturale un task molto complesso !**



Outline

- 1. Sistemi di Web Crawling
 - 1.1 Introduzione
 - 1.2 Strategie di Crawling
 - 1.3 Robot Exclusion Protocol
 - 1.4 Concorrenza
- 2. Tecniche di Parsing ed Estrazione di Informazioni
 - 2.1 Introduzione
 - 2.2 NLP: Cenni Storici 
 - 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
 - 2.4 NLP Tools

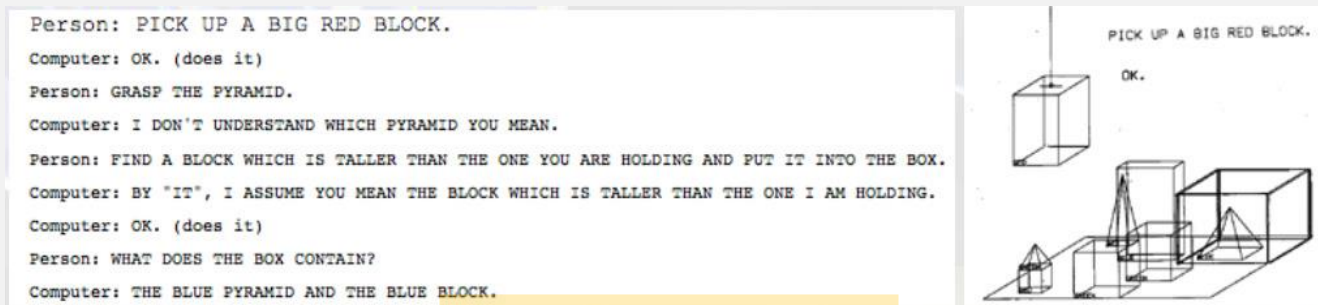
3. Applicazioni



2.2 - Tecniche di Parsing – NLP: Cenni Storici

➤ **SHRDLU** [Winograd - 1971]

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



➤ **Apple Siri** [iOS Siri - 2010]

- ❖ Virtual personal assistant
- ❖ Knowledge navigator
- ❖ User recommendation system.



➤ **Google Assistant**

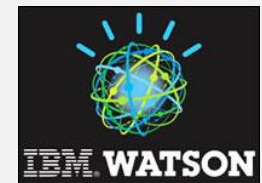


➤ **Microsoft Cortana**



➤ **IBM Watson** [IBM Watson - 2012]

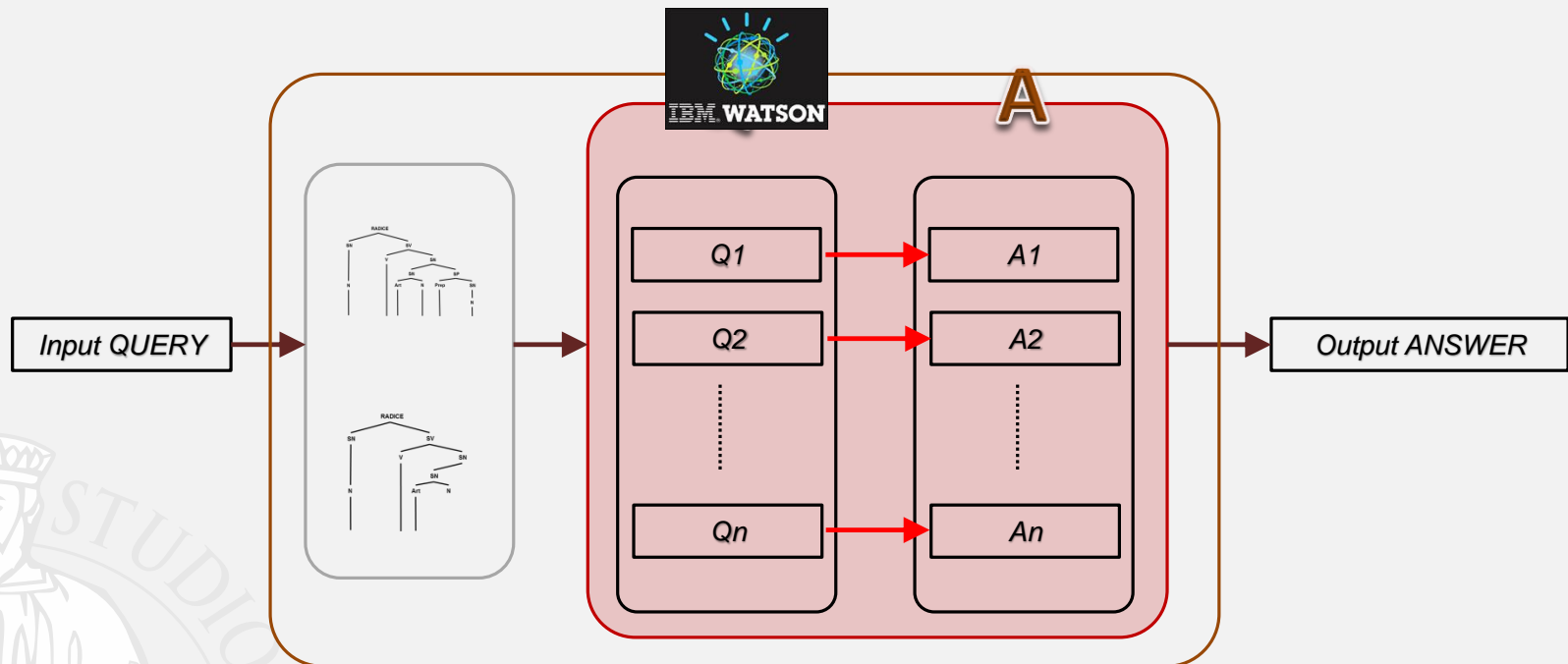
- ❖ Nato come sistema Question & Answer per il quiz televisivo americano *Jeopardy!*, successivamente è stato sviluppato come *cognitive system* completo con meccanismi di auto-apprendimento.



2.2 - Tecniche di Parsing – NLP: Cenni Storici

➤ **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.



Outline


- 1. Sistemi di Web Crawling
 - 1.1 Introduzione
 - 1.2 Strategie di Crawling
 - 1.3 Robot Exclusion Protocol
 - 1.4 Concorrenza
- 2. Tecniche di Parsing ed Estrazione di Informazioni
 - 2.1 Introduzione
 - 2.2 NLP: Cenni Storici
 - 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
 - 2.4 NLP Tools




3. Applicazioni



2.3 – Fasi di Elaborazione in Linguaggio Naturale



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.

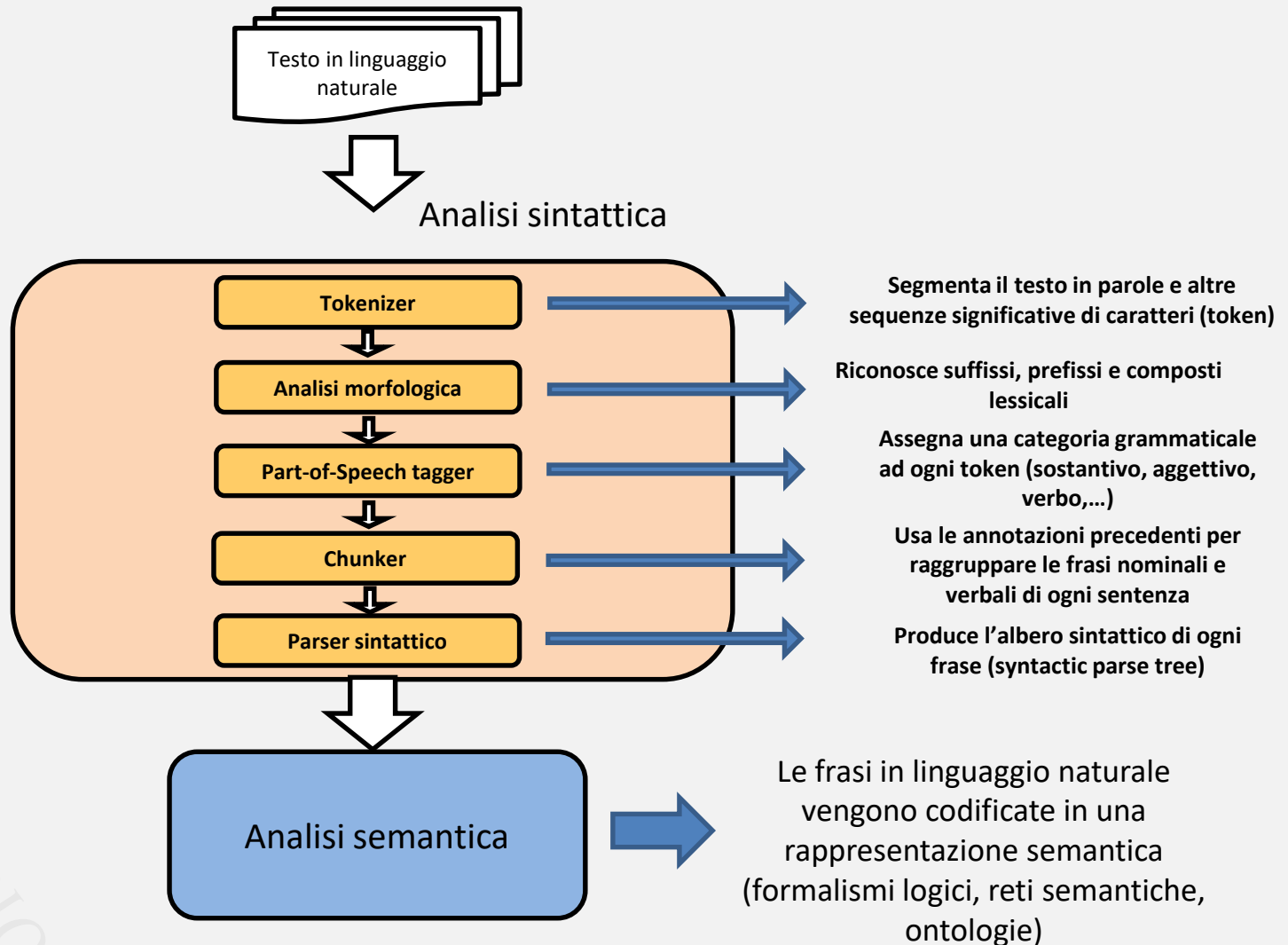
Discourse integration: il significato di una frase spesso dipende dalla frase che la precede e può influenzare quello della frase che la segue.

