Web Crawling, Natural Language Processing & 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
pierfrancesco.bellini@unifi.it
gianni.pantaleo@unifi.it

Security And Knowledge Management
Prof. Pierfrancesco Bellini
A.A. 2019/20

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

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 millioni di DVD all'ora***, con una crescita di 5 volte rispetto al 2015 (*).

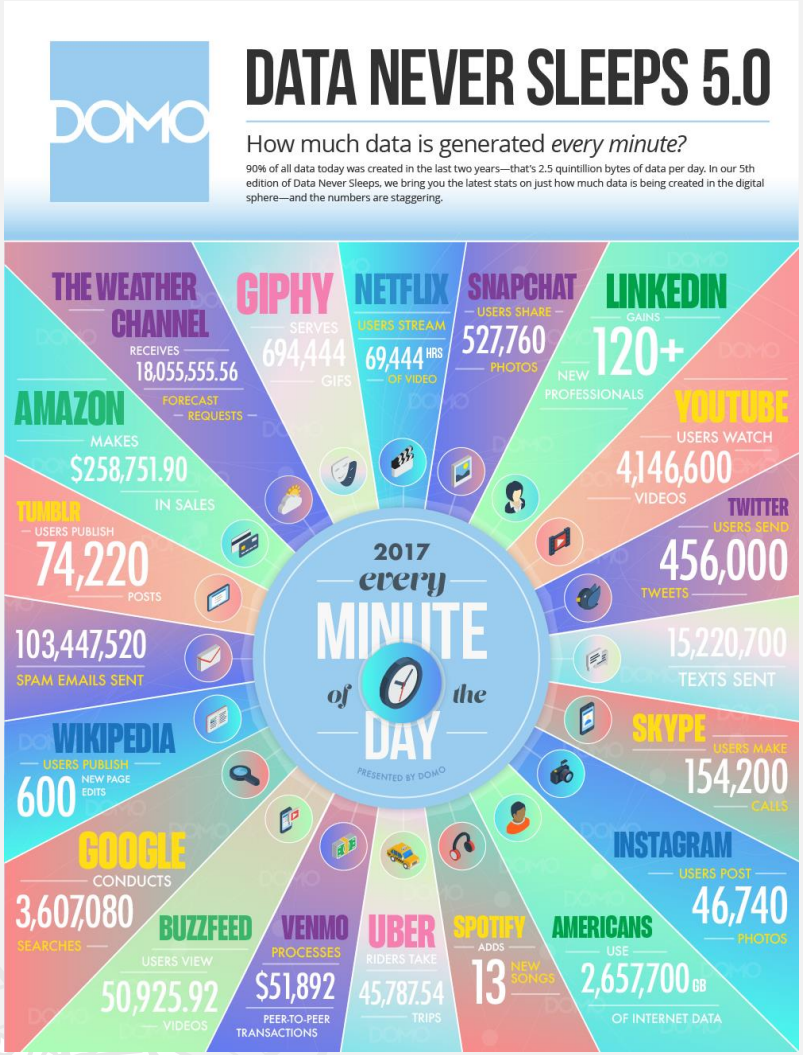
- La quantità di traffico dati nel web globale (*global data center traffic*), arriverà a ***15.3 Zettabyte nel 2020*** (**).

---

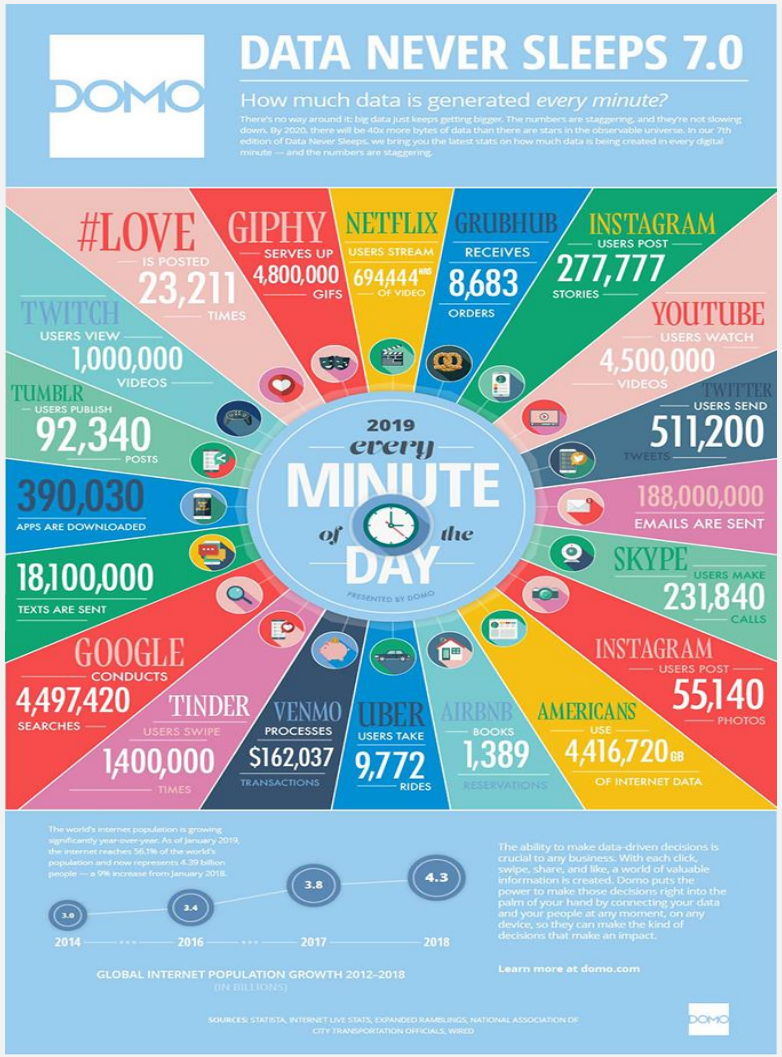
* Cisco Global Cloud Index: Forecast and Methodology, 2015–2020 White Paper (*Cisco Public Knowledge*)

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

Introduzione: Era dell'Information Technology



Data Never Sleeps 5.0 2017 Report
https://www.domo.com/learn/data-never-sleeps-5



Data Never Sleeps 7.0 2019 Report
https://www.domo.com/learn/data-never-sleeps-7

Introduzione: Era dell'Information Technology

LISA GROSSMAN SCIENCE 10.19.10 01:30 PM

TWITTER CAN PREDICT THE STOCK MARKET

The emotional roller coaster captured on Twitter can predict the ups and downs of the stock market, a new study

BUSINESS INSIDER UK FINANCE

The ECB says Twitter can predict the stock market

Oscar Williams-Grut
Jul. 22, 2015, 11:40 AM 1,790

FACEBOOK LINKEDIN TWITTER EMAIL PRINT

The European Central Bank (ECB) just put out an interesting study looking at whether Twitter and Google can be used to predict stock market moves — and its conclusion is, for Twitter, it can.

The ECB says: "Twitter bullishness has a statistically and economically significant predictive value in respect of share prices in the United States, the United Kingdom and Canada."

In other words, what people are

Twitter CEO Dick Costolo, right, celebrates the Twitter IPO with Twitter
ders, from left, Jack Dorsey, Biz Stone and Evan Williams on the

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: Generalità di un Web Crawler

Definizione di ***Web Crawler / Spider / Scraper / Bot***

- 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**, **Web Scraper** 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***)

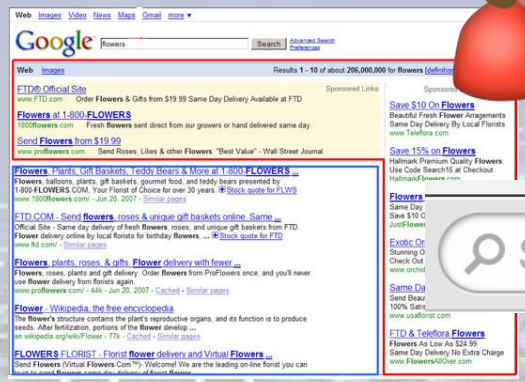
- Archiviare il contenuto di pagine web e documenti (***Web archival***)
- Analizzare il contenuto di pagine e siti web per produrre statistiche e analisi aggregate (***Web characterization***)
- Tenere copie o repliche di siti web (***Web mirroring***)

INTERNET ARCHIVE

- Analisi di siti web (***Web site analysis***)

1.1 - Introduzione: Processo di Crawling

Users



Motori di Ricerca

Recommendations

Web Indexes

The Web

Web Crawlers

& Content Parsing

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

HERITRIX https://webarchive.jira.com/wiki/display/Heritrix

nutch™ http://nutch.apache.org/

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

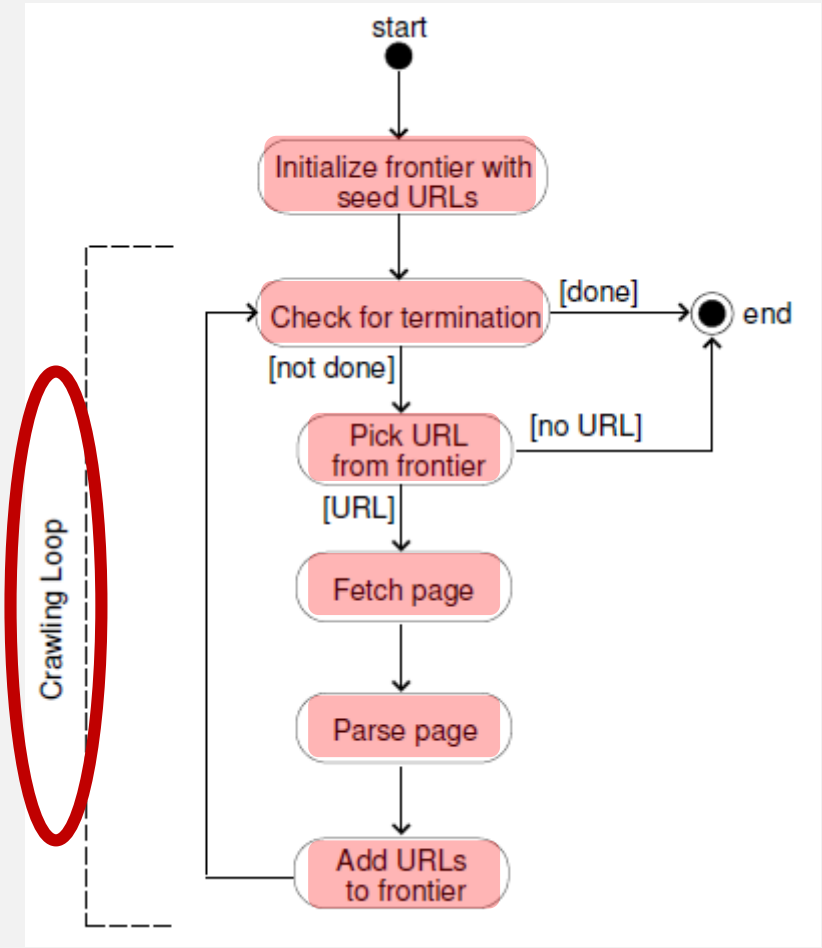
https://portia.readthedocs.io/en/latest/

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***)

- Un crawler viene di solito 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

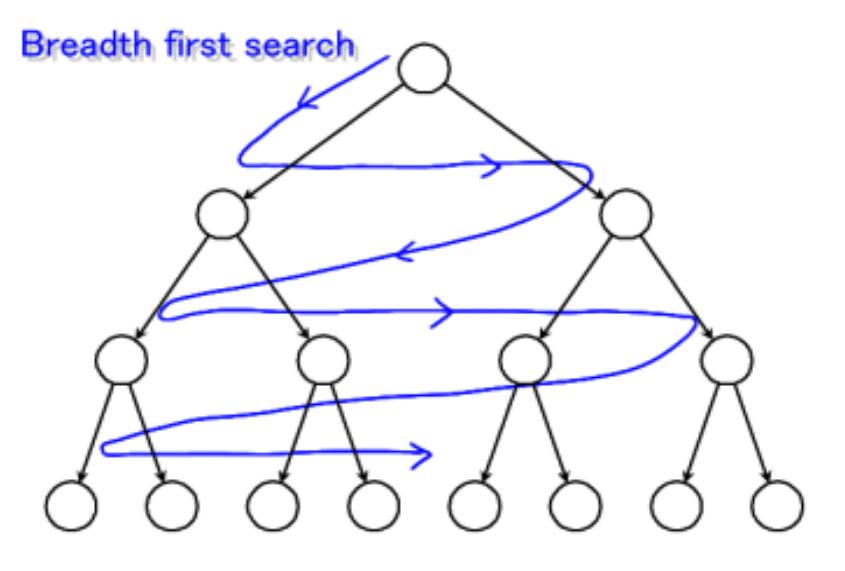
Occorre stabilire un meccanismo per decidere quale URL ignorare in caso di limite raggiunto

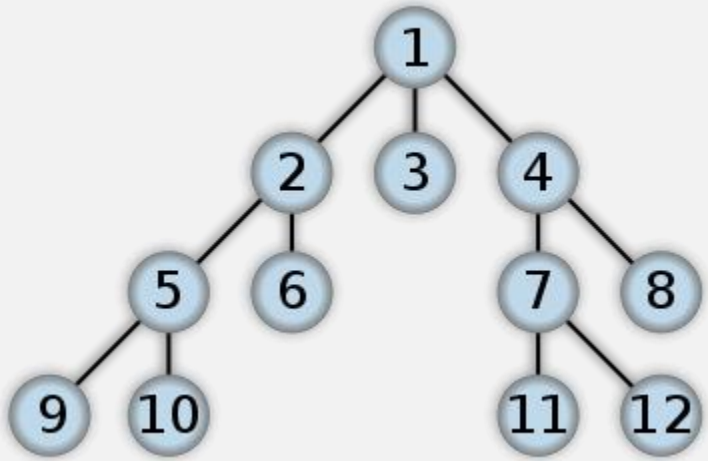
Outline

- 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

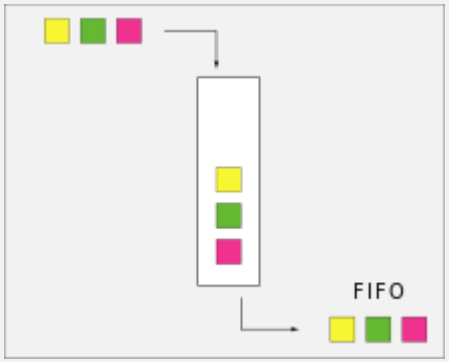
1.2 - Strategie di Crawling: Breadth First Search

- **Breadth First Search**



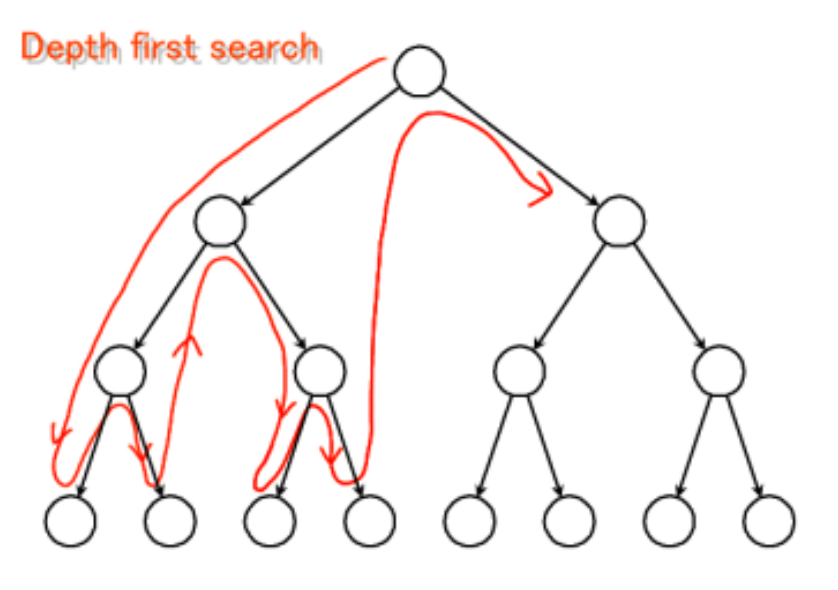


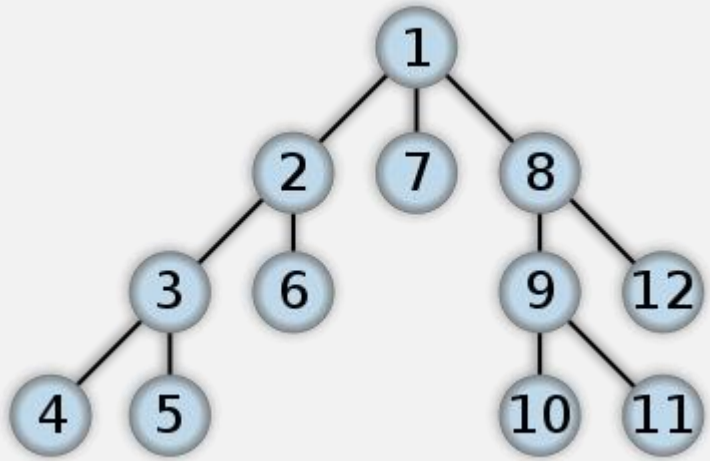
- Struttura dati lineare
  con politica
  FIFO : First In First Out (CODA)



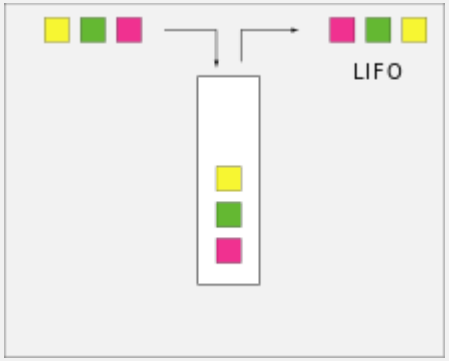
1.2 - Strategie di Crawling: Depth First Search

- **Depth First Search**





- Struttura dati lineare
  con politica
  LIFO: Last In First Out (PILA o STACK)



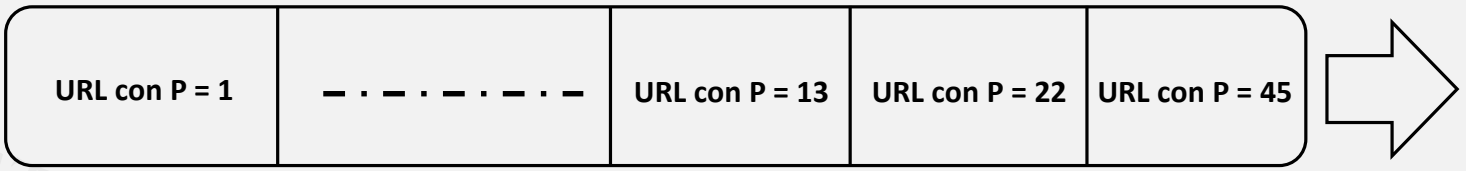
1.2 - Strategie di Crawling

- **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

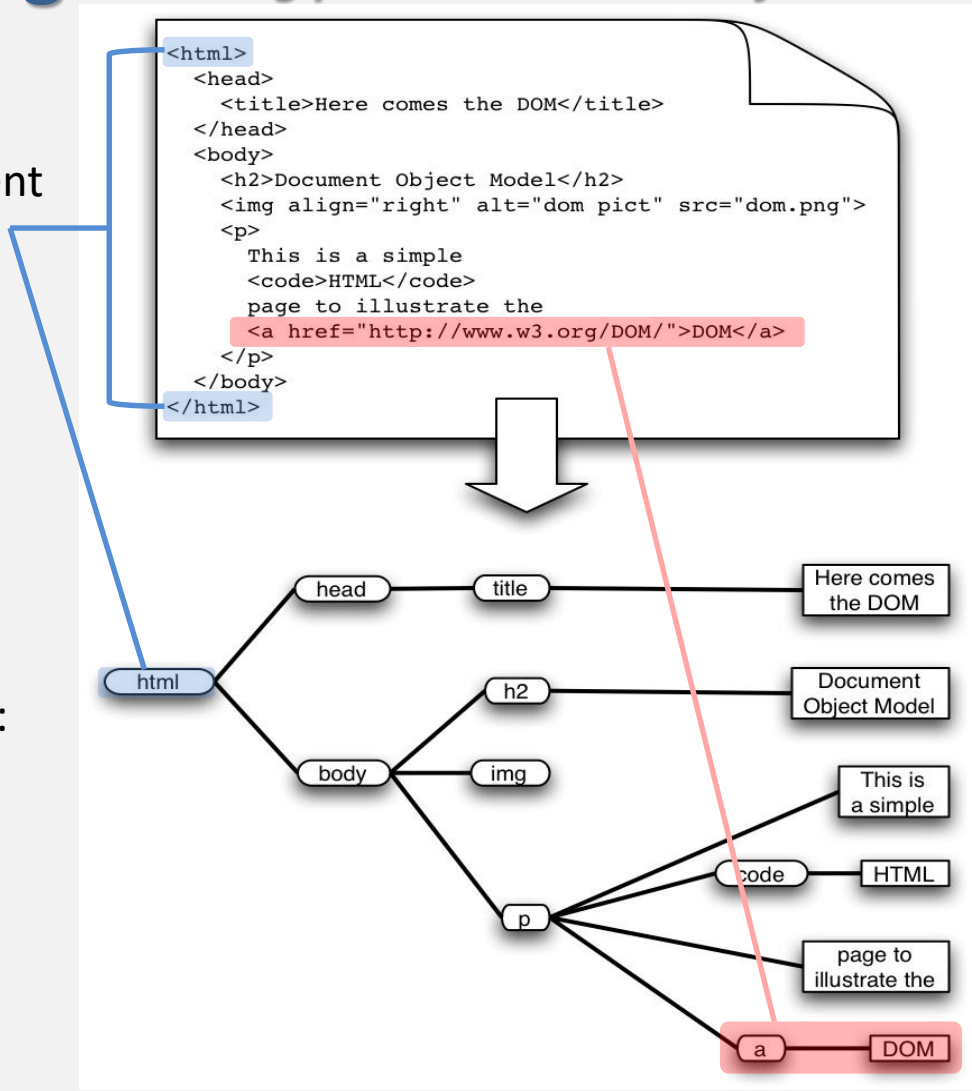


1.2 - Strategie di Crawling

- 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 - *Parsing per estrazione di Informazioni*

- I documenti HTML hanno una struttura ad albero - DOM (Document Object Model) definite da tag
- 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.mspx
Disallow: /communities/blogs/PortalResults.mspx
Disallow: /communities/rss.aspx
Disallow: /downloads/Browse.aspx
Disallow: /downloads/info.aspx
Disallow: /france/formation/centres/planning.asp
Disallow: /france/mnp_utility.mspx
Disallow: /germany/library/images/mnp/
Disallow: /germany/mnp_utility.mspx
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/*

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

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

User-agent: scooter
Disallow:

# all others
User-agent: *
Disallow: /
```

Google crawler può analizzare tutte le pagine eccetto queste

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

AltaVista non ha limiti

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

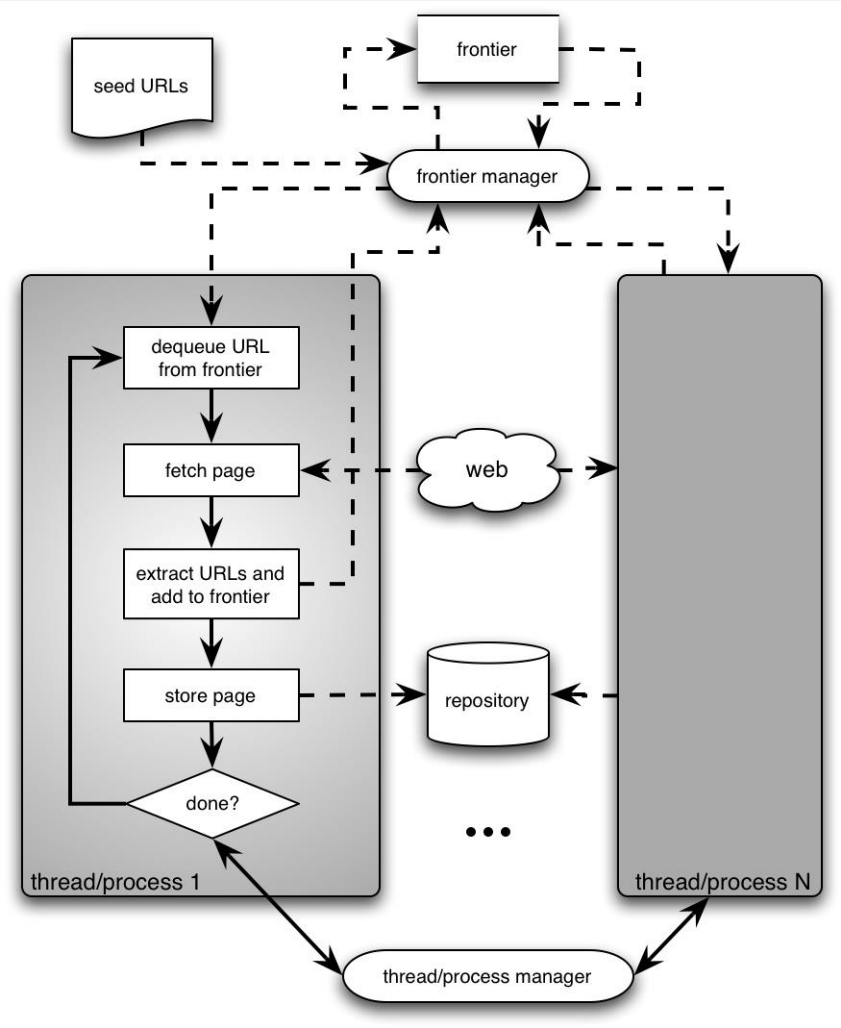
- Le operazioni effettuate dal crawler possono richiedere diverso tempo:
  - Risoluzione indirizzi IP
  - Connessioni al server e invio risposte
  - Ricezione pagina di risposta

- 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 sintattica (strutturale)*: *«Ieri ho visto l'uomo col binocolo»*
    - ✓ *Ambiguità semantica*: *«acuto» (persona intelligente, tipo di suono)*
    - ✓ *Ambiguità pragmatica:* *«la porta è aperta…»*

      L'intensione comunicativa può essere recepita diversamente dagli interlocutori:
      - Invito ad entrare liberamente…
      - Meglio chiudere, si è creata una corrente d'aria…

Ciò rende il processo di elaborazione automatica del linguaggio naturale un task molto complesso !

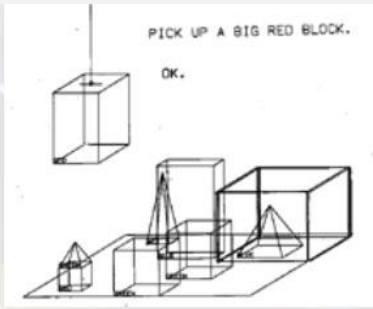
Outline

- 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