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

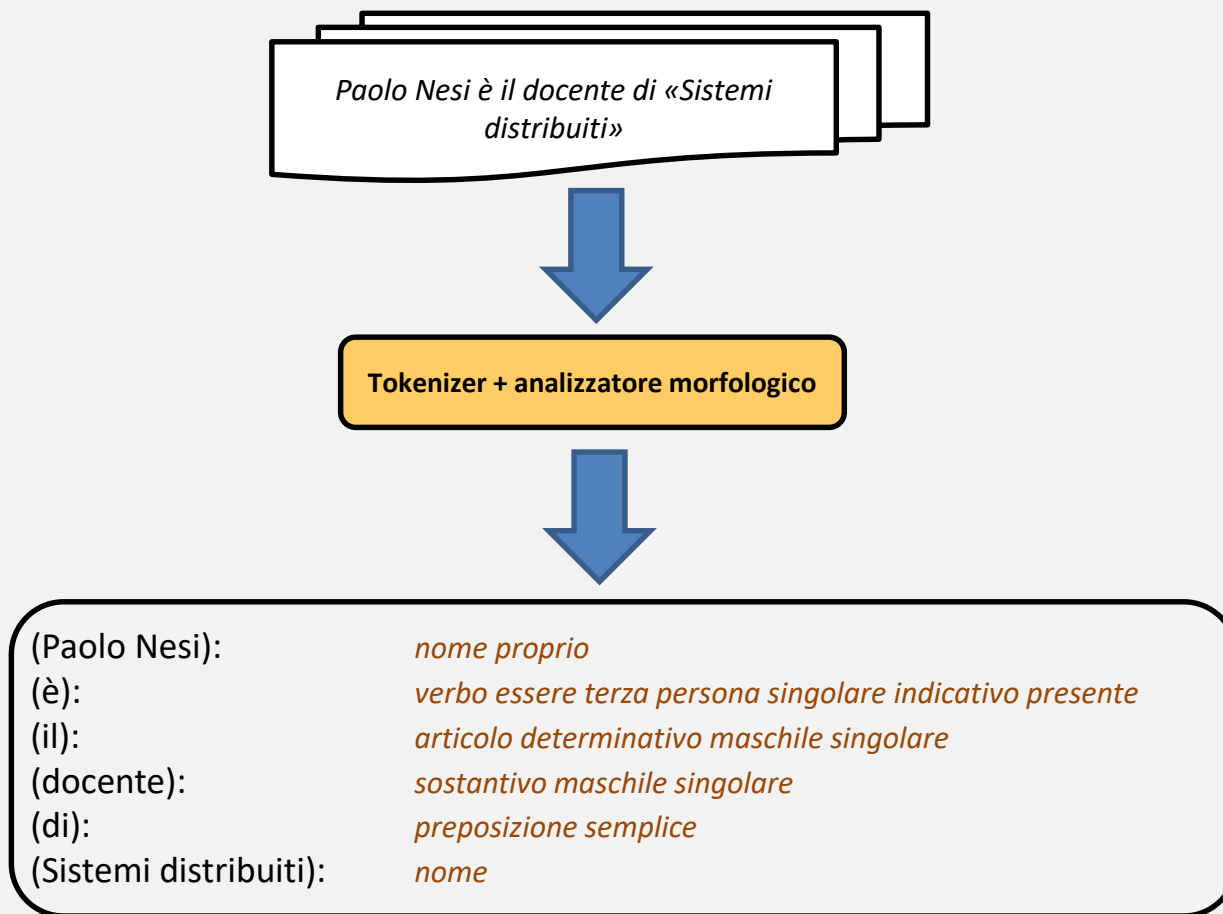
“la porta è aperta” necessita di conoscere quale è stata l'intenzione dell'interlocutore:

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

2.3 – Fasi di Elaborazione in Linguaggio Naturale



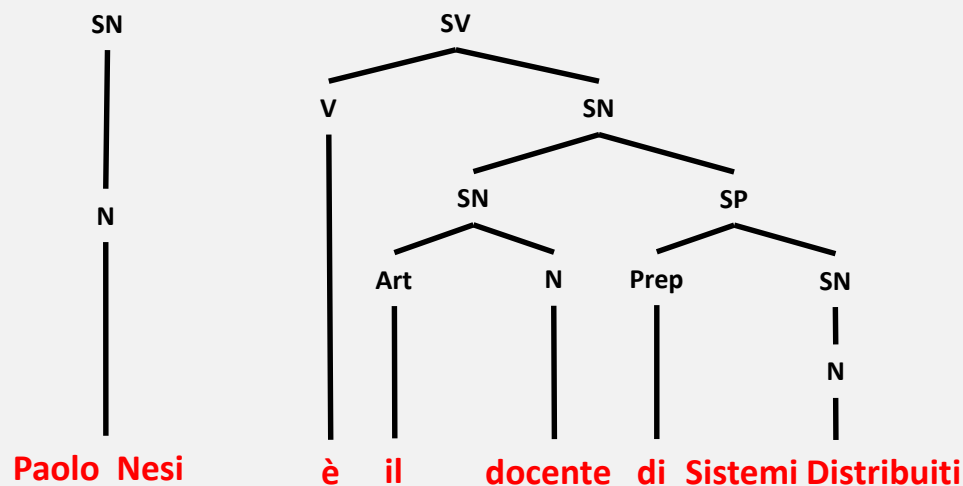
2.3 – Fasi di Elaborazione in Linguaggio Naturale



(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>

Part-of-Speech + chunker +
parser sintattico

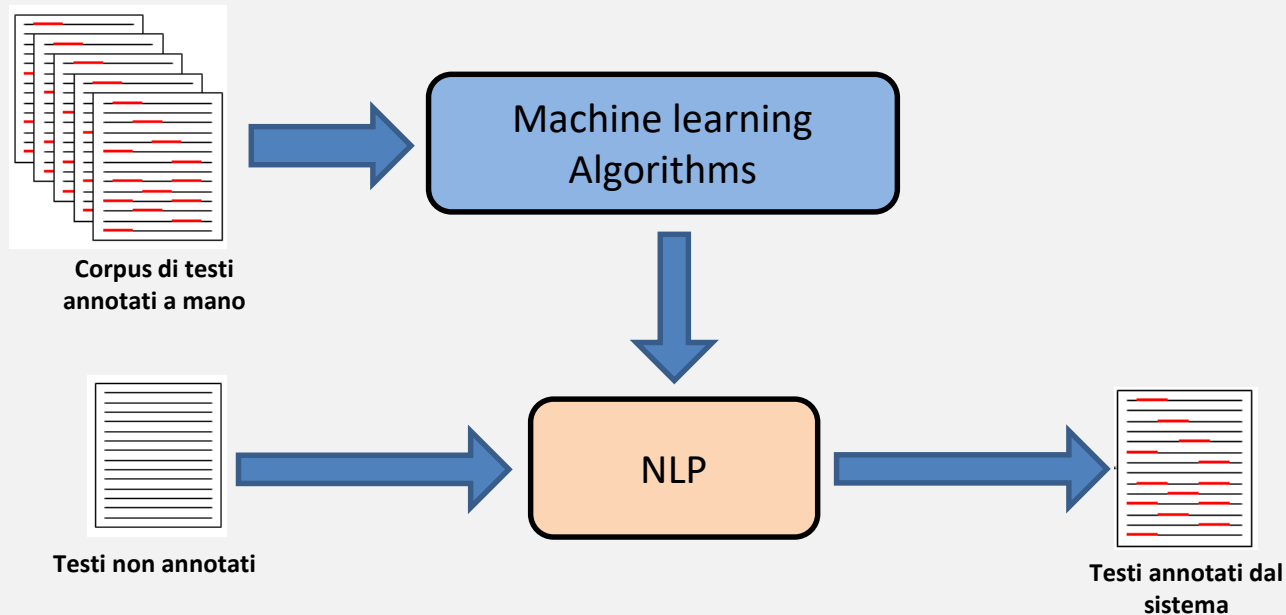
RADICE




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

2.3 – Fasi di Elaborazione in Linguaggio Naturale

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



Outline

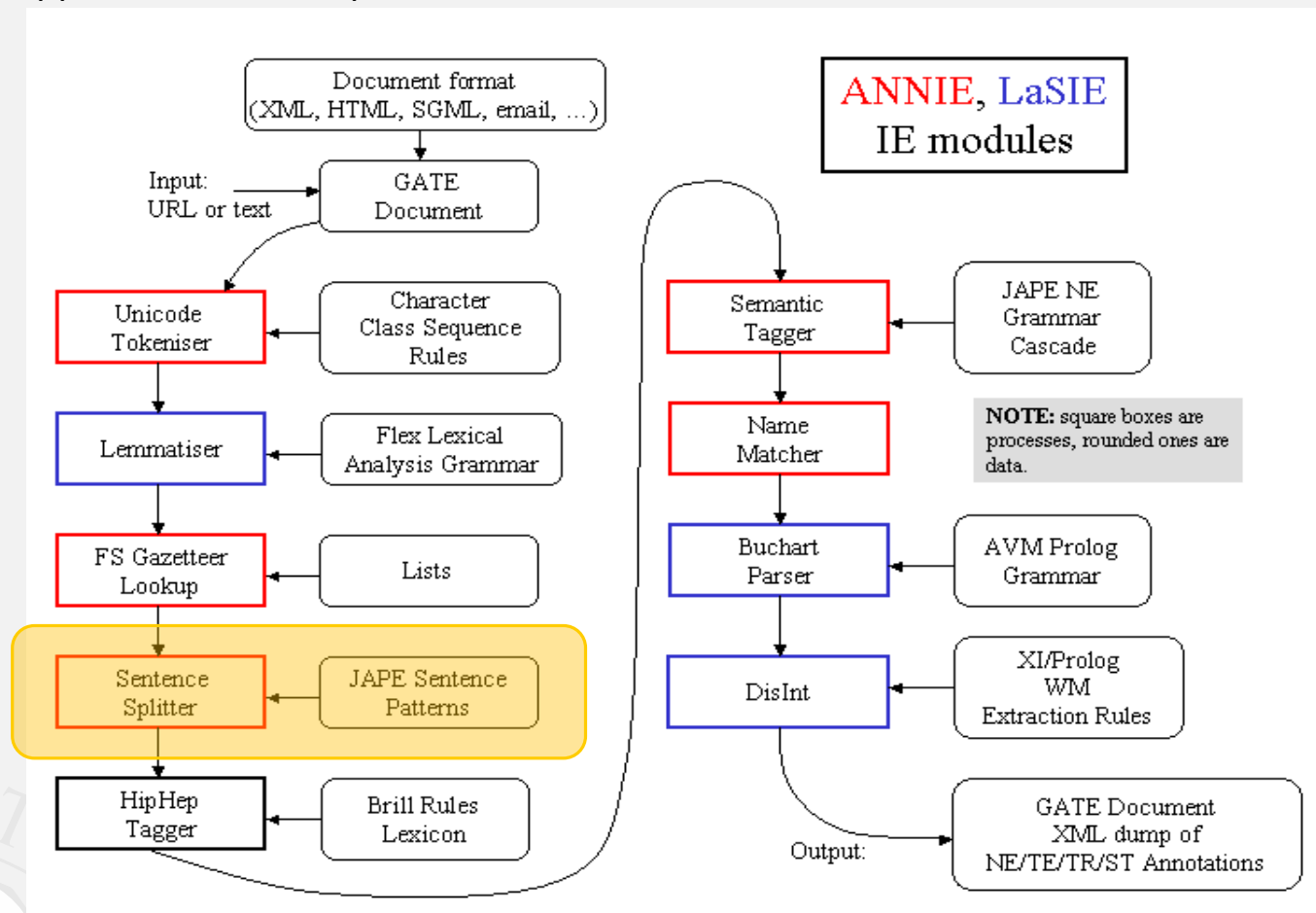
- 1. Sistemi di Web Crawling
 - 1.1 Introduzione
 - 1.2 Strategie di Crawling
 - 1.3 Robot Exclusion Protocol
 - 1.4 Concorrenza
- 2. Tecniche di Parsing ed Estrazione di Informazioni
 - 2.1 Introduzione
 - 2.2 NLP: Cenni Storici
 - 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
 - 2.4 NLP Tools 