```
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.
```

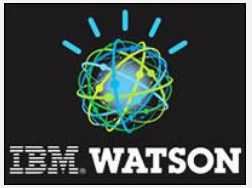


- **Apple Siri** [IOS Siri - 2010]
  - Virtual personal assistant
  - Knowledge navigator
  - User recommendation system.

- **Google Assistant**

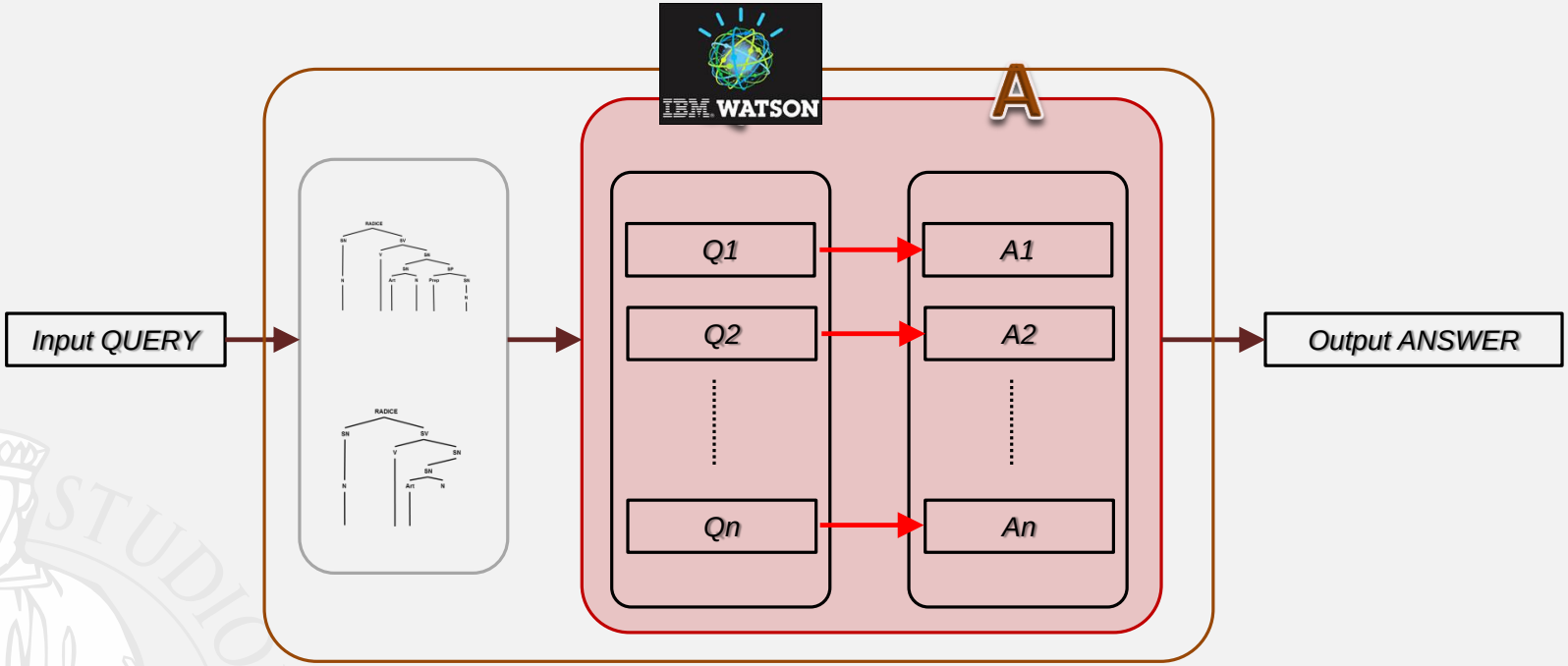
- **Microsoft Cortana**

- **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 .

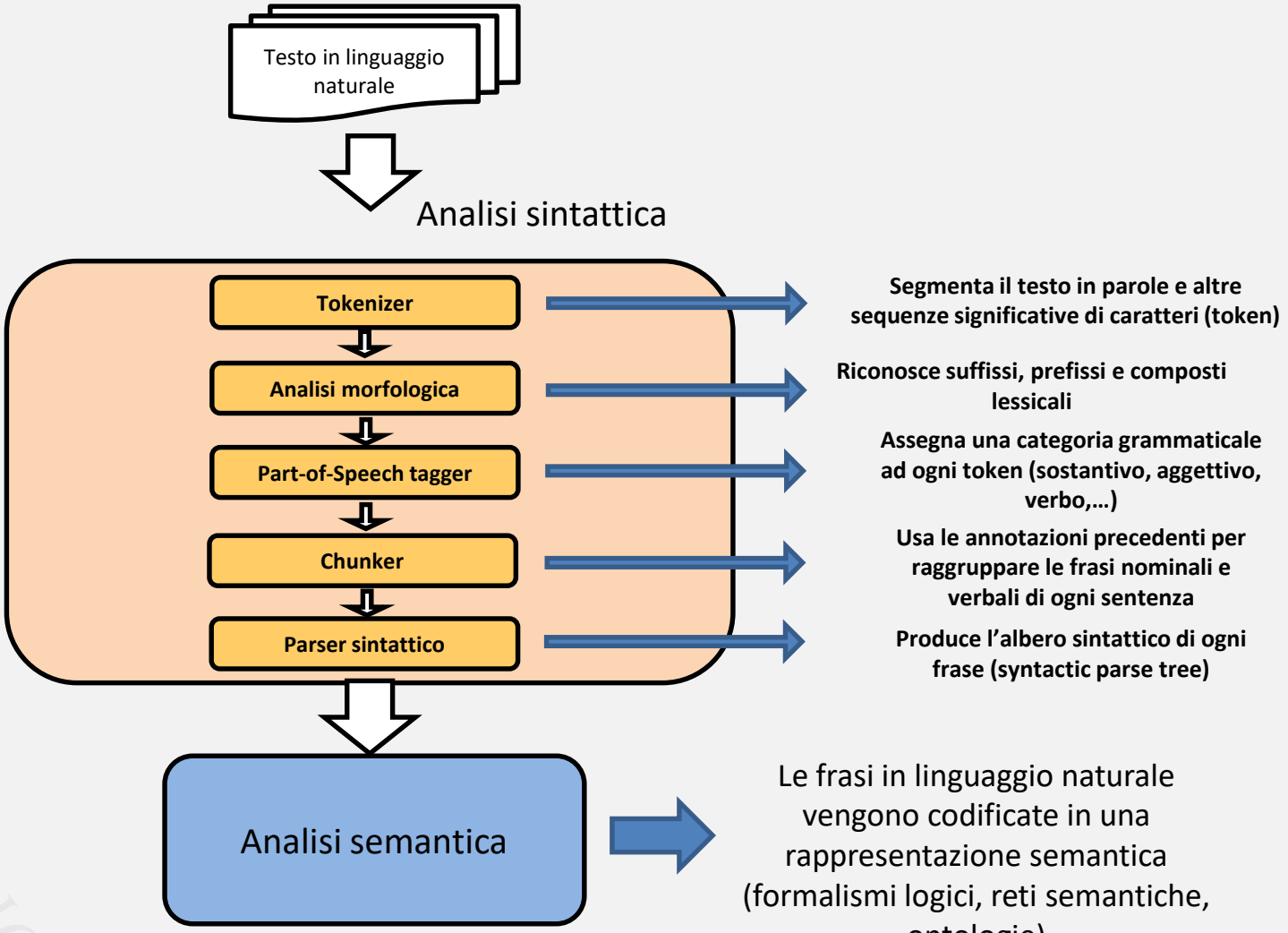
Syntactic Analysis: Le sequenze di parole sono trasformate in strutture che mostrano come le parole sono in relazione l'una con l'altra.

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

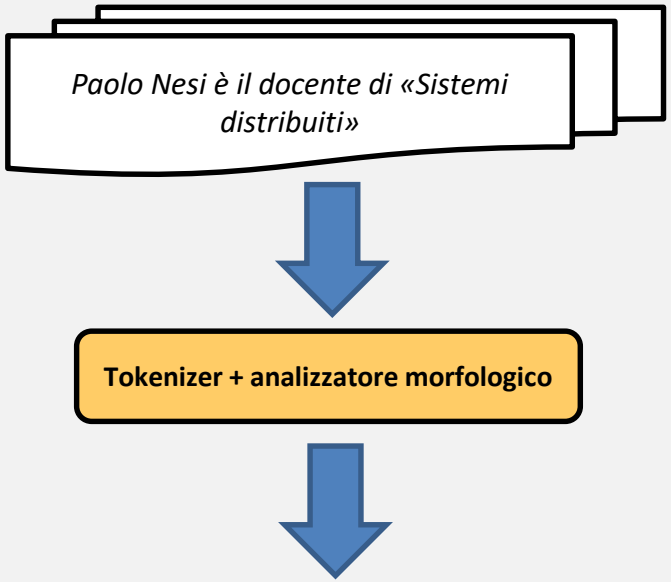
Discourse integration: il significato di una frase **spesso dipende dalla frase che la precede** e può influenzare quello della frase che la segue.

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

2.3 – Fasi di Elaborazione in Linguaggio Naturale



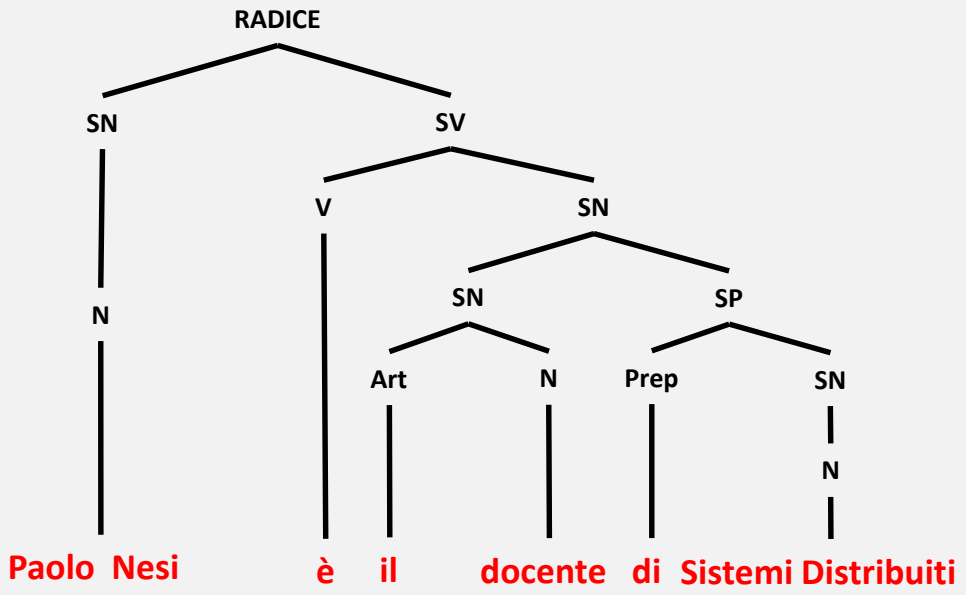
2.3 – Fasi di Elaborazione in Linguaggio Naturale



(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*

(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*

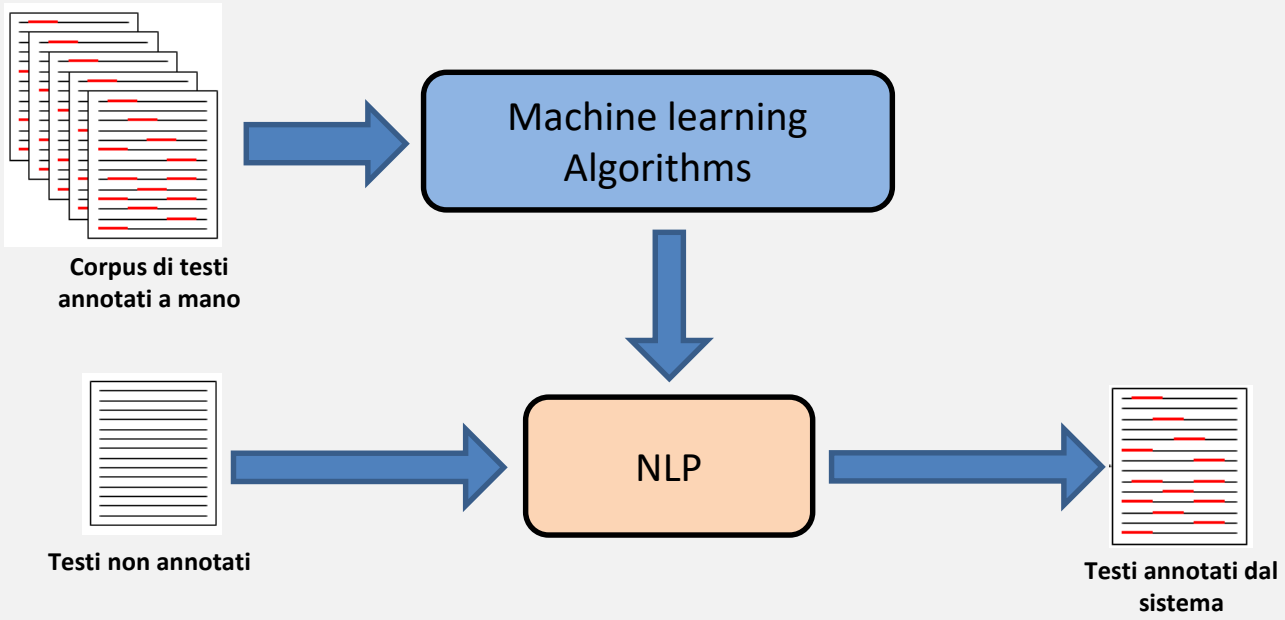
Part-of-Speech + chuncker + parser sintattico



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

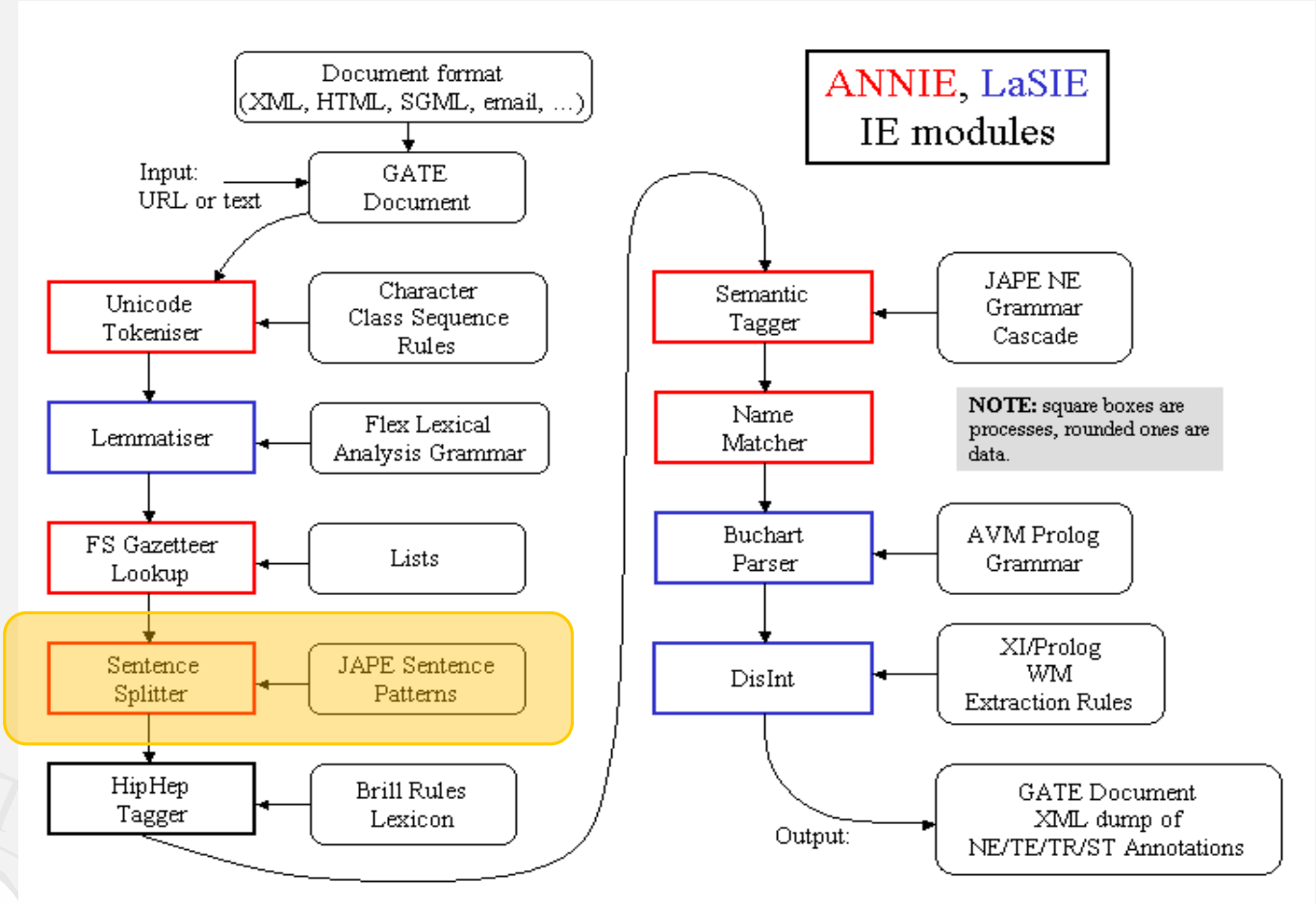


Outline

- 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

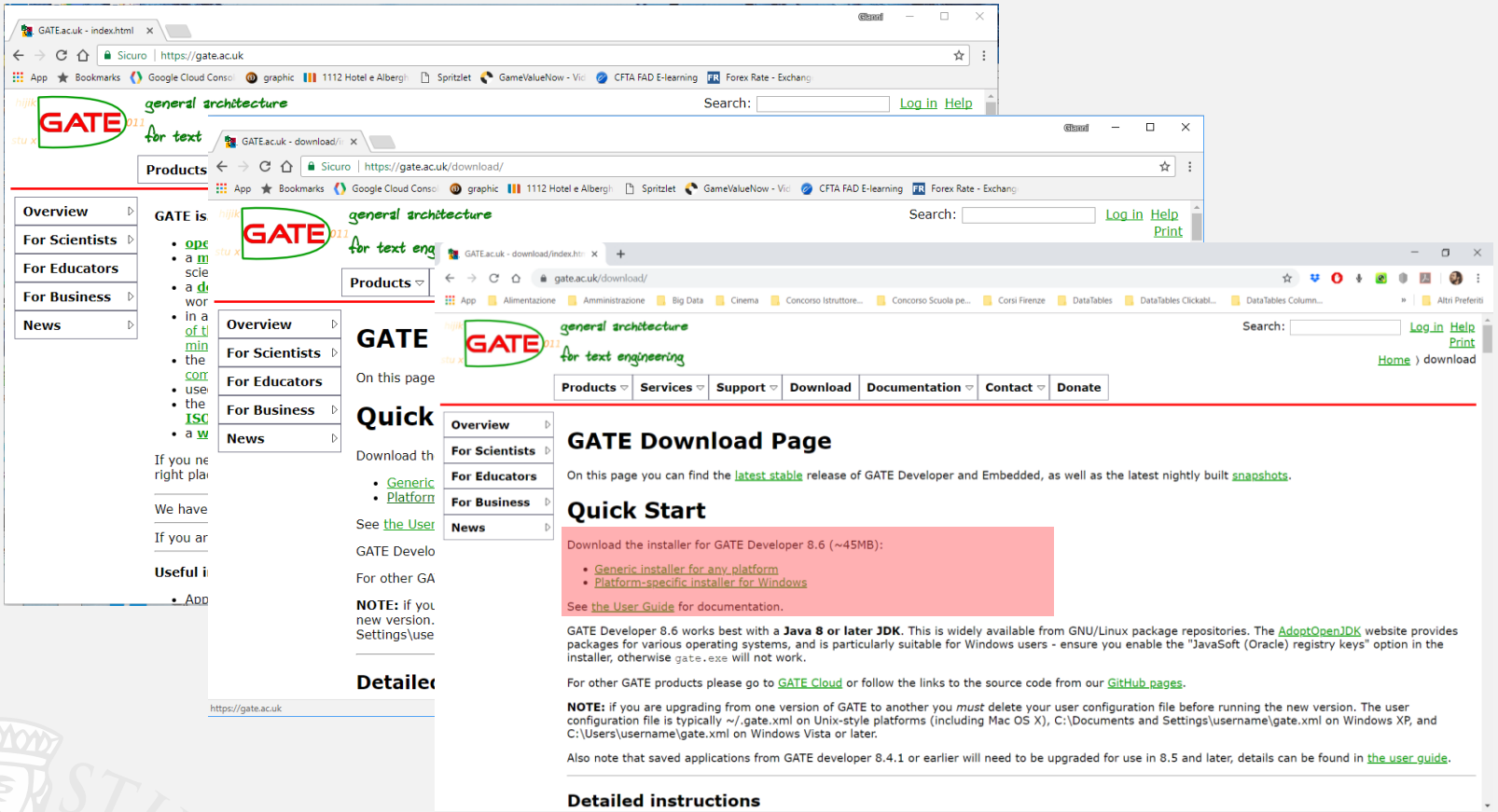
2.4 - NLP Tools: GATE

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



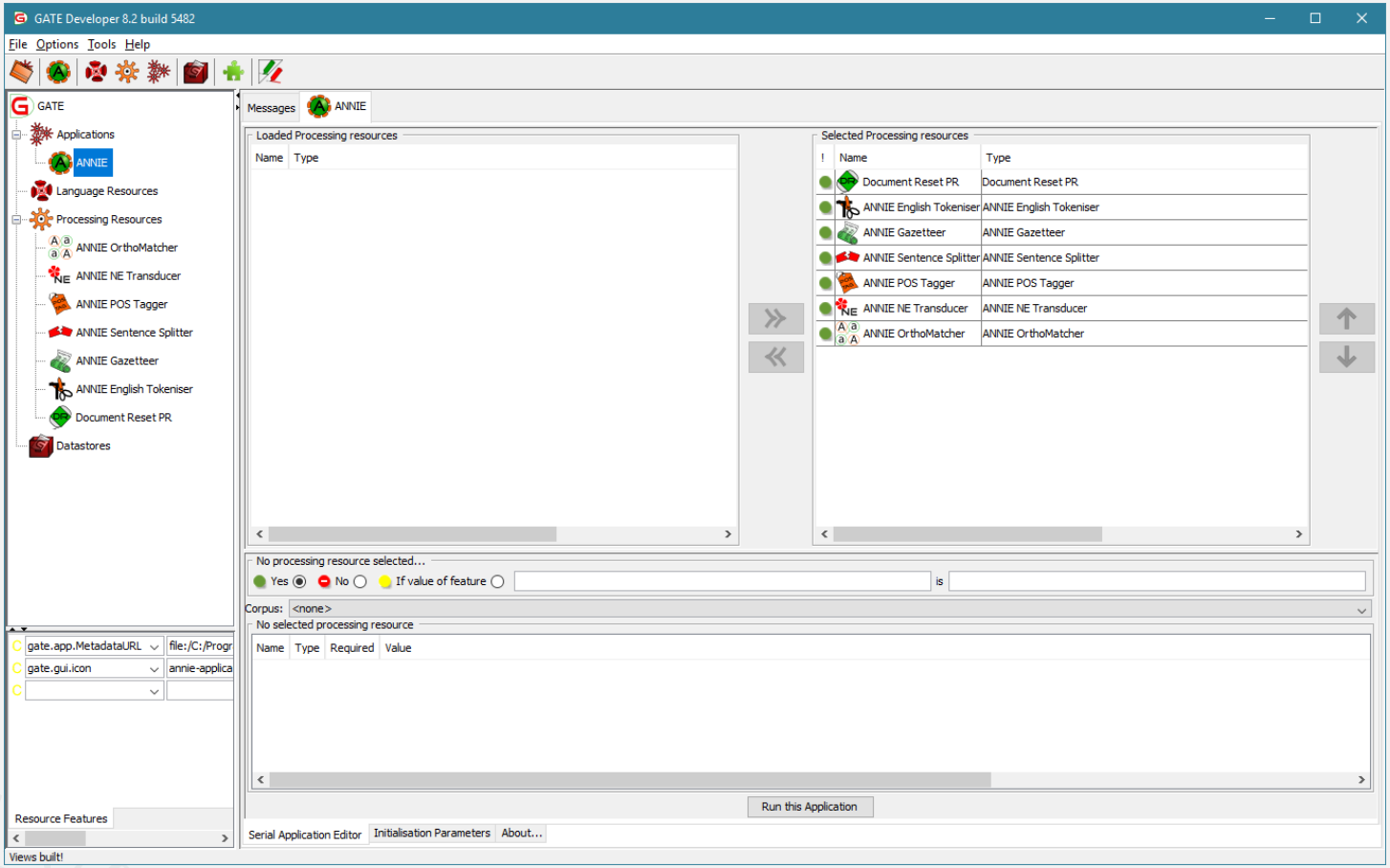
2.4 - NLP Tools: GATE

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



2.4 - NLP Tools: GATE

- **GATE: Interfaccia grafica e caricamento plugin principali**



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 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
  - Guida online: https://gate.ac.uk/sale/thakker-jape-tutorial/GATE%20JAPE%20manual.pdf

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

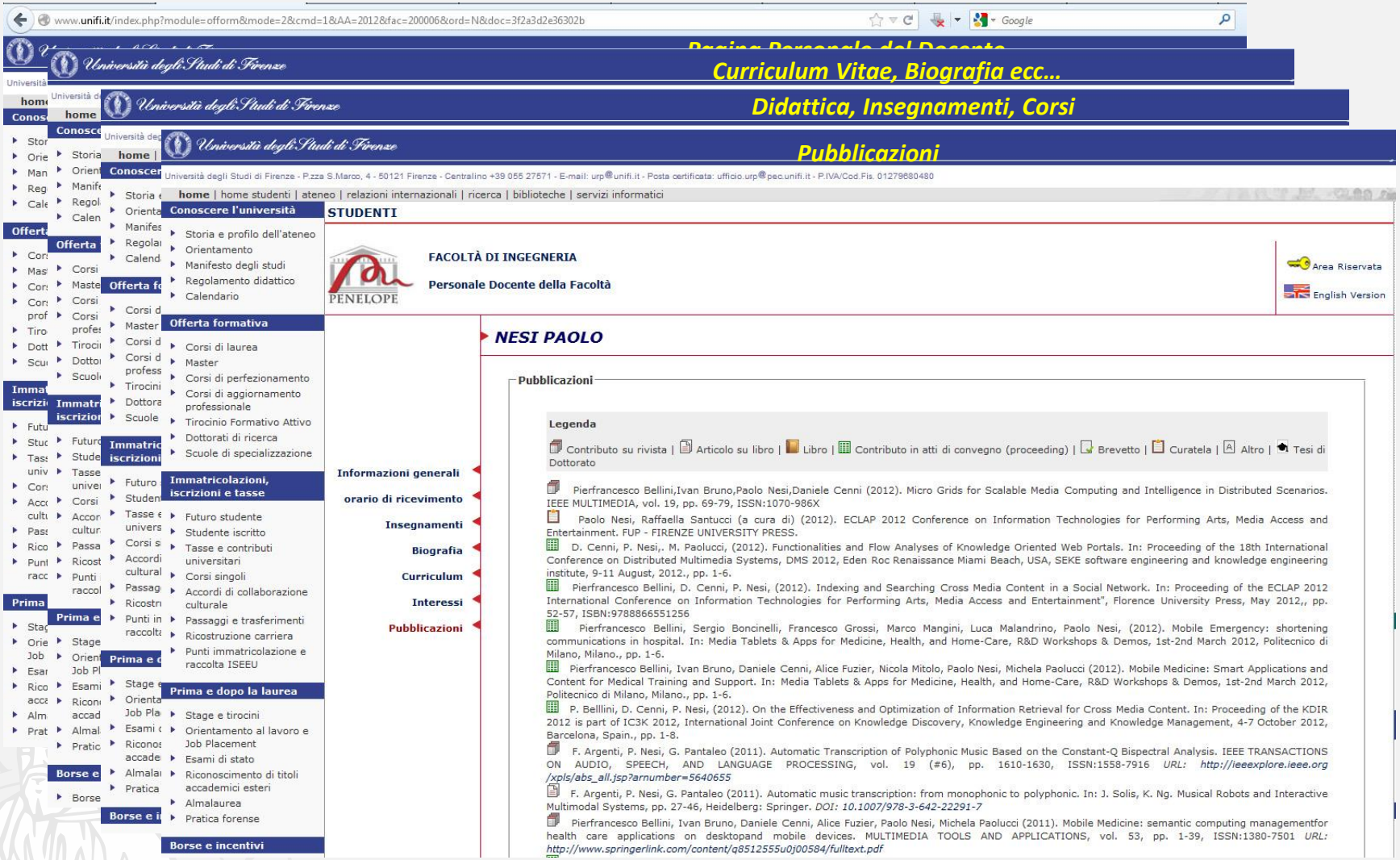
```
Rule: FindStreetAddress
Priority: 20
```
Nome della regola e priorità