3. Applicazioni



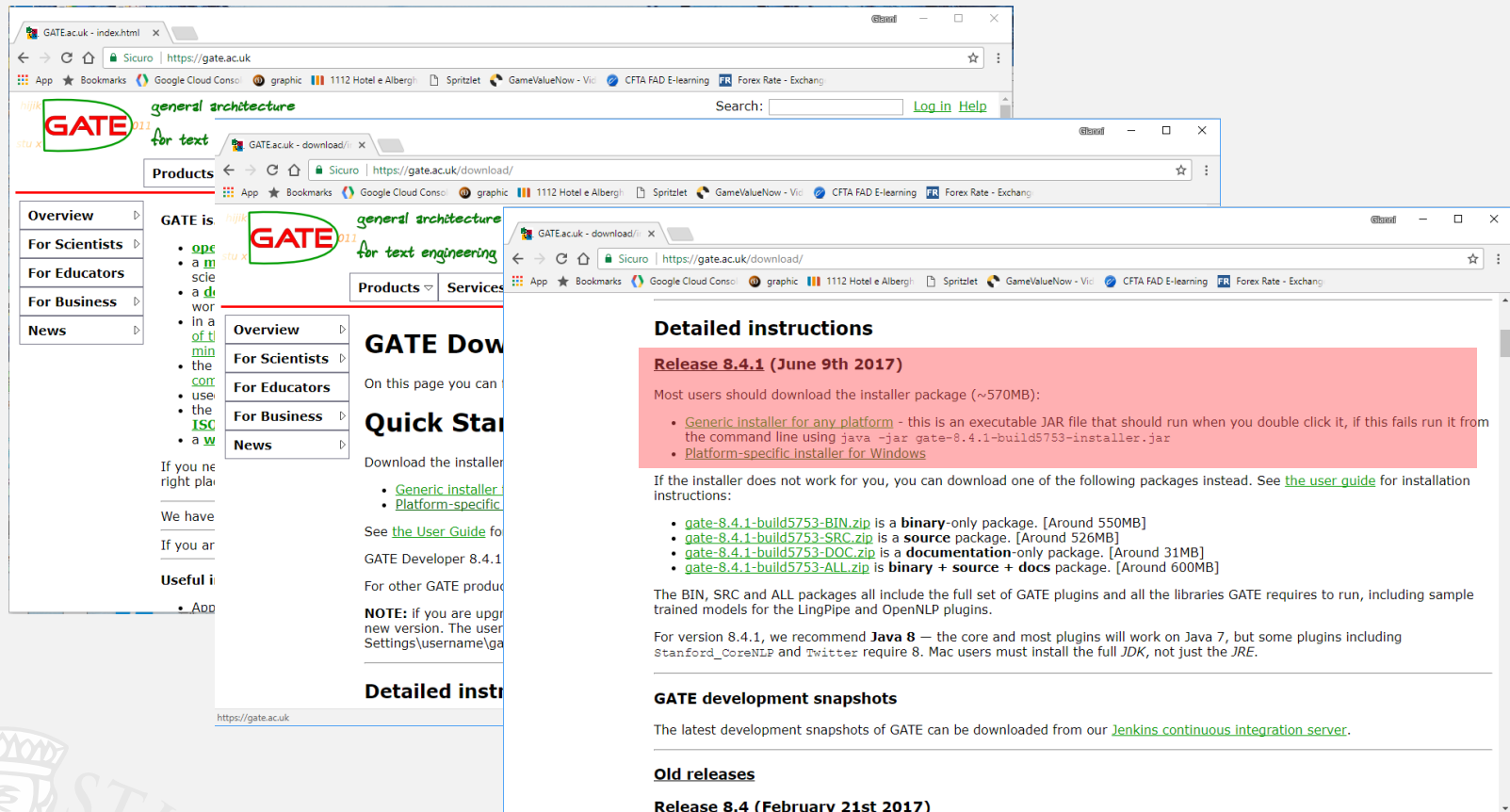
2.4 - NLP Tools: GATE

- **GATE – General Architecture for Text Engineering** (<https://gate.ac.uk/>)
 - ❖ Supporta documenti plain text, HTML, XML ...



2.4 - NLP Tools: GATE

- **GATE – General Architecture for Text Engineering** (<https://gate.ac.uk/>)



The screenshot displays the GATE website's download page. The left sidebar contains navigation links: Overview, For Scientists, For Educators, For Business, and News. The main content area is titled "GATE Download" and includes a "Quick Start" section with links to "Generic installer" and "Platform-specific". Below this, it mentions "GATE Developer 8.4.1" and provides a "NOTE" about upgrading. The right sidebar, titled "Detailed instructions", highlights "Release 8.4.1 (June 9th 2017)" and lists the recommended download package: "Generic installer for any platform". It also provides instructions for alternative packages (BIN, SRC, DOC, ALL) and mentions the requirement for Java 8. The bottom of the page includes sections for "GATE development snapshots" and "Old releases".

Detailed instructions

Release 8.4.1 (June 9th 2017)

Most users should download the installer package (~570MB):

- [Generic installer for any platform](#) - this is an executable JAR file that should run when you double click it, if this fails run it from the command line using `java -jar gate-8.4.1-build5753-installer.jar`
- [Platform-specific installer for Windows](#)

If the installer does not work for you, you can download one of the following packages instead. See [the user guide](#) for installation instructions:

- [gate-8.4.1-build5753-BIN.zip](#) is a **binary**-only package. [Around 550MB]
- [gate-8.4.1-build5753-SRC.zip](#) is a **source** package. [Around 526MB]
- [gate-8.4.1-build5753-DOC.zip](#) is a **documentation**-only package. [Around 31MB]
- [gate-8.4.1-build5753-ALL.zip](#) is **binary + source + docs** package. [Around 600MB]

The BIN, SRC and ALL packages all include the full set of GATE plugins and all the libraries GATE requires to run, including sample trained models for the LingPipe and OpenNLP plugins.

For version 8.4.1, we recommend **Java 8** — the core and most plugins will work on Java 7, but some plugins including `Stanford_CoreNLP` and `Twitter` require 8. Mac users must install the full `JDK`, not just the `JRE`.

GATE development snapshots

The latest development snapshots of GATE can be downloaded from our [Jenkins continuous integration server](#).

Old releases

Release 8.4 (February 21st 2017)

2.4 - NLP Tools: GATE

➤ ANNIE (A Nearly New IE System)

- ❖ Utilizza regole per il **Part of Speech (POS) Tagging**. Tali regole sono predefinite (tag hmtl) 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 hmtl) 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

2.4 - NLP Tools: GATE

➤ JAPE (Java Annotations Pattern Engine)

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

❖ Es. di sintassi di una regola JAPE:

```
Phase: Address_Retrieval
Input: Token Lookup
Options: control = appelt
```

Headers della regola

```
Rule: FindStreetAddress
Priority: 20
```

Nome della regola e priorità

```
(
  ({Token.string == "Via"} | {Token.string == "Piazza"} | {Token.string == "Largo"})
  ( ({Lookup.majorType == NomeProprio} | {Lookup.majorType == Cognome}) |
    ({Lookup.majorType == Cognome}) )
)
```

```
:address
```

Label

```
-->
```

RHS

LHS

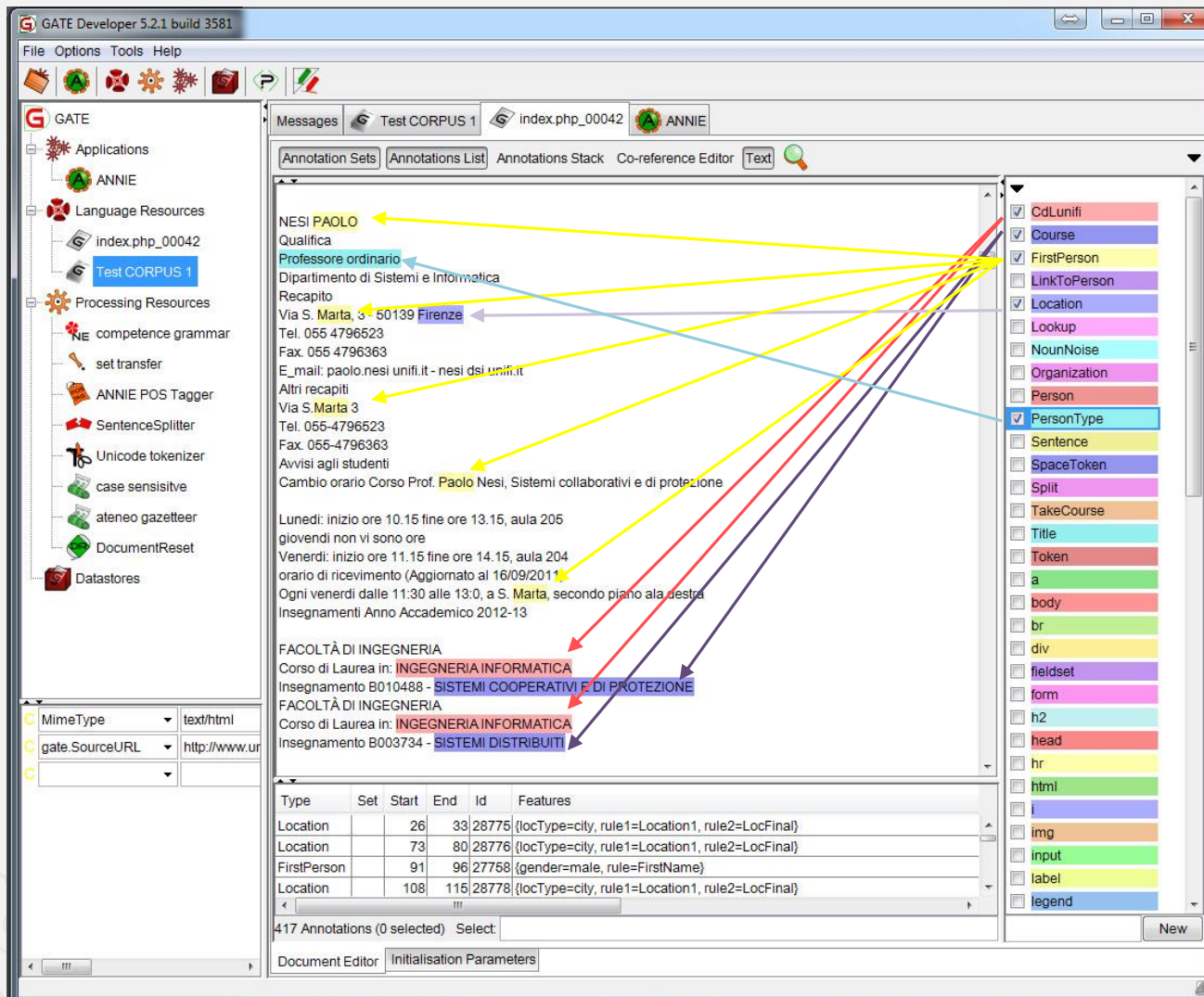
```
: address.indirizzo = {rule = "FindStreetAddress"}
```



2.4 - NLP Tools: GATE

[illegible]

2.4 - NLP Tools: GATE



The screenshot shows the GATE Developer 5.2.1 build 3581 interface. The main window displays a text document with various annotations. The left sidebar shows the project structure, including Applications, Language Resources, and Processing Resources. The right sidebar shows a list of annotation sets, with several checked, including CdLunifi, Course, FirstPerson, LinkToPerson, Location, Lookup, NounNoise, Organization, Person, PersonType, Sentence, SpaceToken, Split, TakeCourse, Title, Token, a, body, br, div, fieldset, form, h2, head, hr, html, i, img, input, label, and legend.