```
(
    ({Token.string == "Via"} | Token.string == "Piazza"} | {Token.string == "Largo"})
    (({Lookup.majorType == NomeProprio} | {Lookup.majorType == Cognome}) |
     ({Lookup.majorType == Cognome}))
)
:address
-->
: address.indirizzo = {rule = "FindStreetAddress"}
```
Label

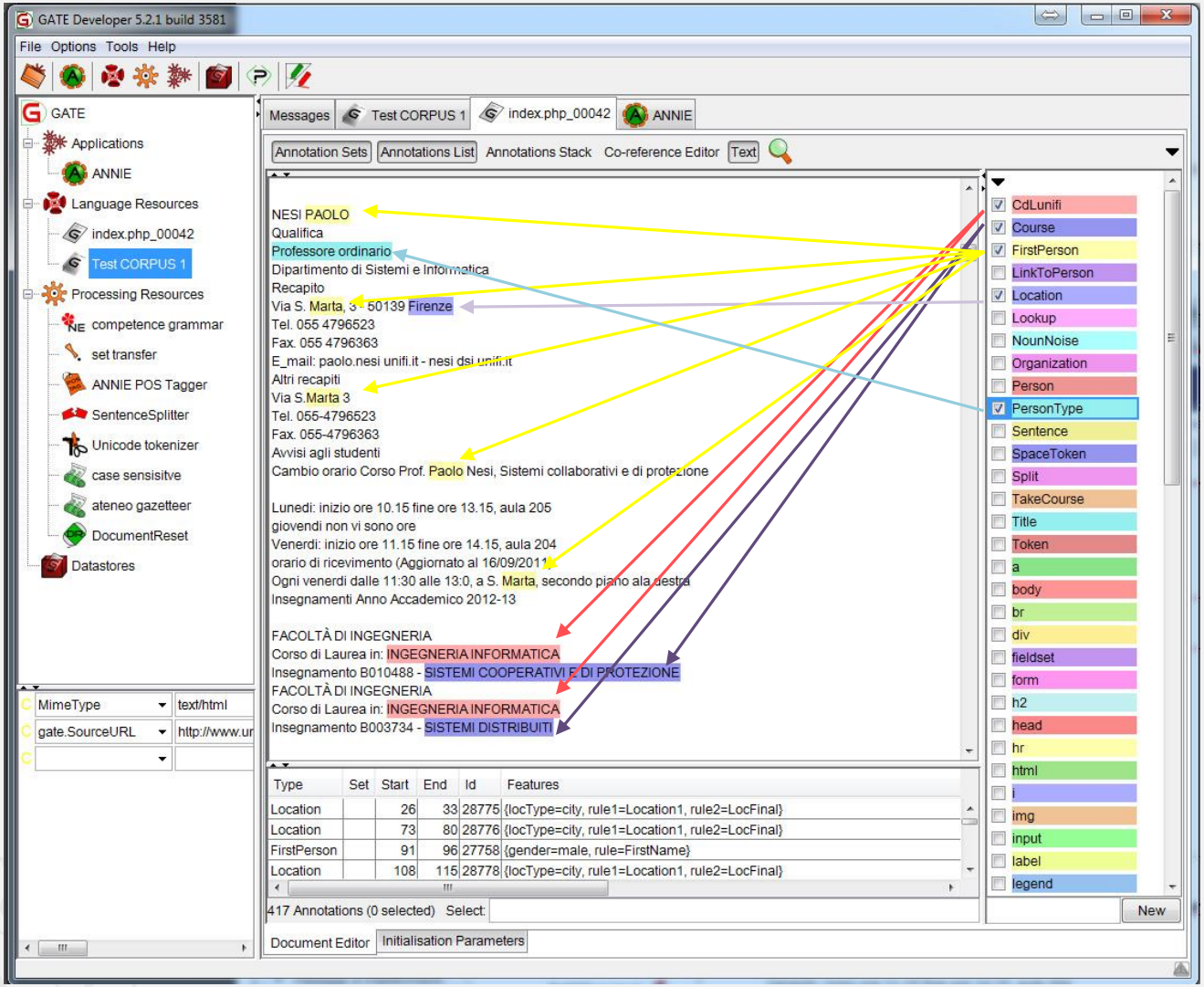
LHS

RHS

2.4 - NLP Tools: GATE



2.4 - NLP Tools: GATE



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 

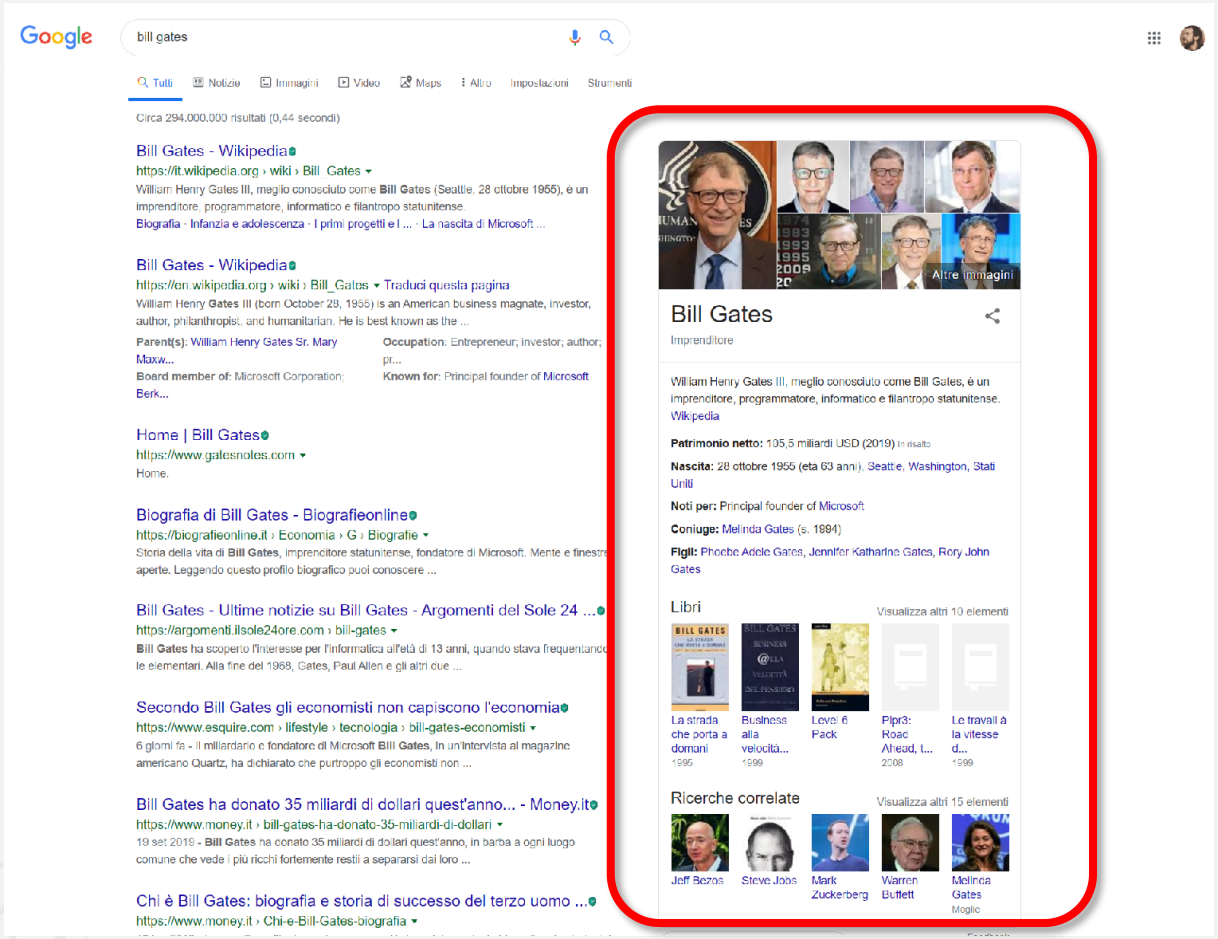
3 - Applicazioni: Motori di Ricerca Semantici

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

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

3 - Applicazioni: Motori di Ricerca Semantici

Introduzione della tecnologia Google Knowledge Graph nelle ricerche su Google:

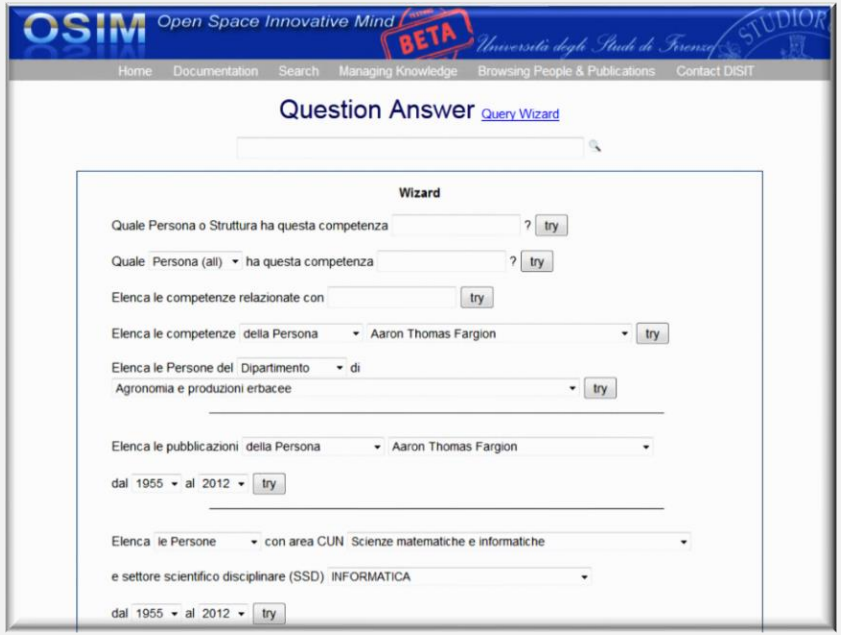


3 - Applicazioni: OSIM (Open Space Innovative Mind)

- **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)***



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 comptenza più genrale rispetto a quella cercata

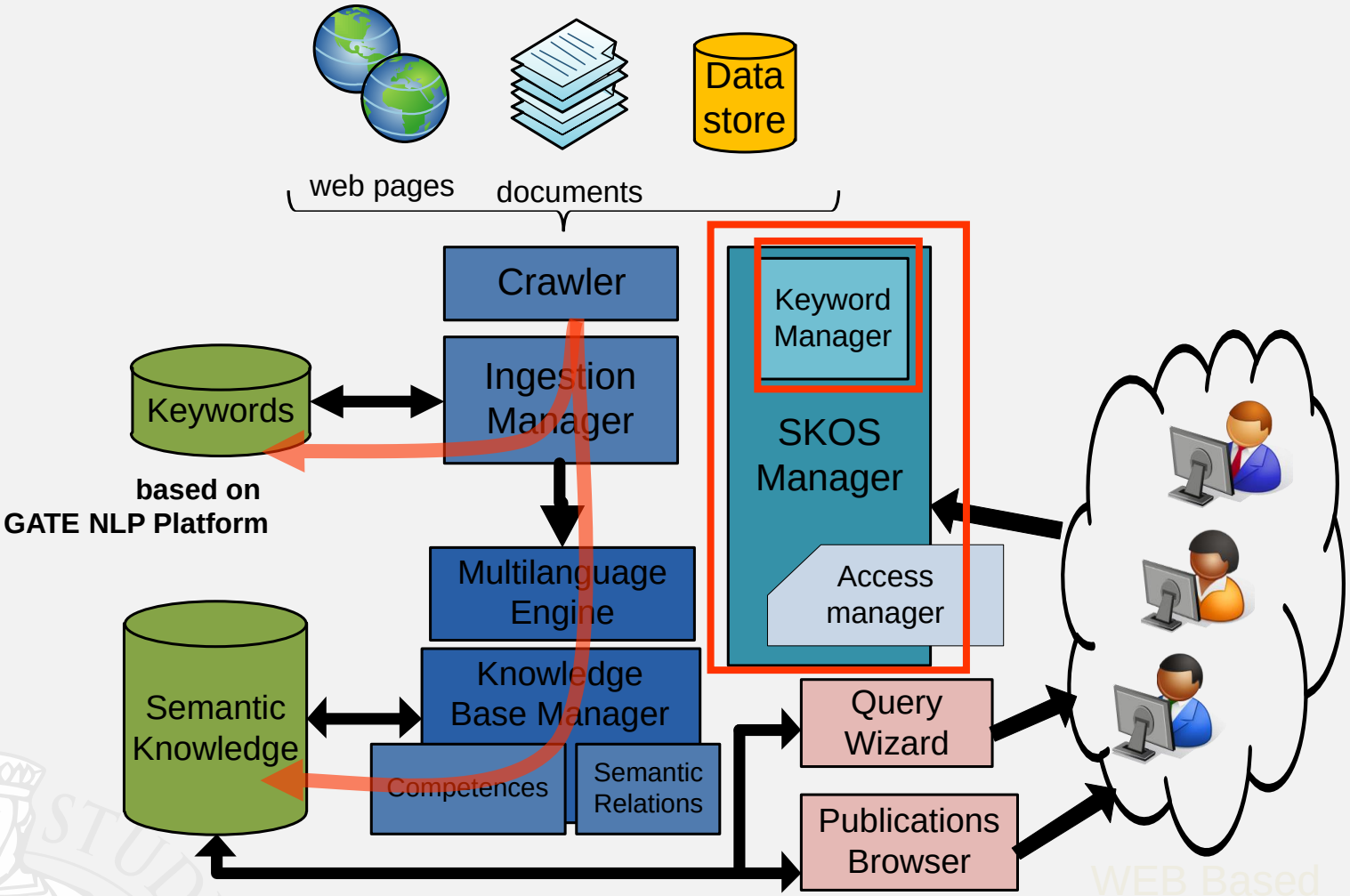
3 - Applicazioni: OSIM (Open Space Innovative Mind)

3 - Applicazioni: OSIM (Open Space Innovative Mind)

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

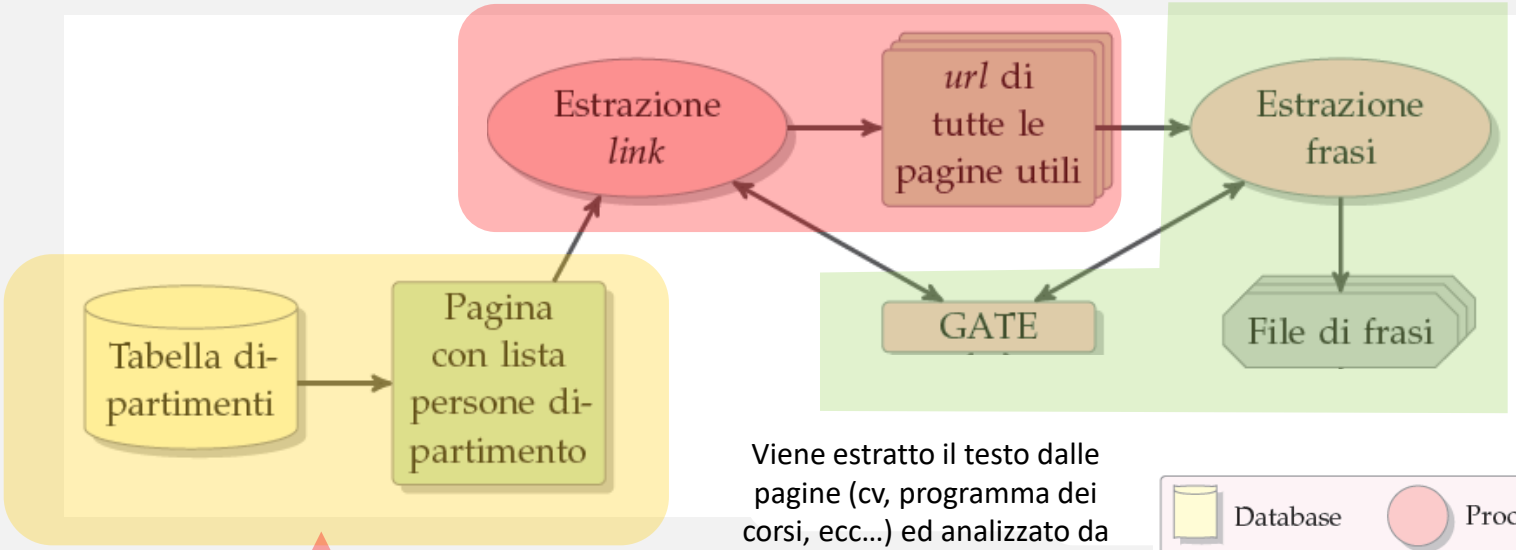
I risultati sono "il corso di Sistemi distribuiti" e la competenza "Sistemi distribuiti"

3 - Applicazioni: OSIM (Open Space Innovative Mind)

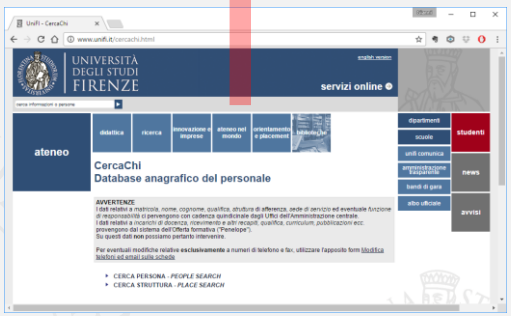


3 - Applicazioni: OSIM (Open Space Innovative Mind)

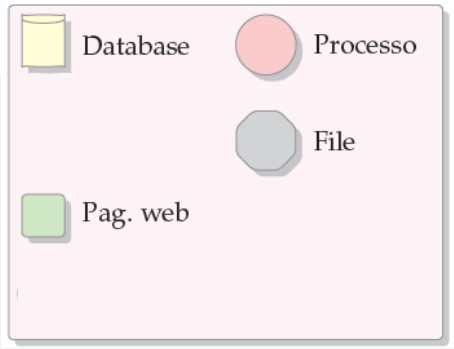
Vettore di URLs. Gli URLs sono scelti dalle apposite sezioni delle pagine, basandosi sui tags HTML delle pagine stesse



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