The main text area contains the following content:

NESI PAOLO
Qualifica
Professore ordinario
Dipartimento di Sistemi e Informatica
Recapito
Via S. Marta, 3 - 50139 Firenze
Tel. 055 4796523
Fax. 055 4796363
E_mail: paolo.nesi.unifi.it - nesi.dsi.unifi.it
Altri recapiti
Via S. Marta 3
Tel. 055-4796523
Fax. 055-4796363
Avvisi agli studenti
Cambio orario Corso Prof. Paolo Nesi, Sistemi collaborativi e di protezione

Lunedì: inizio ore 10.15 fine ore 13.15, aula 205
giovedì non vi sono ore
Venerdì: inizio ore 11.15 fine ore 14.15, aula 204
orario di ricevimento (Aggiornato al 16/09/2011)
Ogni venerdì dalle 11:30 alle 13.0, a S. Marta, secondo piano ala destra
Insegnamenti Anno Accademico 2012-13

FACOLTÀ DI INGEGNERIA
Corso di Laurea in: INGEGNERIA INFORMATICA
Insegnamento B010488 - SISTEMI COOPERATIVI E DI PROTEZIONE
FACOLTÀ DI INGEGNERIA
Corso di Laurea in: INGEGNERIA INFORMATICA
Insegnamento B003734 - SISTEMI DISTRIBUITI

The bottom section 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}

417 Annotations (0 selected) Select: [New]

Document Editor Initialisation Parameters

Outline

- 1. Sistemi di Web Crawling
 - 1.1 Introduzione
 - 1.2 Strategie di Crawling
 - 1.3 Robot Exclusion Protocol
 - 1.4 Concorrenza
- 2. Tecniche di Parsing ed Estrazione di Informazioni
 - 2.1 Introduzione
 - 2.2 NLP: Cenni Storici
 - 2.3 Fasi dell'Elaborazione in Linguaggio Naturale
 - 2.4 NLP Tools
- 3. Applicazioni ←

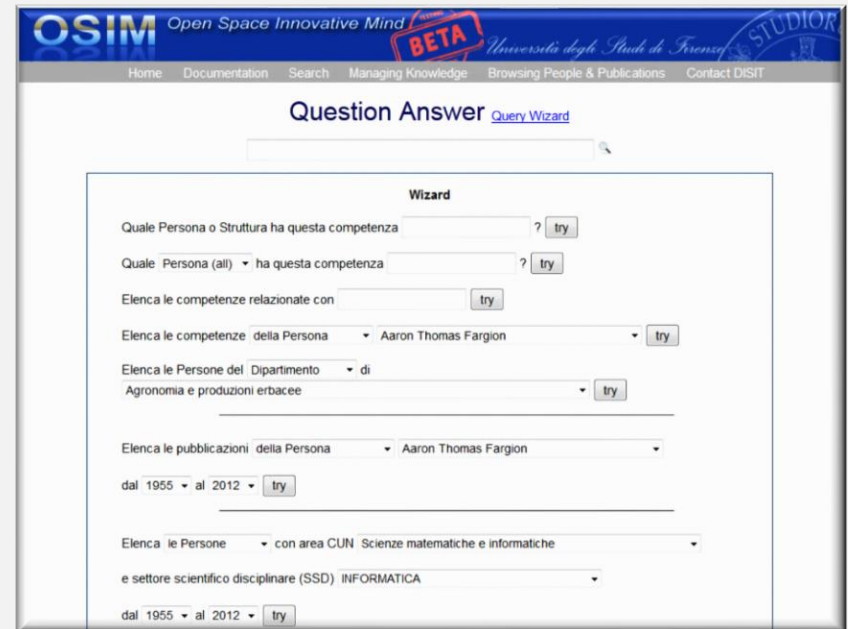


3 - Applicazioni: OSIM (Open Space Innovative Mind)

- **Open Space Innovative Mind** (<http://openmind.disit.org>) [Bellandi et al., 2012] è un progetto per la realizzazione di un portale nel quale industrie, istituti di ricerca, 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 presentate:

- ❖ **Data Mining**
- ❖ **Web Crawling**
- ❖ **NLP**
- ❖ **Semantic Web**
- ❖ **Interrogazione della conoscenza e Reasoning (OSIM Query Wizard)**



The screenshot shows the OSIM Query Wizard interface. At the top, there's a navigation bar with links: Home, Documentation, Search, Managing Knowledge, Browsing People & Publications, and Contact DISIT. Below this is a 'Question Answer' section with a 'Query Wizard' link. The main area is a form titled 'Wizard' with several input fields and 'try' buttons. The fields include: 'Quale Persona o Struttura ha questa competenza', 'Quale Persona (all) ha questa competenza', 'Elenca le competenze relazionate con', 'Elenca le competenze della Persona' (with a dropdown menu showing 'Aaron Thomas Fargion'), 'Elenca le Persone del Dipartimento di' (with a dropdown menu showing 'Agronomia e produzioni erbacee'), 'Elenca le pubblicazioni della Persona' (with a dropdown menu showing 'Aaron Thomas Fargion'), 'dal 1955 al 2012', 'Elenca le Persone con area CUN Scienze matematiche e informatiche', and 'e settore scientifico disciplinare (SSD) INFORMATICA'. Each field has a 'try' button next to it.

3 - Applicazioni: OSIM (Open Space Innovative Mind)

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 o ricercatori 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 esempio, alla ricerca di chi possieda una specifica competenza, un docente/ricercatore può avere più rilevanza se possiede una competenza più specifica rispetto a quella cercata, di un docente/ricercatore che possiede una competenza più generale rispetto a quella cercata

3 - Applicazioni: OSIM (Open Space Innovative Mind)



ha cercato nell'area di INGEGNERIA

Cerca

Ricerca base

Forse cercavi: sistemi distribuiti

Dottorato in Informatica, Sistemi
... Il corso punterà a completare la formazione in progettazione ed analisi. ... Sicurezza
www.unifi.it/dist/cmpro-v-p-6.html - 23

Dottorato in Informatica, Sistemi
... Sono di interesse i sistemi di gestione software, i sistemi distribuiti, il software
www.unifi.it/dist/cmpro-v-p-16.html - 2

Dipartimento di Energetica - Laboratorio IBIS
... F., Morin F., Miglioramento delle prestazioni tramite sistemi
Manutenzione ... della catena logistica tramite simulazione
www.ibis.unifi.it/cmpro-l-s-4.html - 32k - 2009-05-19

Informatica e Applicazioni

Referente: Rosario Pugliese

Obiettivi:

Scopo del Dottorato è la formazione verso gli aspetti applicativi. Questo (che debbono fare i conti con frequenze soprattutto ad allargare la base culturale lavorative non collegate alla ricerca. Il corso punterà a completare la formazione.

- Algoritmi per sistemi distribuiti
- Elaborazione delle immagini
- Metodi formali per la specificazione
- Progettazione di algoritmi distribuiti
- Progettazione ed analisi di sistemi
- Sicurezza di sistemi distribuiti
- Strumenti formali per l'analisi
- Trattamento numerico e modellazione

Visto l'esiguo numero di dottorandi interessati e che garantisce gli obiettivi.

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



Condition Monitoring & 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 (previsione vita utile residua del bene).

Service Management and Engineering

Analisi del valore, sviluppo service concept, ingegnerizzazione servizi, due diligence tecnologica, progettazione 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.

Primo risultato

Secondo risultato

3 - Applicazioni: OSIM (Open Space Innovative Mind)



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

Home Documentation Search Managing Knowledge Browsing People & Publications Contact DISIT

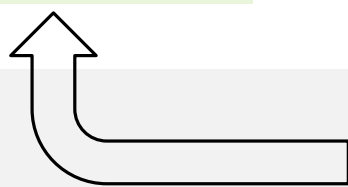
Question Answer [Query Wizard](#)