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



Welcome root Logout OSIM Managing Knowledge HOME

dipartimento di matematica per le decisioni - crawler is running english

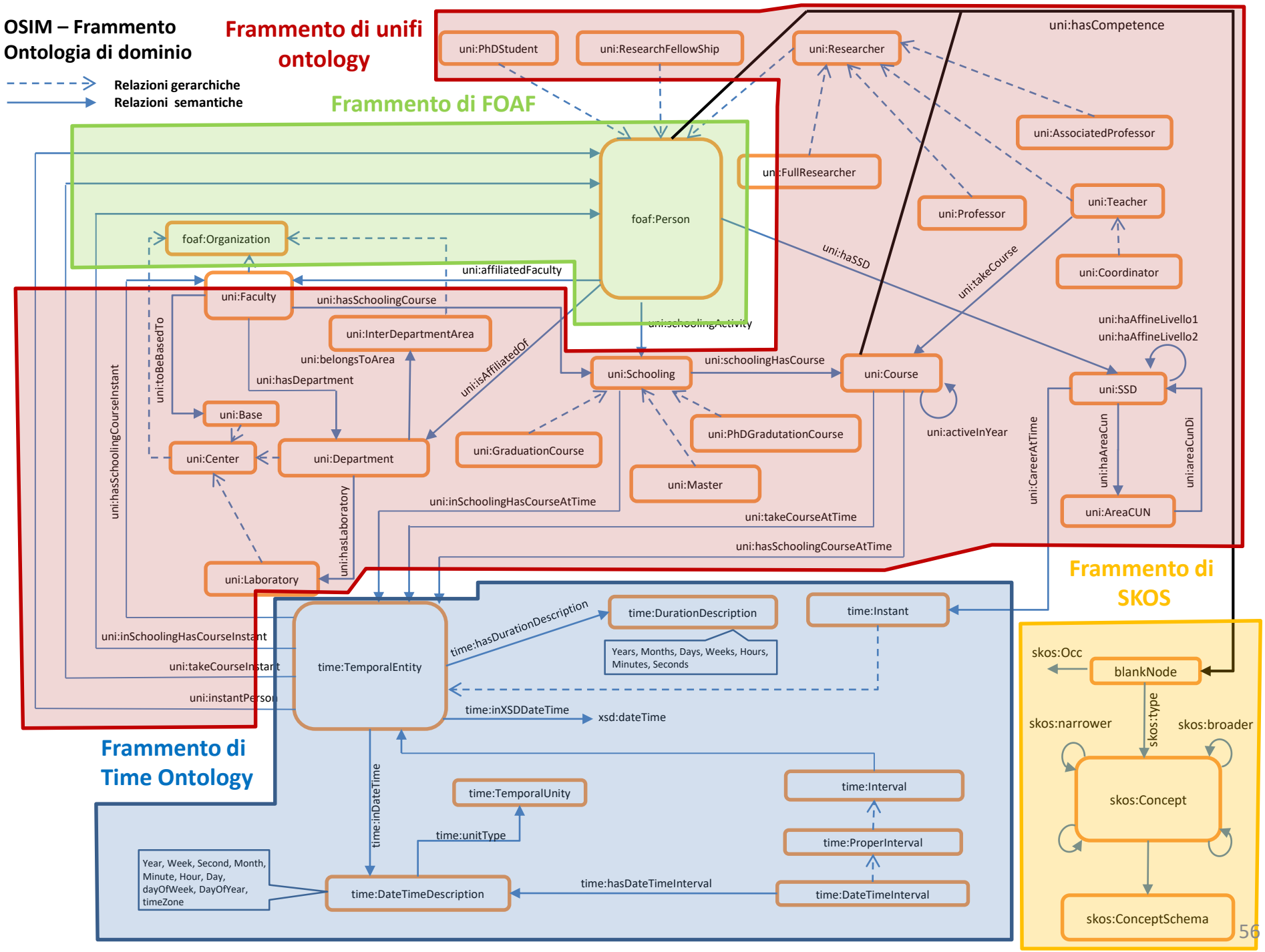
ONTOLOGY MANAGER | KEYS SELECTION | RELATIONS MANAGER

all 20/113 (#2258)

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

Trova: mate Successivo Precedente Evidenzia Maiuscole/minuscole

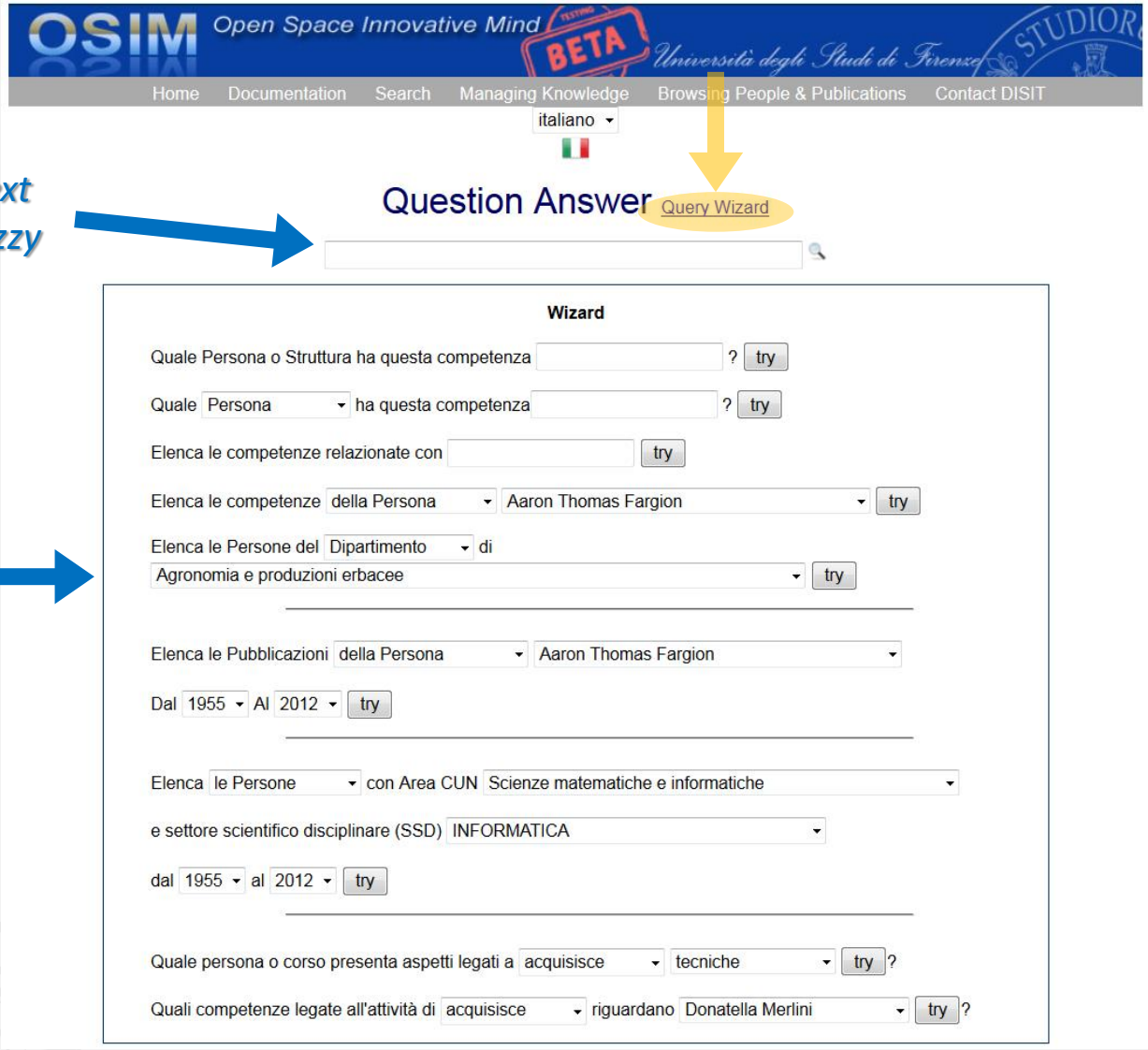

Welcome root Logout
OSIM Managing Knowledge HOME
dipartimento di sistemi ed informatica - crawler is running
english
ONTOLOGY MANAGER
KEYS SELECTION
RELATIONS MANAGER
ISTANCES
filtered by black list
5
Concepts Repository
2
A
able (18)
academic (25)
access (20)
access methods (5)
acm (21)
acm multimedia (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=&cognome=&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=&cognome=&nome=&f=s
7. Related Person:
8. Carlo Colombo (6)
CoSKOSAM


OSIM – Frammento
Ontologia di dominio
Relazioni gerarchiche
Relazioni semantiche
Frammento di unifi
ontology
Frammento di FOAF
uni:PhDStudent
uni:ResearchFellowShip
uni:Researcher
uni:hasCompetence
uni:AssociatedProfessor
uni:FullResearcher
uni:Professor
uni:Teacher
uni:Coordinator
foaf:Person
foaf:Organization
uni:haSSD
uni:takeCourse
uni:affiliatedFaculty
uni:Faculty
uni:hasSchoolingCourse
uni:InterDepartmentArea
uni:schoolingActivity
uni:haAffineLivello1
uni:haAffineLivello2
uni:toBeBasedTo
uni:belongsToArea
uni:isAffiliatedOf
uni:schoolingHasCourse
uni:Schooling
uni:Course
uni:SSD
uni:hasSchoolingCourseInstant
uni:hasDepartment
uni:Base
uni:Center
uni:Department
uni:GraduationCourse
uni:PhDGradutationCourse
uni:activeInYear
uni:CareerAtTime
uni:haAreaCun
uni:areaCunDi
uni:Master
uni:AreaCUN
uni:hasLaboratory