sistemi distribuiti

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

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			

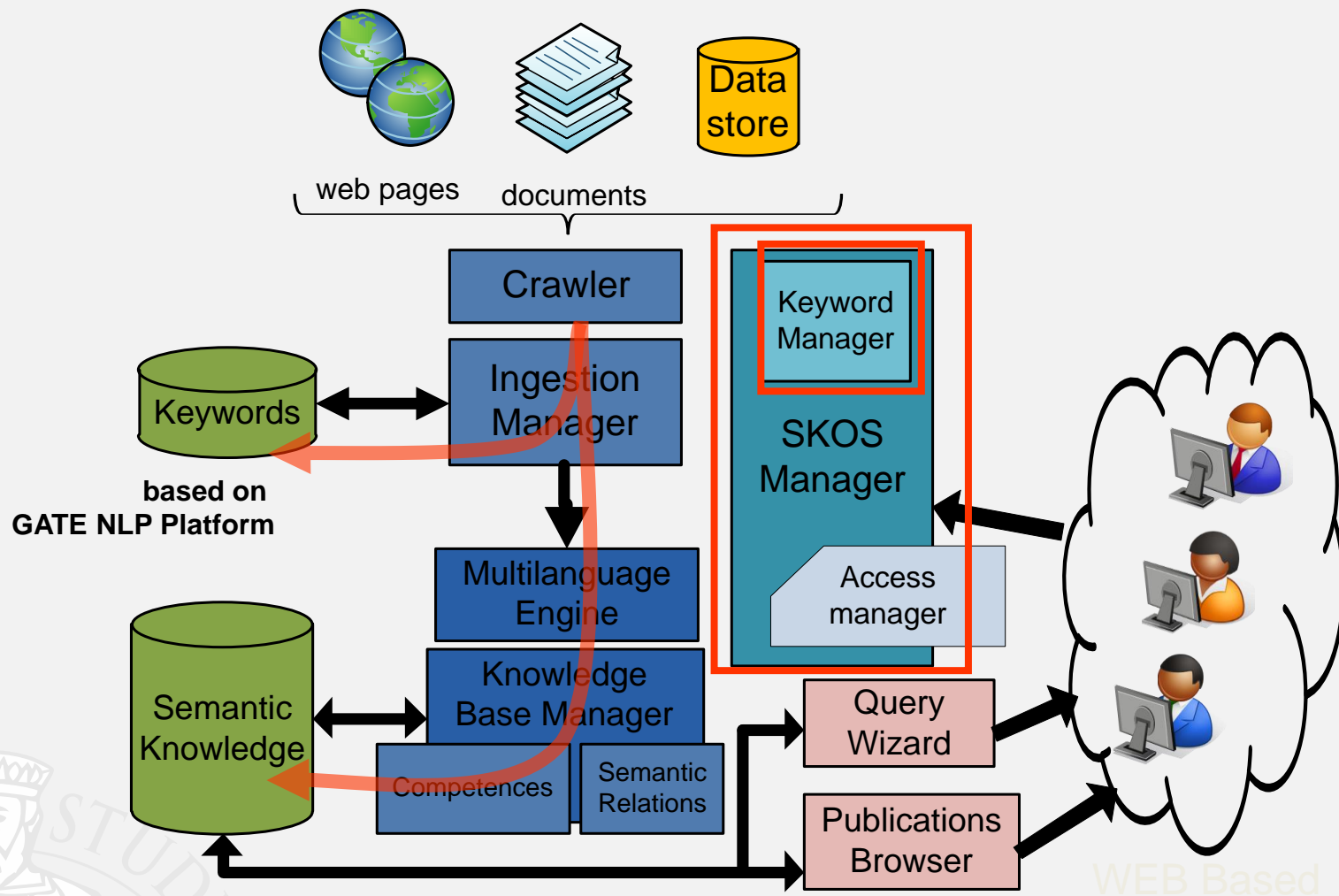
Previous 1 Next



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

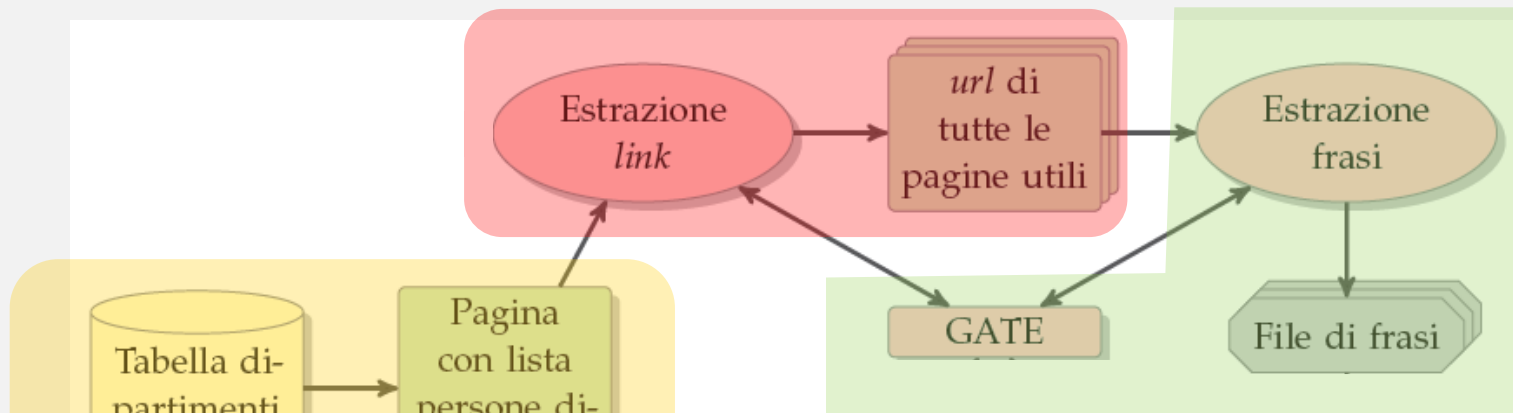
I risultati sono "il corso di Sistemi distribuiti" e la competenza "Sistemi distribuiti"

3 - Applicazioni: OSIM (Open Space Innovative Mind)



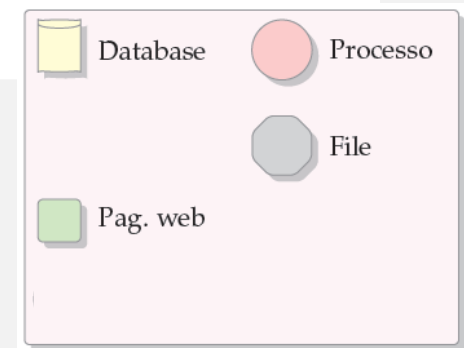
3 - Applicazioni: OSIM (Open Space Innovative Mind)

Vettore di URLs. Gli URLs sono scelti dalle apposite sezioni delle pagine, basandosi sui tags HTML delle pagine stesse



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

La seed list del crawler sono gli URLs delle pagine dipartimentali del Cerca Chi di UniFI



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

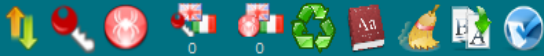
dipartimento di matematica per le decisioni - crawler is running

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: mate [Successivo](#) [Precedente](#) [Evidenzia](#) ☐ Maiuscole/minuscole

ONTOLOGY MANAGER

KEYS SELECTION

RELATIONS MANAGER



INSTANCES

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

1



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)
- architetture (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

- [INFO]: LOOKUP FOR fornire strumenti (11)
- Related Subject:
- 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=&cognome=&nome=&f=s
- 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=&cognome=&nome=&f=s
- 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=&cognome=&nome=&f=s
- 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=&cognome=&nome=&f=s
- 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=&cognome=&nome=&f=s

CoSKOSAM

OSIM – Frammento
Ontologia di dominio

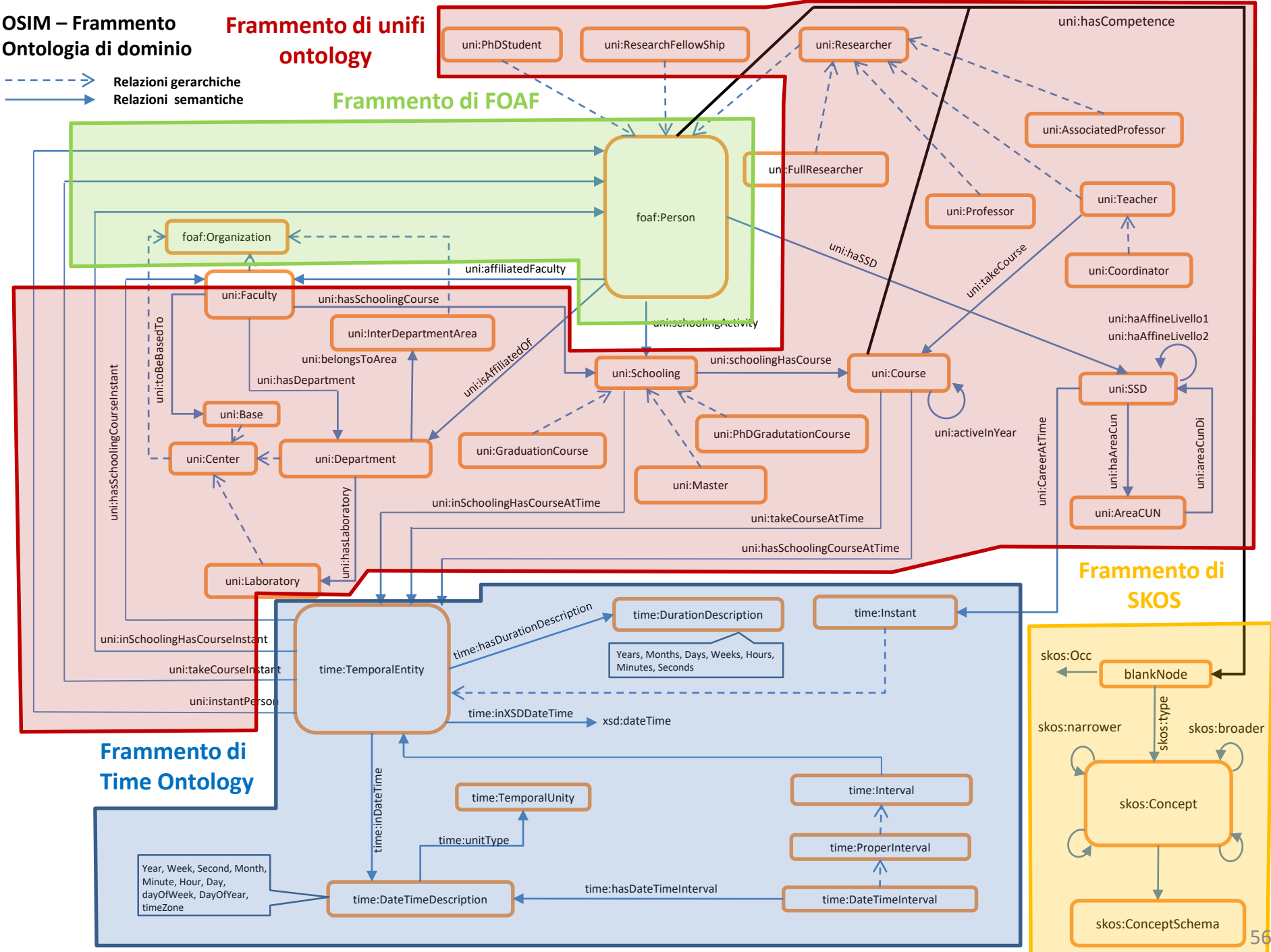
Frammento di unifi
ontology

Frammento di FOAF

Frammento di
SKOS

Frammento di
Time Ontology

Relazioni gerarchiche
Relazioni semantiche



3 - Applicazioni: OSIM (Open Space Innovative Mind)

*Ricerca Full Text
con Logica Fuzzy*

Question Answer

Query Wizard

*Ricerca
Assistita*

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

Home Documentation Search Managing Knowledge Browsing People & Publications Contact DISIT

italiano

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 riguardano ?

3 - Applicazioni: Motori di Ricerca Semantici



<https://www.google.it/intl/it/insidesearch/features/search/knowledge.html>

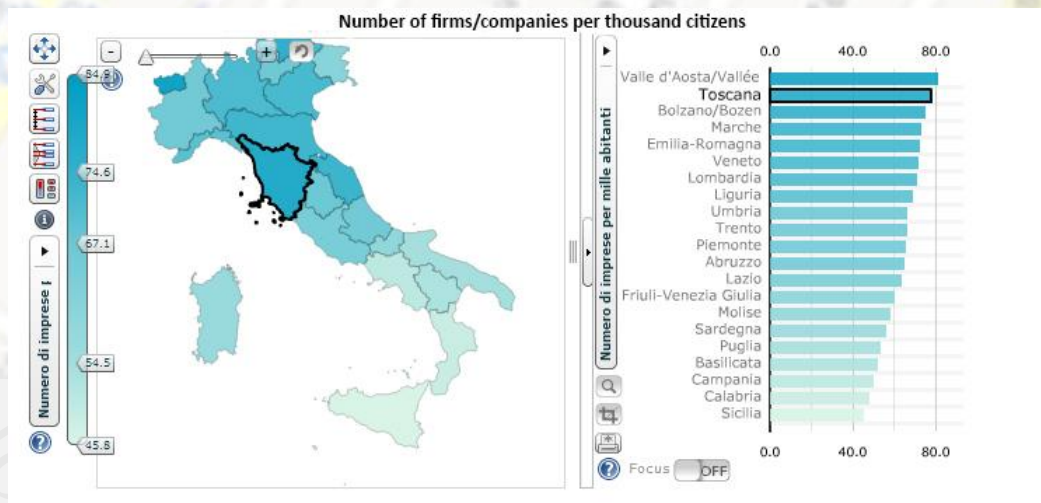


<https://searchengineland.com/library/bing/bing-satori>

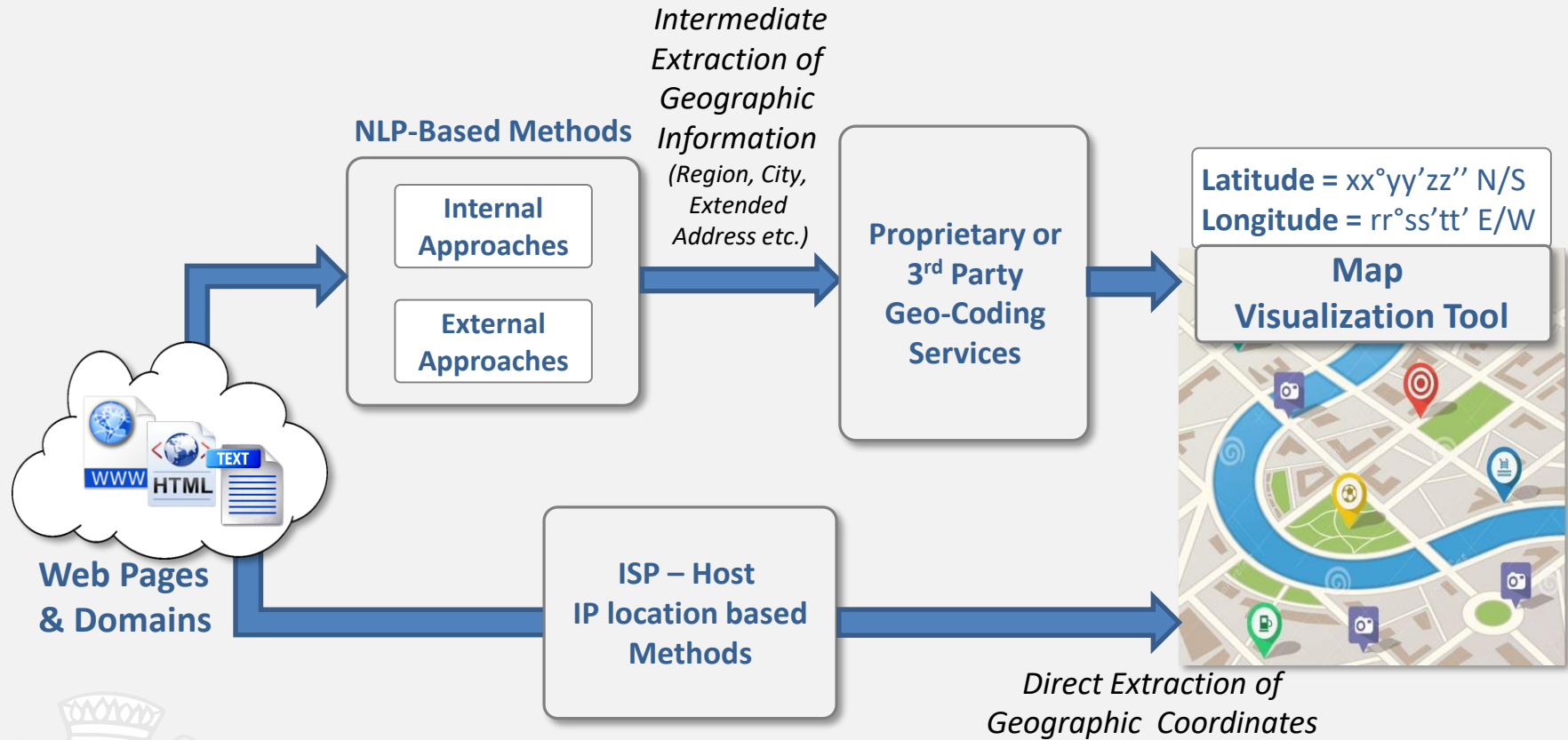
3 - Applicazioni: GeoLocator

- La Geolocalizzazione puntuale di un numero di servizi sempre crescent (di cui è possibile reperire risorse attraverso il web o gli Open Data delle Pubbliche Amministrazioni) sta diventando un requisito sempre più importante.
- Molti servizi, sepcialmente in ambito Smart City, sono basati su dati non strutturati e Open Data, spesso in formati diversi e non standardizzati tra loro, e soprattutto non geolocalizzati.

In 2015, Only 30000 commercial operators appears to be active in the Tuscany region, according Public Administration Open Data and among GoodRelations Ontology users.



3 - Applicazioni: GeoLocator



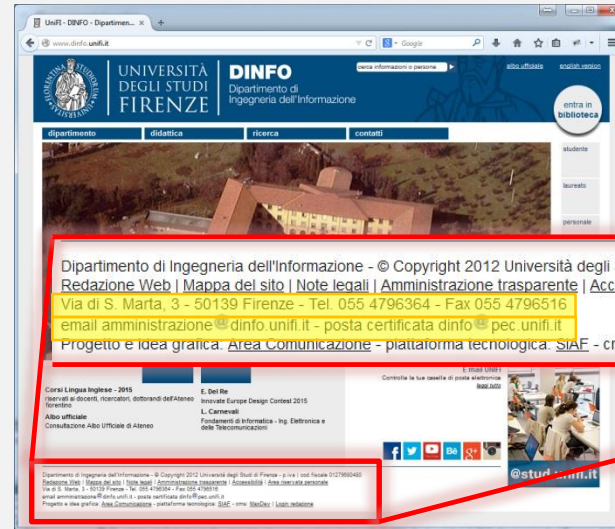
3 - Applicazioni: GeoLocator

GeoLocator [Nesi et al., 2016]

Address Extractor

I tag HTML (in particolare, *footer* e *header*) sono usati generalmente per contenere e visualizzare informazioni amministrative e fisiche dell'azienda / organizzazione / servizio proprietario del dominio web.

Estrazione degli indirizzi tramite tecniche di NLP, in particolare Pattern Matching:



- *General (high level detail) address pattern:*

[REGION] + [PROVINCE] + [POSTAL_CODE] + [CITY].

- *Specific (low level detail) address pattern:*

[STREET_IDENTIFIER] + [STREET_NAME] + [STREET_NUMBER].

- *Pattern per estrarre attributi speciali (e.g.: "Scala A, Interno 4"):*

[INNER_BLOCK_ID1] + [ID1_VALUE] + [INNER_BLOCK_ID2] + [ID2_VALUE].


- *Pattern per estrarre coordinate geografiche (se presenti):*


[LATITUDE_IDENTIFIER] + [LATITUDE_VALUE] + [LONGITUDE_IDENTIFIER] + [LONGITUDE_VALUE].


3 - Applicazioni: Paval - A Location Aware Virtual Personal Assistant -


Query Form Page


https://paval.disit.org/Paval/query.jsp


UNIVERSITÀ
DEGLI STUDI
FIRENZE


DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE


UNIVERSITÀ
DEGLI STUDI
FIRENZE


DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE


DISIT
DISTRIBUTED SYSTEMS
AND INTERNET
TECHNOLOGIES LAB

www.disit.org

Will you allow paval.disit.org to access your location?
[Learn more...](#)
☐ Remember this decision

Allow Location Access

Don't Allow

m4City Semantic
Service Search

www.disit.org

A quale servizio sei interessato...?
Cosa vuoi fare? Dove?

vorrei mangiare una quattro stagioni

Select Language: ITA

Specificare Coordinate Utente
(opzionale):

Latitudine:

Longitudine:

Submit

Il servizio permette la ricerca tramite query in linguaggio naturale di servizi di interesse per l'utente nella provincia di Firenze, restituendo le aziende legate al tipo di servizio cercato attraverso interrogazione del repository semantico di [Km4City](#), piattaforma Smart City ideata, progettata e realizzata da [DISIT Lab](#) dell'Università di Firenze. Se l'utente accetta di inviare le coordinate relative alla sua posizione attuale, il sistema restituisce i risultati dei servizi considerati attinenti alla ricerca effettuata ordinati in base alla vicinanza con la posizione dell'utente stesso. In alternativa, e' possibile per l'utente specificare una coppia di coordinate negli appositi box; in questo caso, le coordinate inserite verranno prese come riferimento di posizione al posto delle coordinate di posizione rilevate. Se inoltre l'utente inserisce nella query di ricerca il riferimento geografico ad un luogo di interesse (via, comune, quartiere), i risultati restituiti saranno ordinati per vicinanza con tale destinazione e non piu' in base al riferimento posizionale dell'utente.

Esempi di domande:


Esempio 1: "Vorrei vedere un museo in centro"

Esempio 2: "Sono in Via di Novoli e voglio fare la manicure"

Esempio 3: "Mi fanno male i denti"

Esempio 4: "Voglio tagliarmi i capelli a Firenze"

3 - Applicazioni: Paval - A Location Aware Virtual Personal Assistant -




UNIVERSITÀ
DEGLI STUDI
FIRENZE

DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

DISIT
DISTRIBUTED SYSTEMS
AND INTERNET
TECHNOLOGIES LAB

www.disit.org

**DISIT - Km4City Semantic
Service Search**



UNIVERSITÀ
DEGLI STUDI
FIRENZE

DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

DISIT
DISTRIBUTED SYSTEMS
AND INTERNET
TECHNOLOGIES LAB

www.disit.org

[Home](#)

Query:

Lingua impostata: ITA

Query Results:

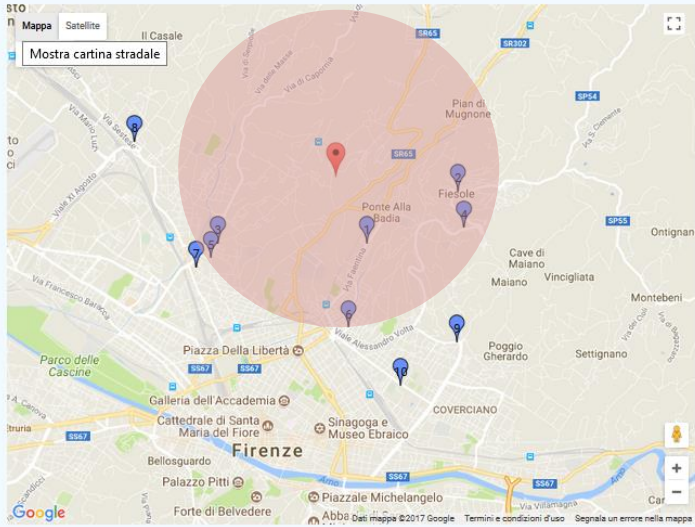
1. Nome: L'antica Badia
Indirizzo: VIA FAENTINA
n°: 342/A
CAP: 50133
Comune: FIRENZE
Prov: FIRENZE
Latitudine: 43.7992
Longitudine: 11.2756
Descrizione: Pizza sul Panaro
Settore: Pizzeria
[ServiceMap URL](#)

User Rating: ★★★★★

2. Nome: Oliva Filippo
Indirizzo: PIAZZA MINO DA FIESOLE
n°: 2
CAP: 50014
Comune: FIESOLE
Prov: FIRENZE
Latitudine: 43.8067
Longitudine: 11.2936
Descrizione:
Settore: Pizzeria
[ServiceMap URL](#)

User Rating: ★★★★★

3. Nome: I Sette Peccati
Indirizzo: VIA TADDEO ALDEROTTI
n°: 87
CAP: 50139
Comune: FIRENZE
Prov: FIRENZE
Latitudine: 43.7992
Longitudine: 11.246
Descrizione:
Settore: Pizzeria



Coordinate Utente (impostate manualmente dall'utente) - LAT.: 43.8087582, LONG.: 11.2693842
Dall'analisi della query risulta che sei interessato a trovare aziende e servizi relativi a: PIZZERIA nei dintorni della suddetta posizione.