uni:inSchoolingHasCourseAtTime
uni:takeCourseAtTime
uni:hasSchoolingCourseAtTime
uni:Laboratory
Frammento di
SKOS
uni:inSchoolingHasCourseInstant
uni:takeCourseInstant
uni:instantPerson
time:TemporalEntity
time:hasDurationDescription
time:DurationDescription
Years, Months, Days, Weeks, Hours,
Minutes, Seconds
time:Instant
time:inXSDDateTime
xsd:dateTime
Frammento di
Time Ontology
time:inDateTime
time:TemporalUnity
time:unitType
time:Interval
time:ProperInterval
Year, Week, Second, Month,
Minute, Hour, Day,
dayOfWeek, DayOfYear,
timeZone
time:DateTimeDescription
time:hasDateTimeInterval
time:DateTimeInterval
skos:Occ
blankNode
skos:narrower
skos:type
skos:broader
skos:Concept
skos:ConceptSchema

3 - Applicazioni: OSIM (Open Space Innovative Mind)

Ricerca Full Text con Logica Fuzzy

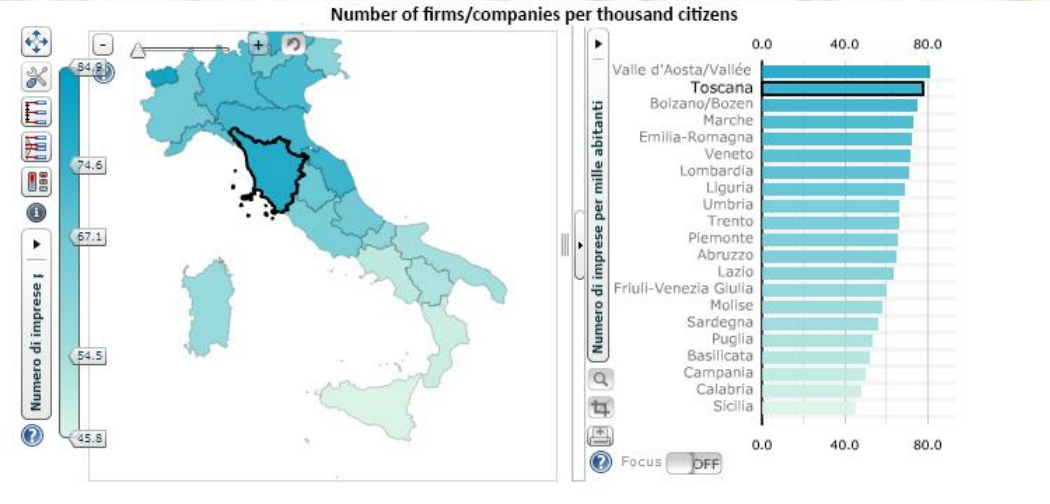
Ricerca Assistita



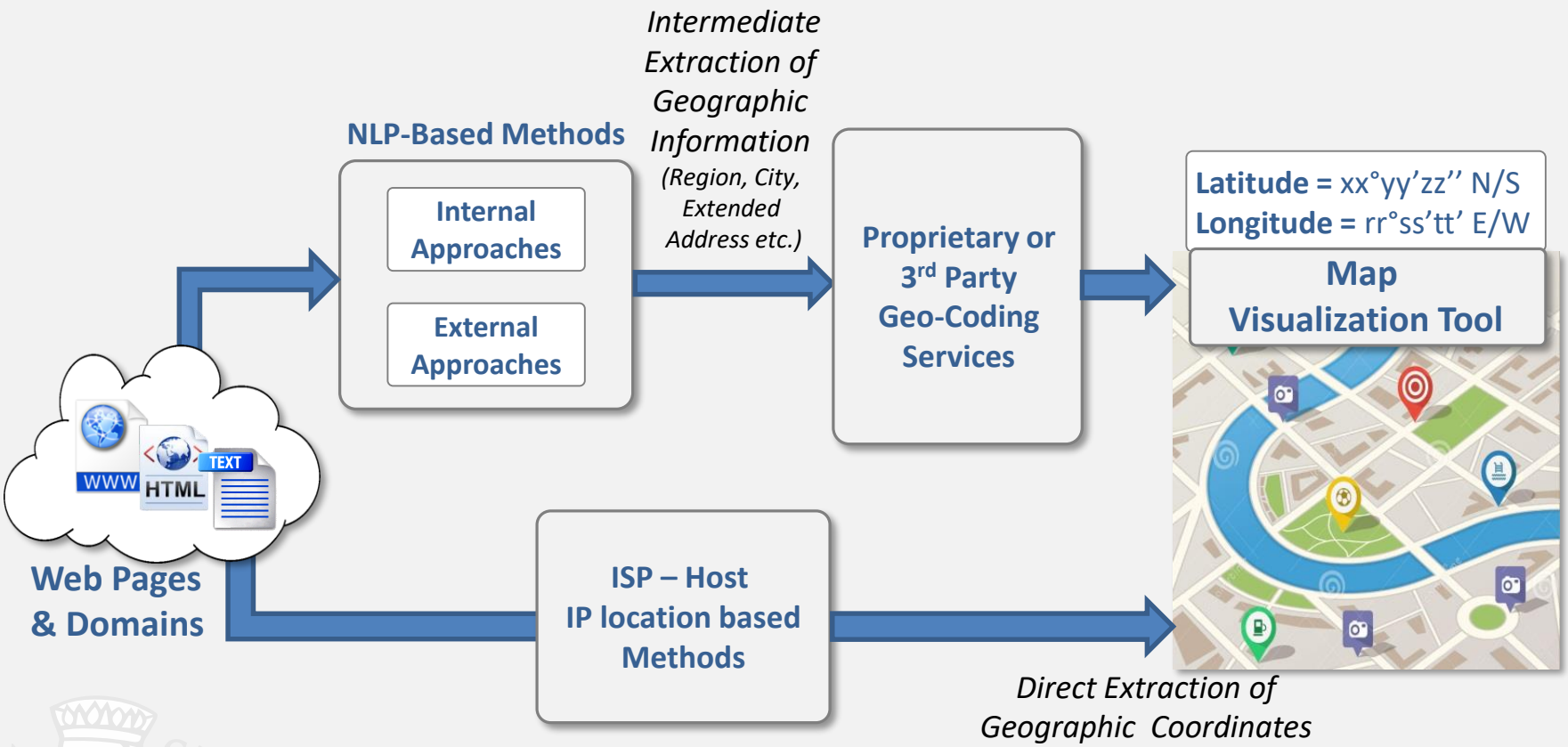
3 - Applicazioni: GeoLocator

- 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

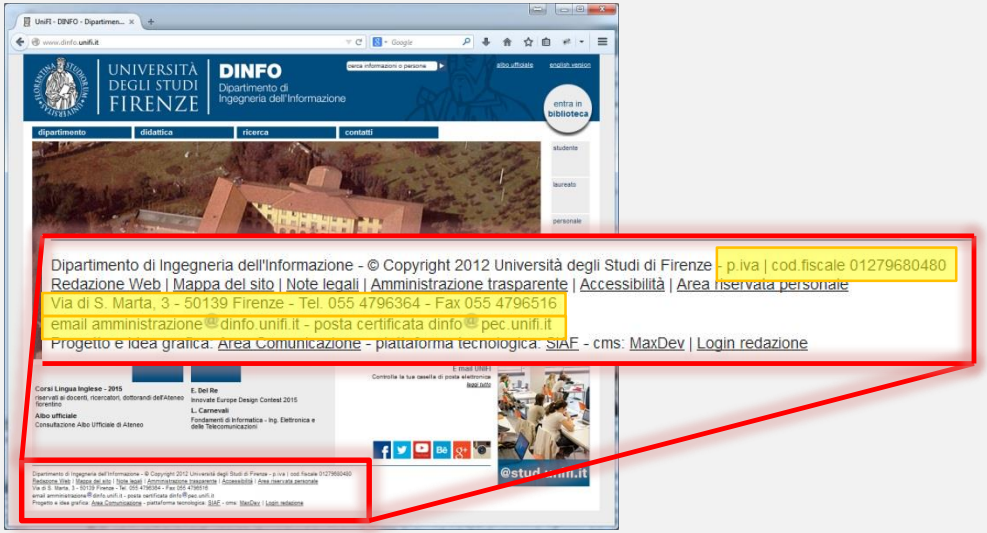


3 - Applicazioni: GeoLocator

GeoLocator [Nesi et al., 2016]

Address Extractor

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

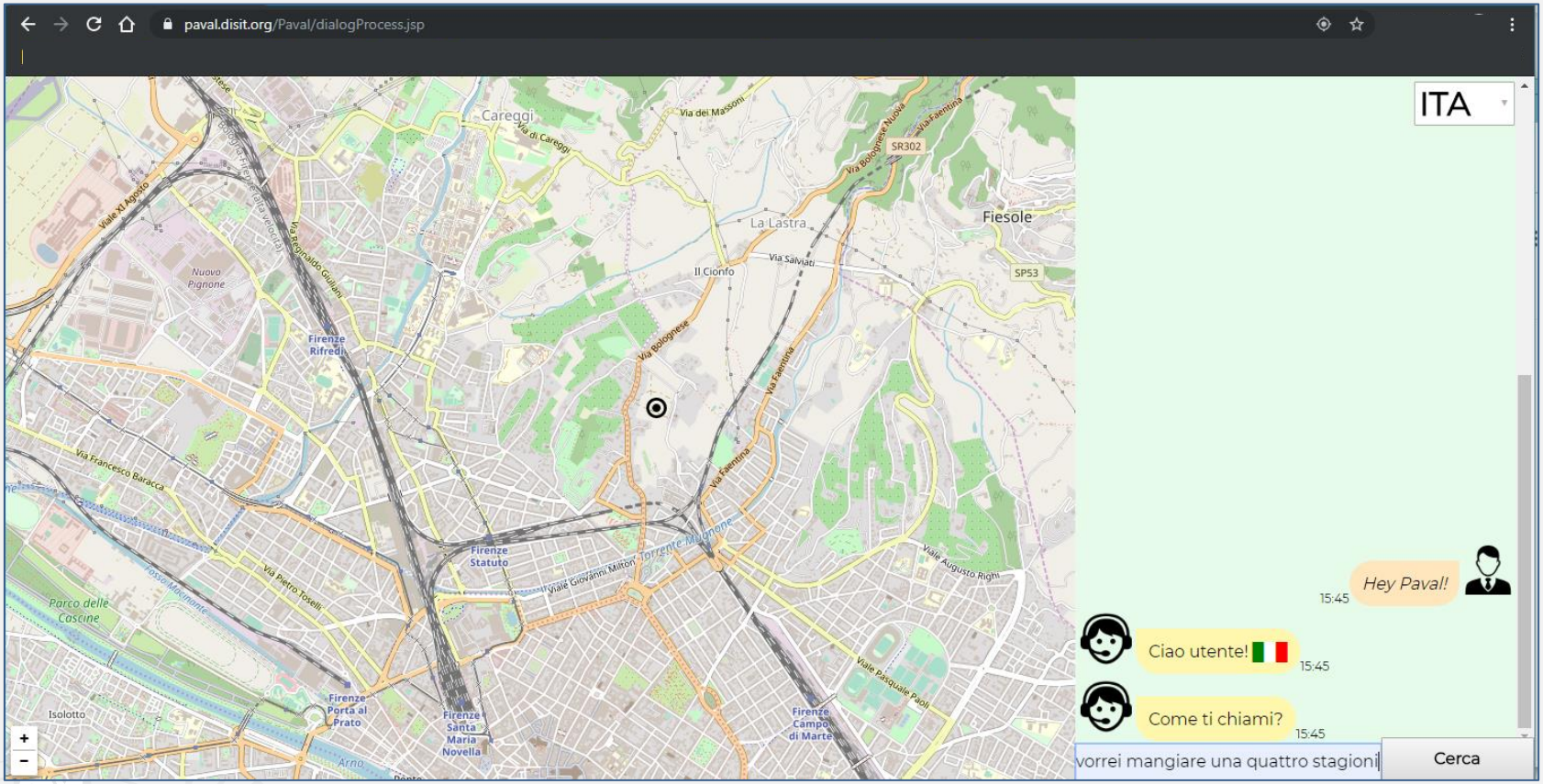


Estrazione degli inidirizzi tramite tecniche di NLP, in particolare Pattern Matching:

- *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].`

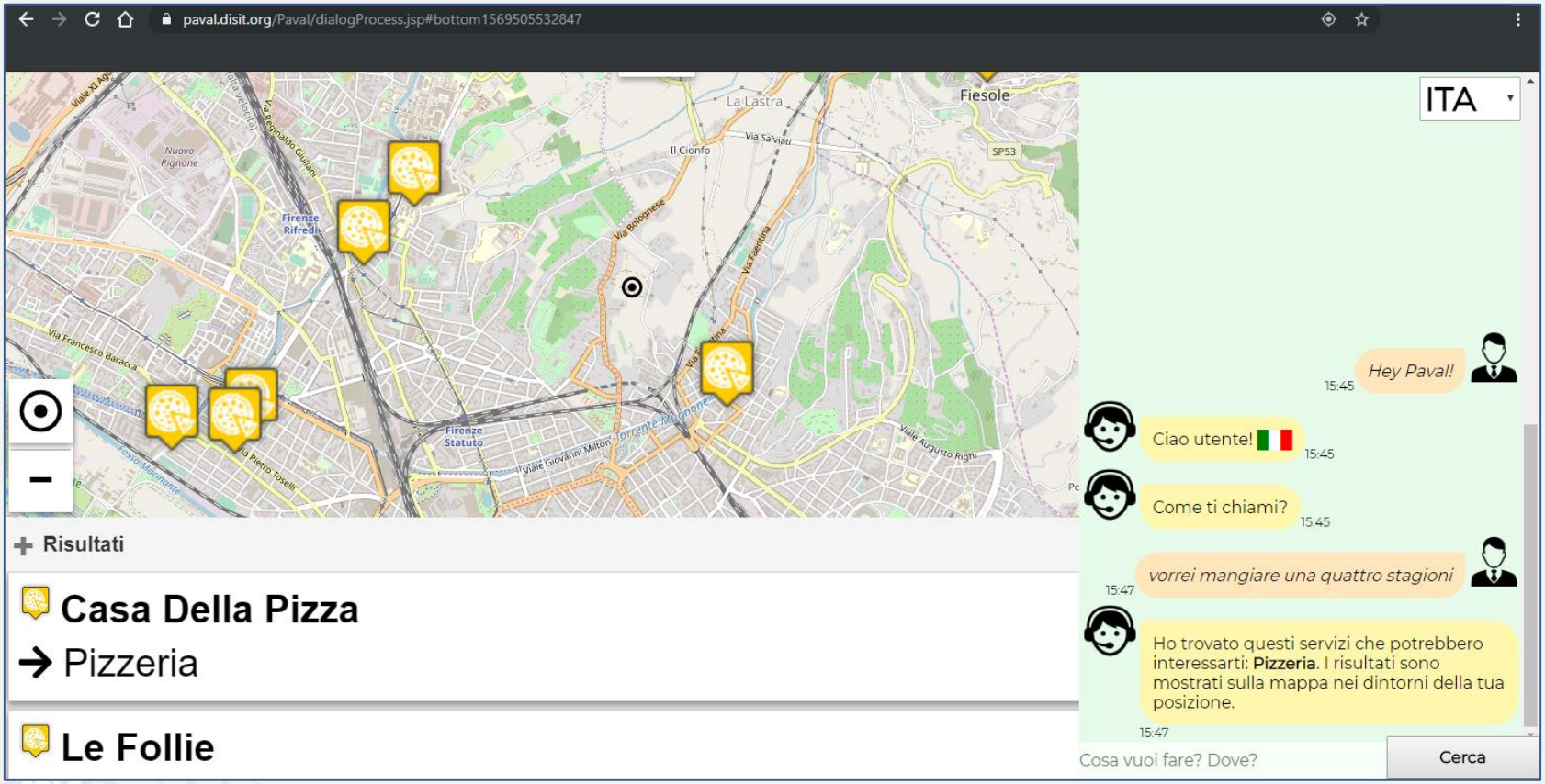
3 - Applicazioni: Paval - *A Location Aware Virtual Personal Assistant -*

Paval [Massai et al., 2019] https://paval.disit.org/



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

Paval [Massai et al., 2019] https://paval.disit.org/

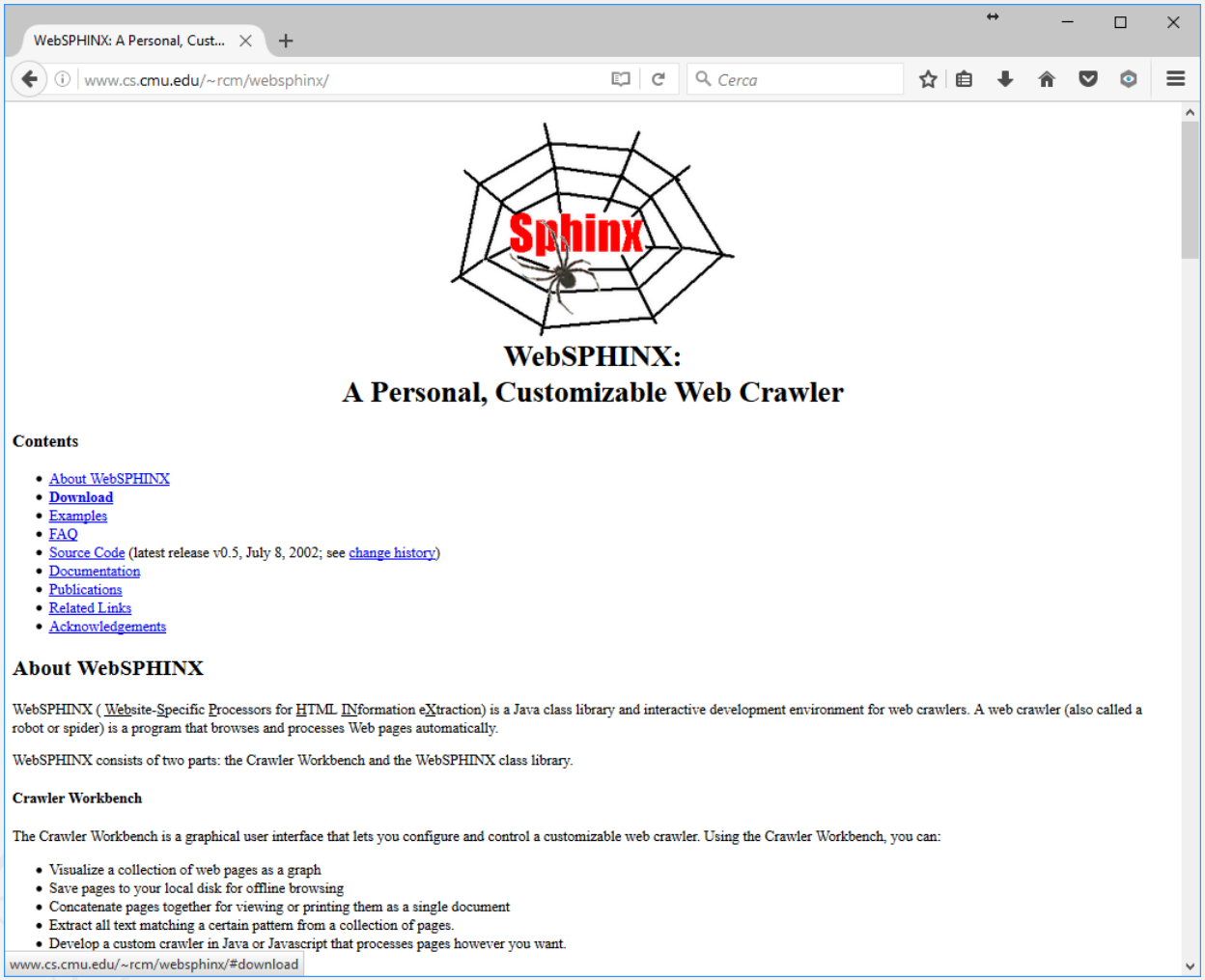


3 - Applicazioni: Paval - *A Location Aware Virtual Personal Assistant* -