Vengono restituiti 10 risultati per ogni tipologia di servizio considerata attinente alla ricerca effettuata dall'utente. Se l'utente ha accettato di trasmettere la sua posizione geografica (oppure ha inserito manualmente le coordinate), i risultati vengono ordinati per distanza decrescente rispetto a tale posizione; se l'utente non ha accettato di trasmettere la sua posizione geografica ma ha specificato un luogo geografico nella query di ricerca allora i risultati sono ordinati in base alla distanza (sempre decrescente) con il luogo geografico richiesto (se riconosciuto dal sistema); altrimenti, vengono restituiti i primi 10 risultati delle aziende appartenenti al servizio identificato nella provincia di Firenze nell'ordine con cui sono indicizzati nel repository Km4City.

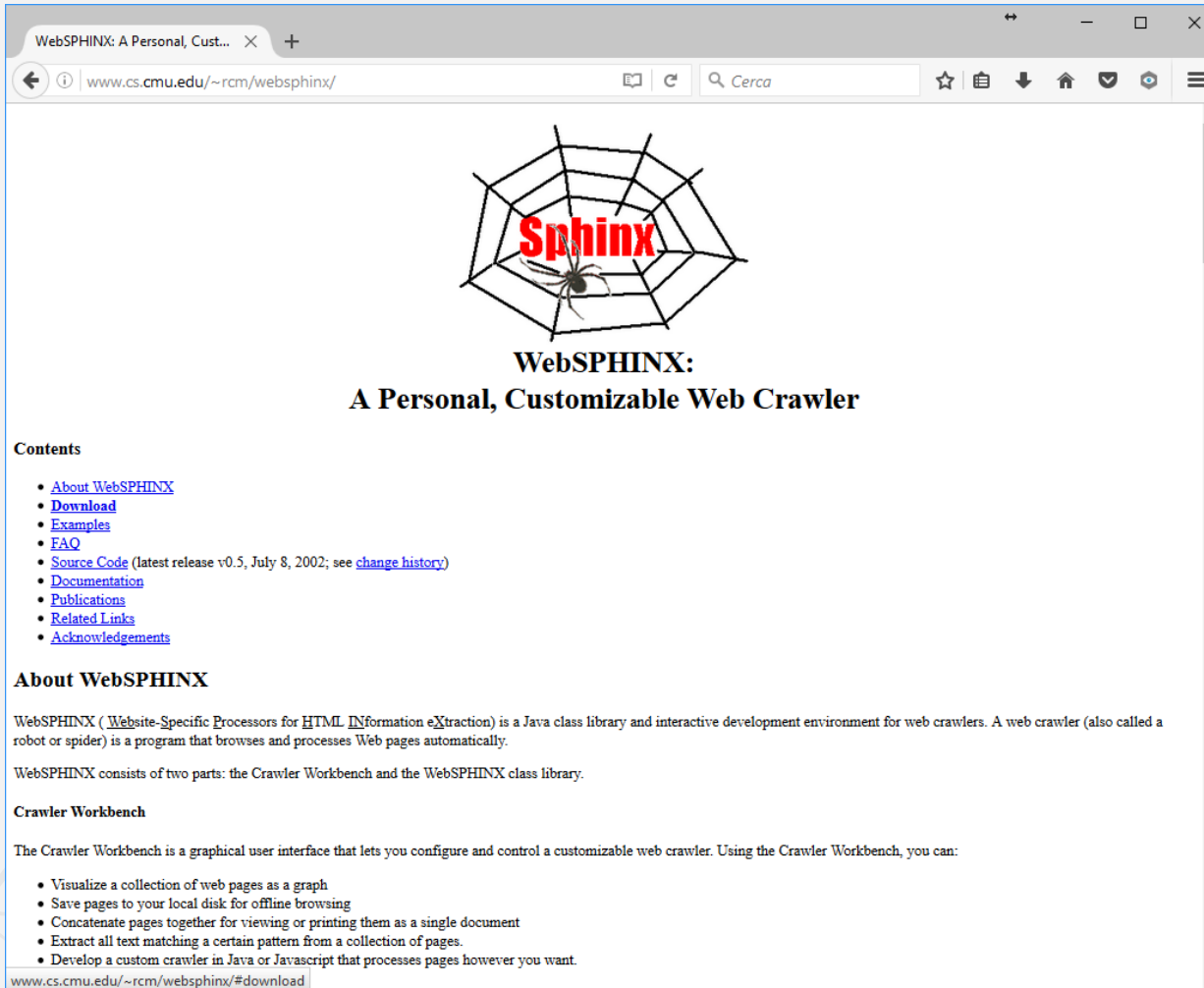
- Il marker rosso identifica la posizione utente (se trasmessa) o, in alternativa, le coordinate inserite manualmente. Se presente, i risultati sono ordinati per distanza da questo.

- Il marker giallo rappresenta un riferimento geografico eventualmente presente nella richiesta dell'utente. Se presente, i risultati sono ordinati per distanza

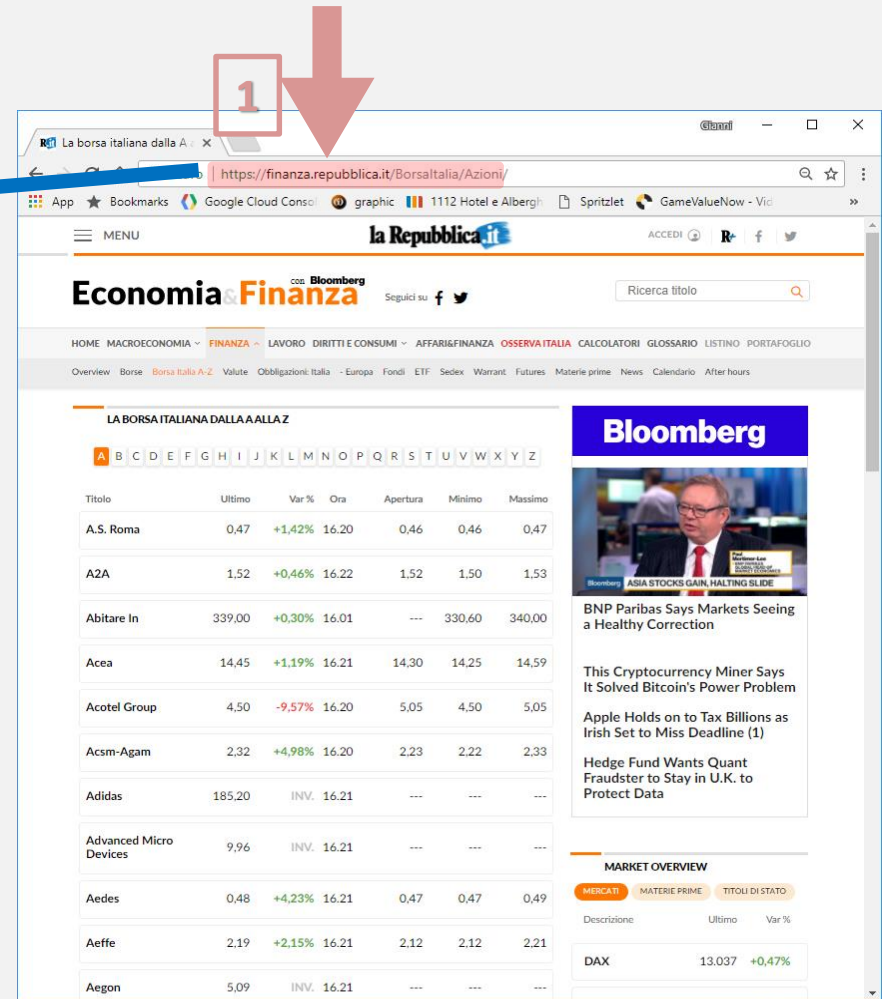
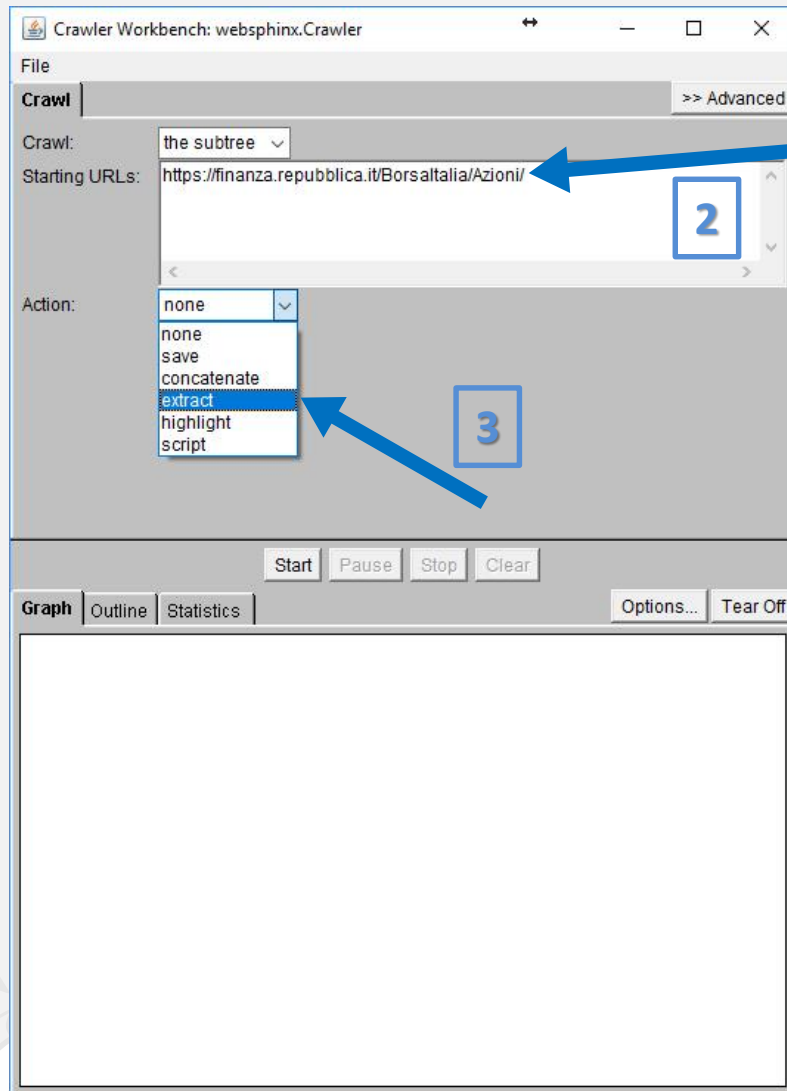
62

3 - Applicazioni: WebSPHINX

<http://www.cs.cmu.edu/~rcm/websphinx/>

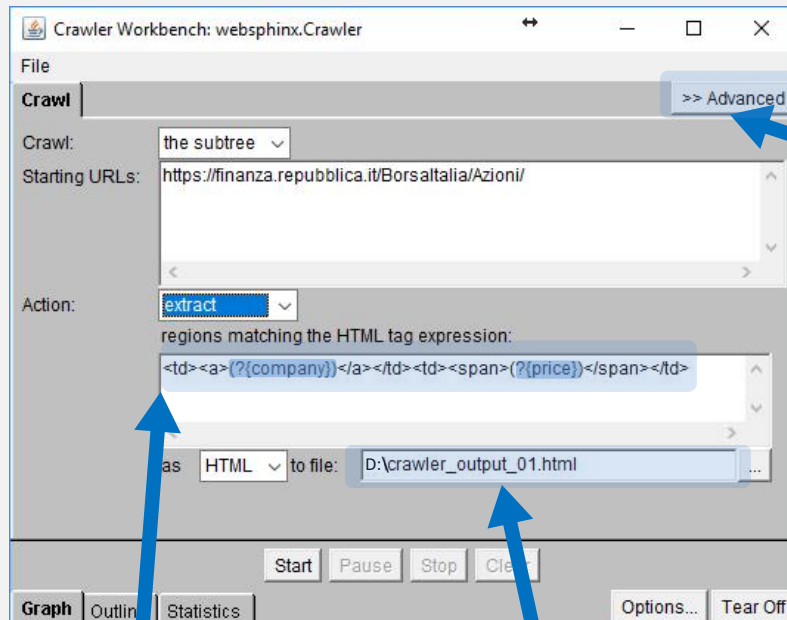


3 - Applicazioni: WebSPHINX



<https://finanza.repubblica.it/Borsaitalia/Azioni/>

3 - Applicazioni: WebSPHINX



5 Identificare la/le risorsa/e informative da Estrarre nel codice HTML e individuare un pattern HTML significativo con cui addestrare il crawler

6 Selezionare la gestione delle opzioni "Advanced"

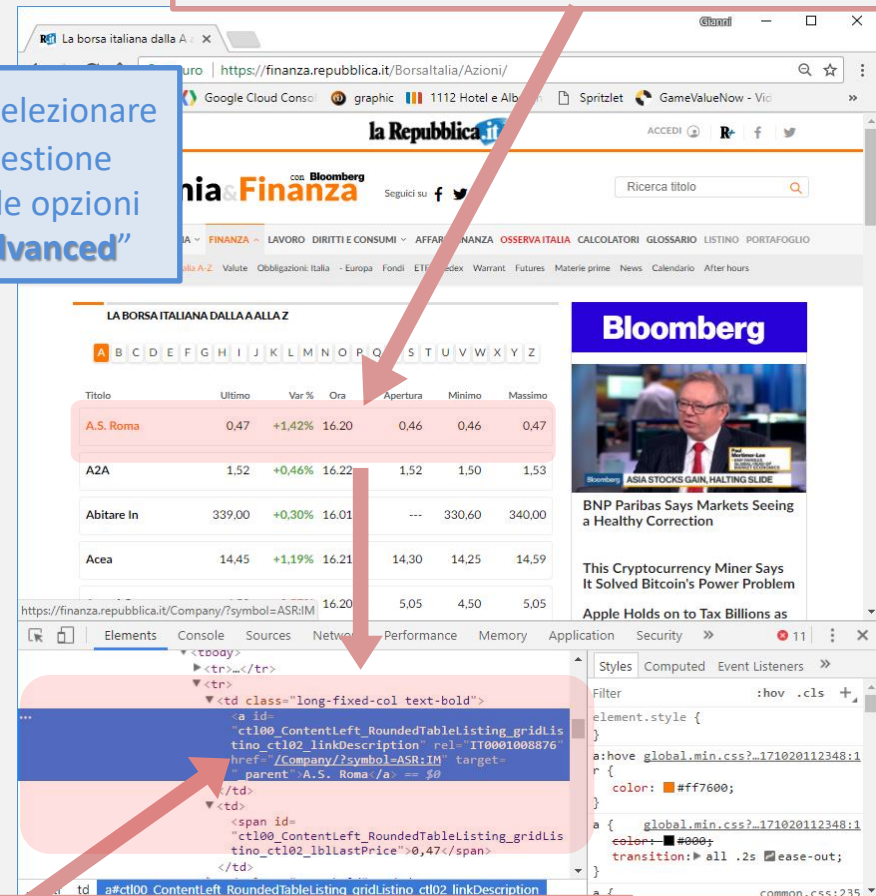
7a Inserire il pattern HTML scelto

8 Specificare il nome sel file in cui salvare il risultato dell'elaborazione

7b Per stampare nel file di uscita delle etichette per le descrivere risorse informative estratte:

```
<td><a>(?{company})</a></td>
<td><span>(?{price})</span></td>
```

4 Visualizzare il codice HTML della pagina:
F12 (Explorer, Firefox, Google Chrome)
Ctrl + U (Safari, Opera)



3 - Applicazioni: WebSPHINX

Crawler Workbench: websphinx.Crawler

File

Crawl Links Pages Classifiers Limits << Simple

Crawl: the subtree Using: all links

Starting URLs: <https://finanza.repubblica.it/Borsaitalia/Azioni/>

9 Impostare la profondità di navigazione (numero di livelli del grafo web)

10 Impostare la strategia di navigazione

Depth: 5 hops Breadth first

Start Pause Stop Clear

Graph Outline Statistics Options... Tear Off

La borsa italiana dalla A alla Z

la Repubblica.it

Economia con Bloomberg

HOME MACROECONOMIA FINANZA LAVORO DIRITTI E CONSUMI AFFARI&FINANZA OSSERVAITALIA CALCOLATORI GLOSSARIO LISTINO PORTAFOGLIO

Overview Borse Borsa Italia A-Z Valute Obbligazioni Italia Europa Fondi ETF Sedex Warrant Futures Materie prime News Calendario After hours

Titolo	Ultimo	Var.%	Ora	Apertura	Minimo	Massimo
A.S. Roma	0.47	+1.42%	16.20	0.46	0.46	0.47
A2A	1.52	+0.46%	16.22	1.52	1.50	1.53
Abitare In	339.00	+0.30%	16.01	---	330.60	340.00
Acea	14.45	+1.19%	16.21	14.30	14.25	14.59

LA BORSA ITALIANA DALLA A ALLA Z

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Elements Console Sources Network Performance Memory Application Security

```

<tr></tr>
<tr>
  <td class="long-fixed-col text-bold">
    <a id="
      "ct100_ContentLeft_RoundedTableListing_gridListino_ct102_linkDescription" rel="IT0001008876"
      href="/Company/?symbol=ASR:IM" target=
        "parent" A.S. Roma /a> == $0
    </td>
    <td>
      <span id=
        "ct100_ContentLeft_RoundedTableListing_gridListino_ct102_iblastPrice">0,47</span>
    </td>
  </tr>
  <tr>
    <td>
      <a#ct100_ContentLeft_RoundedTableListing_gridListino_ct102_linkDescription
  
```

Styles Computed Event Listeners

Filter :hov .cls +

```

element.style {
  a:hove global.min.css?_171020112348:1
  r {
    color: #ff7600;
  }
  a {
    global.min.css?_171020112348:1
    color: #000;
    transition: all .2s ease-out;
  }
}
common.css:235
  
```


3 - Applicazioni: WebSPHINX

Crawler Workbench: websphinx.Crawler

File

Crawl: **the subtree** >> Advanced

Starting URLs: <https://finanza.repubblica.it/Borsaitalia/Azioni/>

Action: **extract**

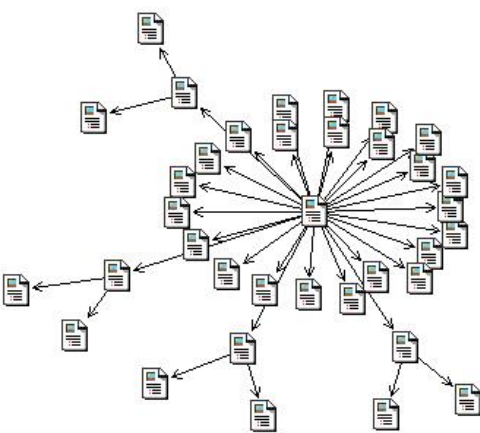
regions matching the HTML tag expression:

```
<td><a>(?{company})</a></td><td><span>(?{price})</span></td>
```

as HTML to file: D:\crawler_output_01.html

Start Pause Stop Clear

Graph Outline Statistics Options... Tear Off




Extracted Records

file:///D:/crawler_output_01.html

	company	price
1.	A.S. Roma	0,46
2.	A2A	1,52
3.	Abitare In	339,00
4.	Acea	14,39
5.	Acotel Group	4,57
6.	Acsm-Agam	2,32
7.	Adidas	185,20
8.	Advanced Micro Devices	9,96
9.	Aedes	0,48
10.	Aeffe	2,20
11.	Aegon	5,09
12.	Aeroporto Guglielmo Marconi di Bologna	15,36
13.	Agatos	0,28
14.	Ageas	40,43
15.	Ahold Del	16,98
16.	Air Liquide	107,20
17.	Airbus	85,50
18.	Alerion	2,93
19.	Alfio Bardolla	7,40
20.	Allianz	196,80
21.	Alphabet Classe A	884,50
22.	Altaba	59,25
23.	Ambienthesis	0,39
24.	Ambromobiliare	3,65
25.	Amgen	150,00
26.	Amplifon	12,42
27.	Anheuser-Busch	97,55
28.	Anima Holding	5,59

- Riferimenti

[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.

[IBM Watson, 2012]

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

[IOS Siri, 2010]

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

[Bellandi et al., 2012]

A. Bellandi, P. Bellini, A. Cappuccio, P. Nesi, G. Pantaleo, N. Rauch, *Assisted Knowledge Base Generation, Management and Competence Retrieval*, Int. Journal of Software Engineering and Knowledge Engineering, World Scientific Publishing Company, press, Vol. 32(8), pp.1007-1038, 2012.

[Nesi et al., 2016]

P. Nesi, G. Pantaleo and M. Tenti, *Geographical localization of web domains and organization addresses recognition by employing natural language processing, Pattern Matching and clustering*, Engineering Applications of Artificial Intelligence, Vol. 51, pp. 202-211, 2016.



- Link Utili

➤ **NLP**

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

The Stanford University NLP Group: <https://nlp.stanford.edu/>

Italian NLP Lab: <http://www.italianlp.it/>

➤ **Web Crawlers**

Websphinx: <https://www.cs.cmu.edu/~rcm/websphinx/>

Heritrix: <http://crawler.archive.org/index.html>

Apache Nutch: <http://nutch.apache.org/>

➤ **Semantic Technologies**

The Semantic Web: https://en.wikipedia.org/wiki/Semantic_Web

FOAF (Friend-Of-a-Friend) Ontology: <http://xmlns.com/foaf/spec/>

Time Ontology: <https://www.w3.org/TR/owl-time/>

SKOS (Simple Knowledge Organization System) Ontology: <https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html>