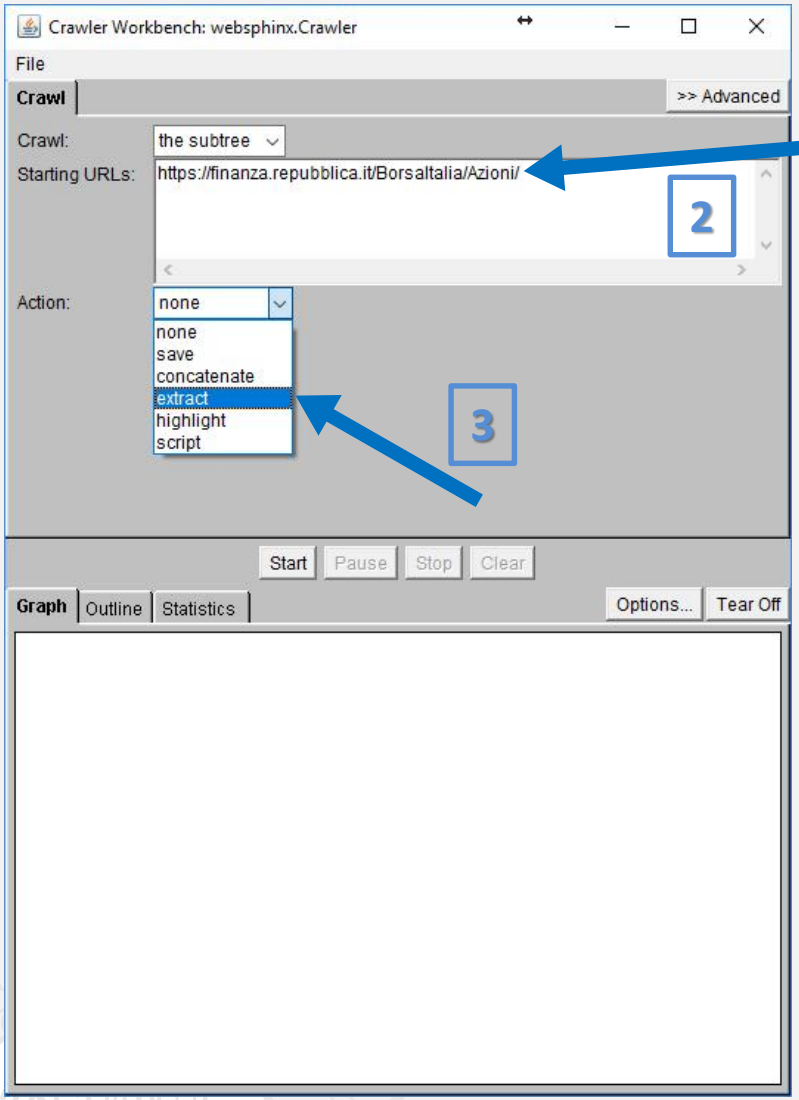
Paval [Massai et al., 2019] https://paval.disit.org/

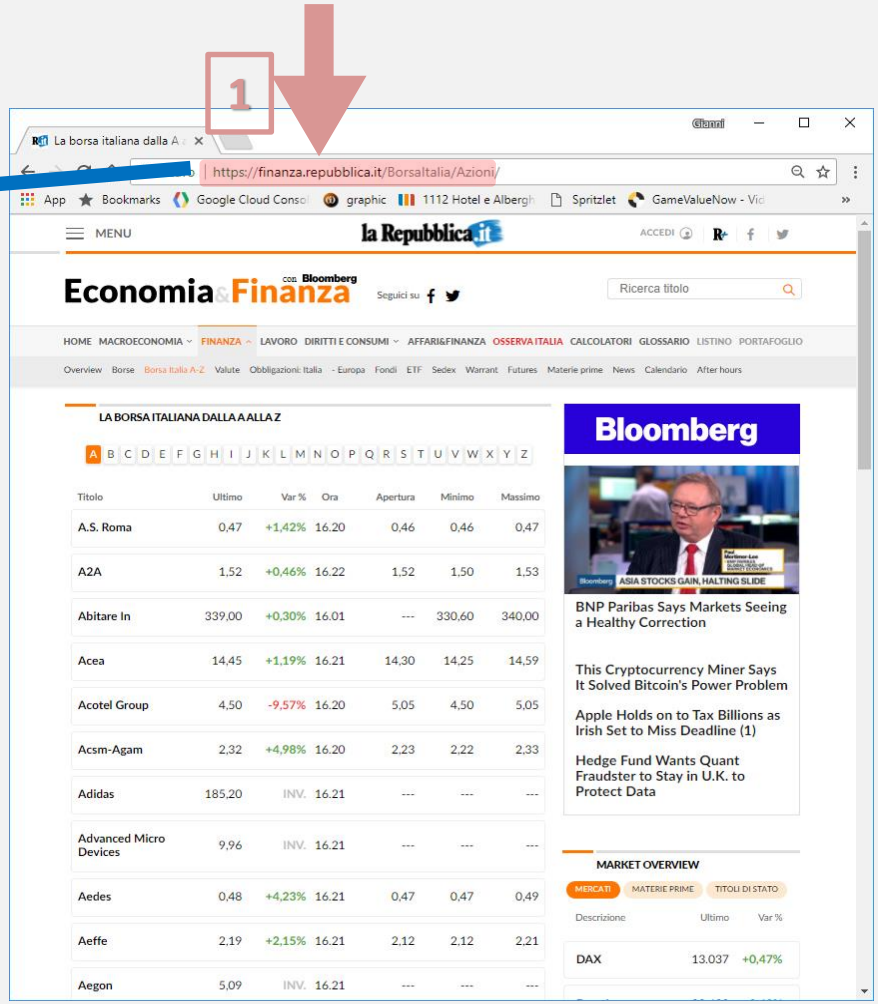
3 - Applicazioni: WebSPHINX

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



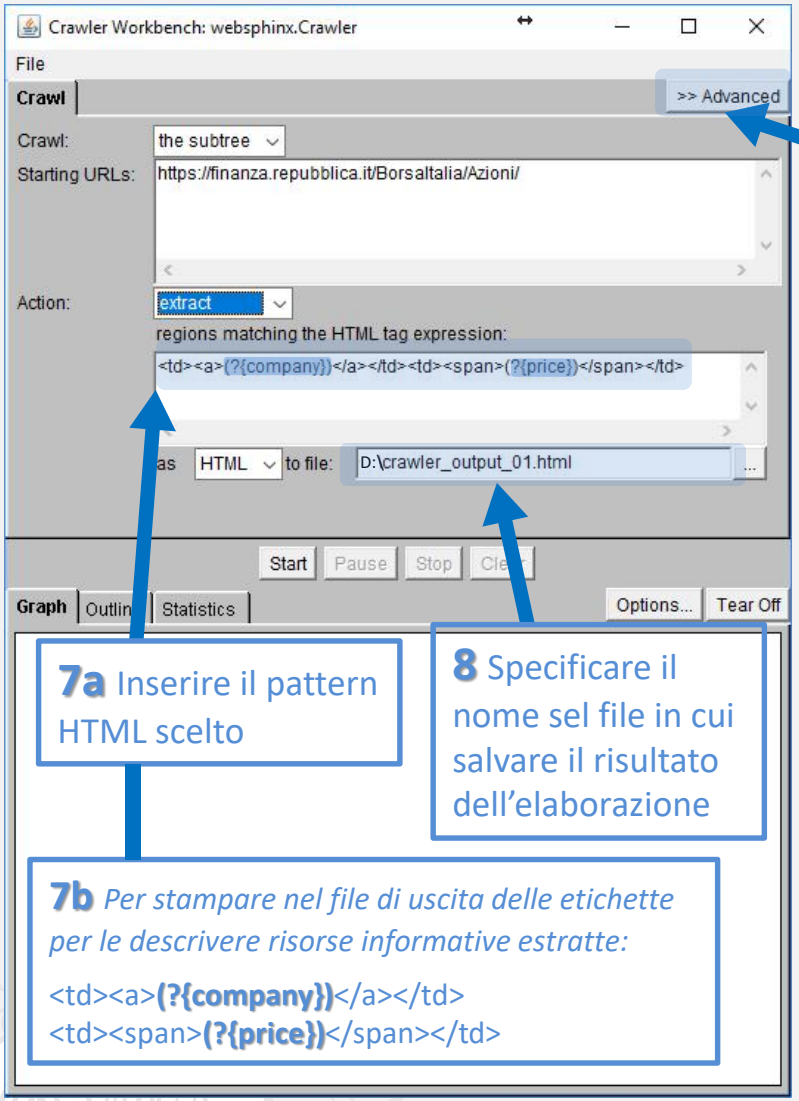
3 - Applicazioni: WebSPHINX





https://finanza.repubblica.it/Borsaltalia/Azioni/

3 - Applicazioni: WebSPHINX



7a Inserire il pattern HTML scelto

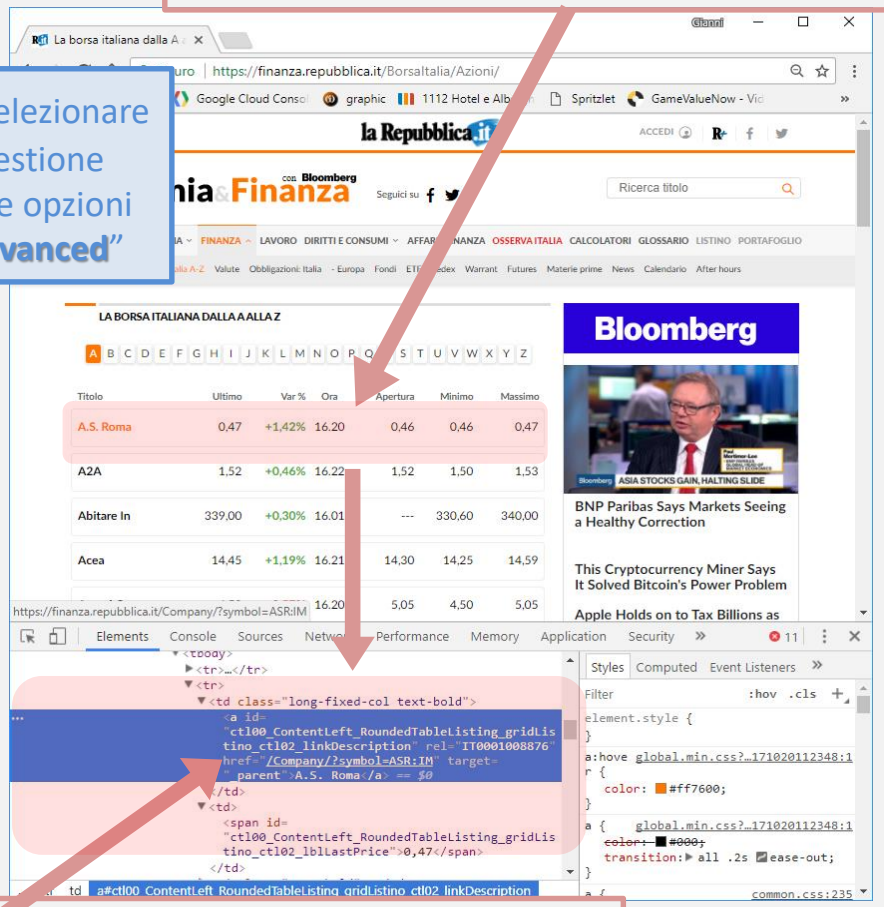
7b *Per stampare nel file di uscita delle etichette per le descrivere risorse informative estratte:*

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

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

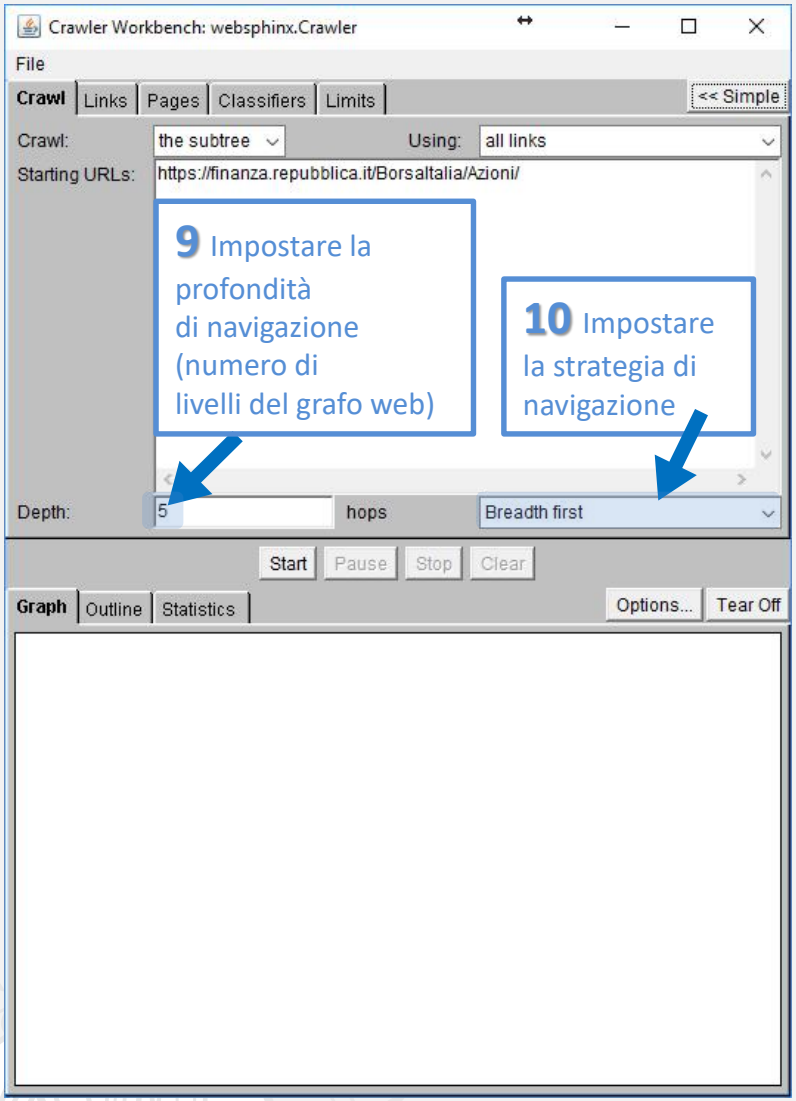
6 Selezionare la gestione delle opzioni "**Advanced**"

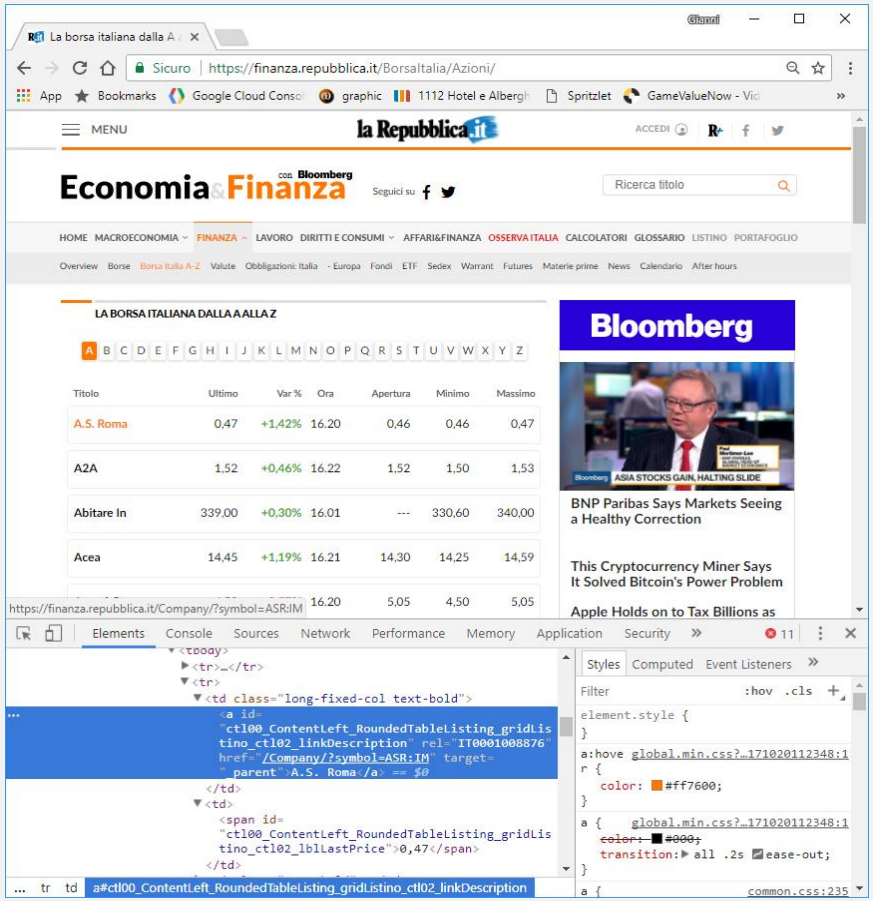
5 Identificare la/le risorsa/e informative da Estrarre nel codice HTML e individuare un pattern HTML significativo con cui addestrare il crawler



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

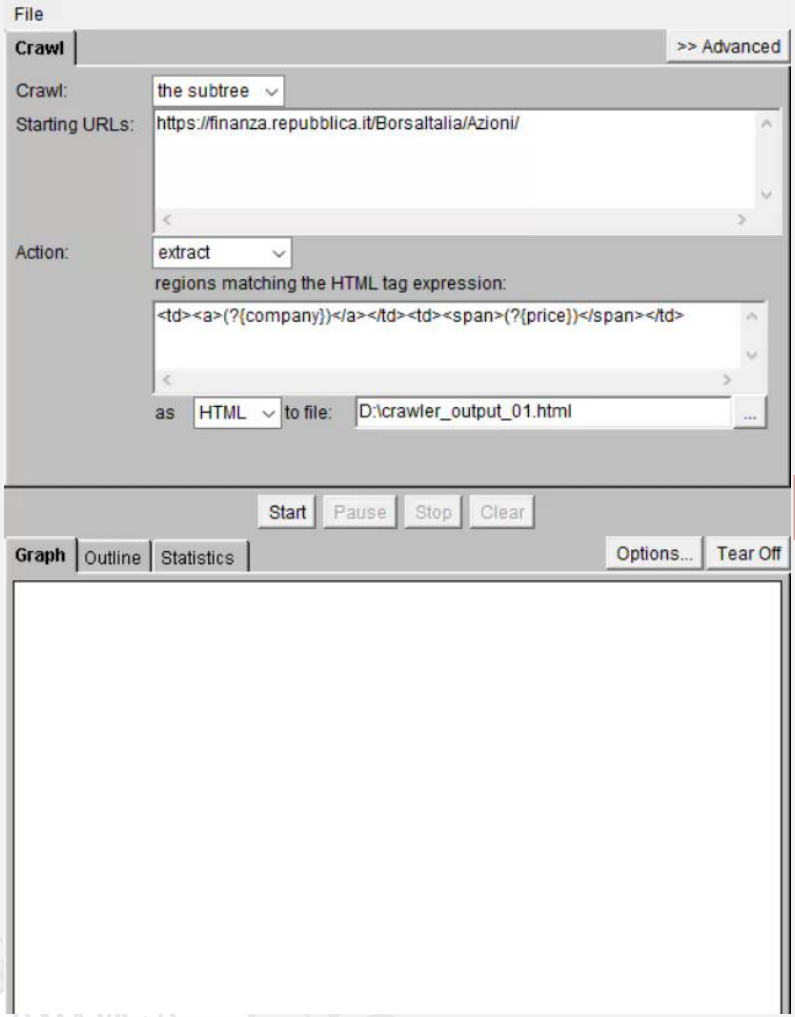
3 - Applicazioni: WebSPHINX

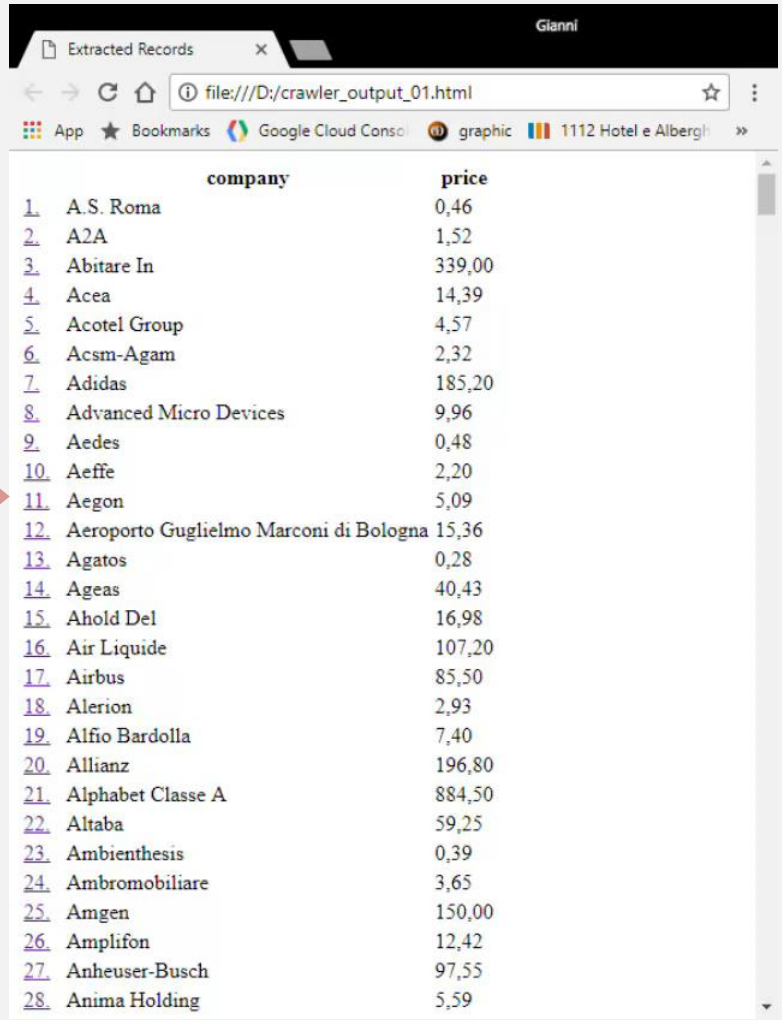




3 - Applicazioni: WebSPHINX

Risultato in Output:

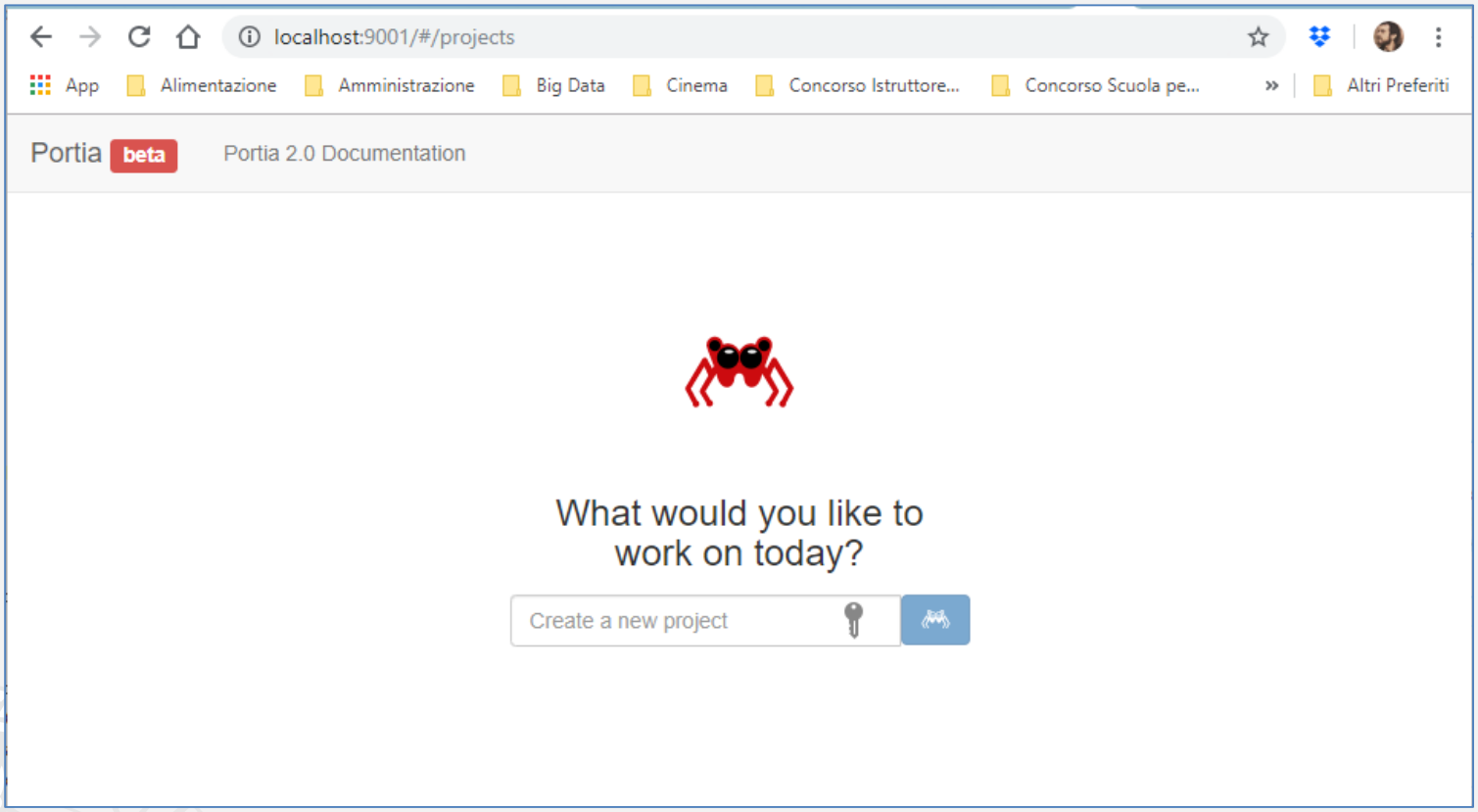




	company	price
1.	A.S. Roma	0,46
2.	A2A	1,52
3.	Abitare In	339,00
4.	Acea	14,39
5.	Acotel Group	4,57
6.	Acsm-Agam	2,32
7.	Adidas	185,20
8.	Advanced Micro Devices	9,96
9.	Aedes	0,48
10.	Aeffe	2,20
11.	Aegon	5,09
12.	Aeroporto Guglielmo Marconi di Bologna	15,36
13.	Agatos	0,28
14.	Ageas	40,43
15.	Ahold 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

3 - Applicazioni: Portia Web Scraper

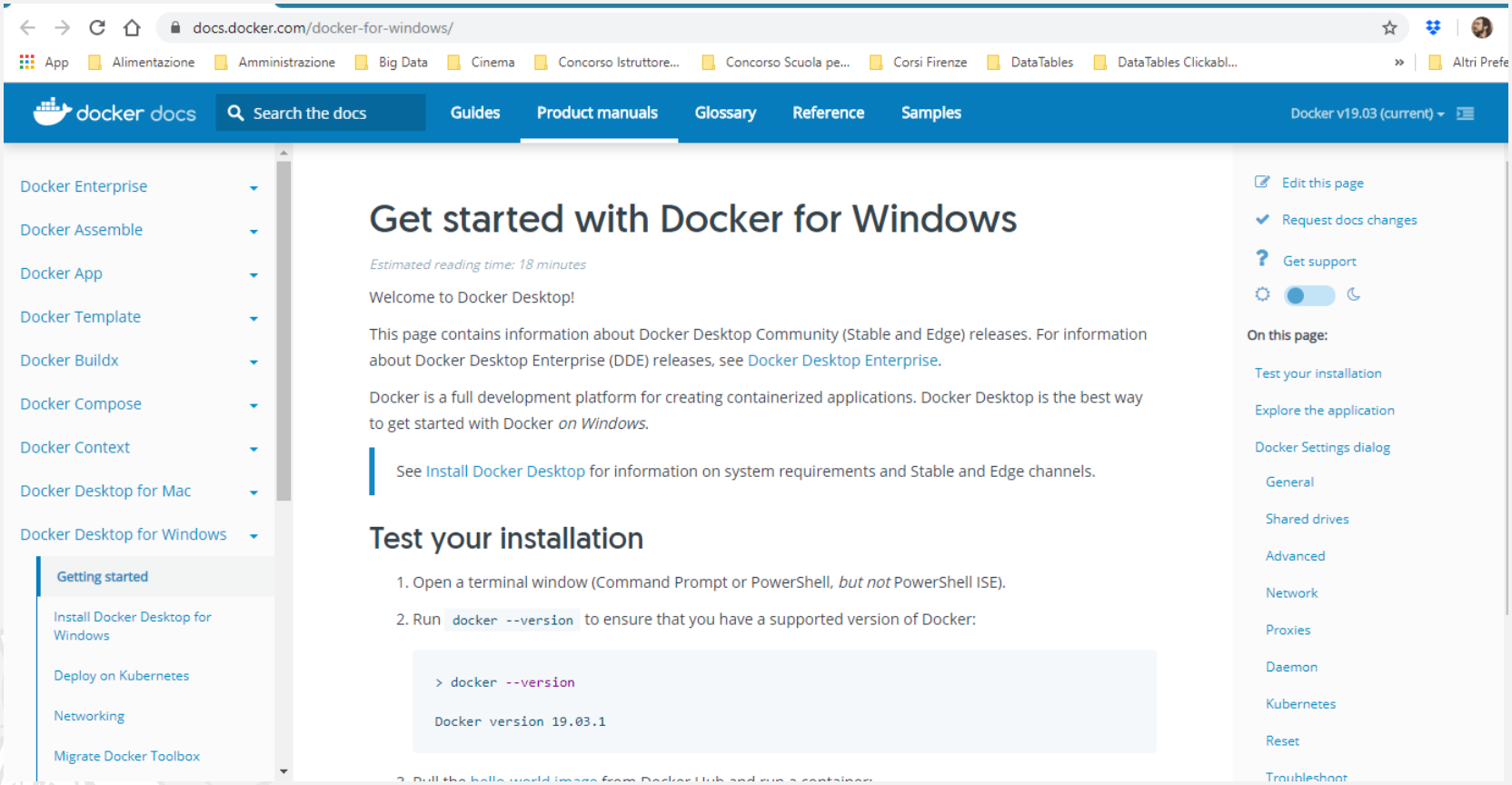
Strumento di crawling web orientato alla configurazione assistita tramite interfaccia grafica



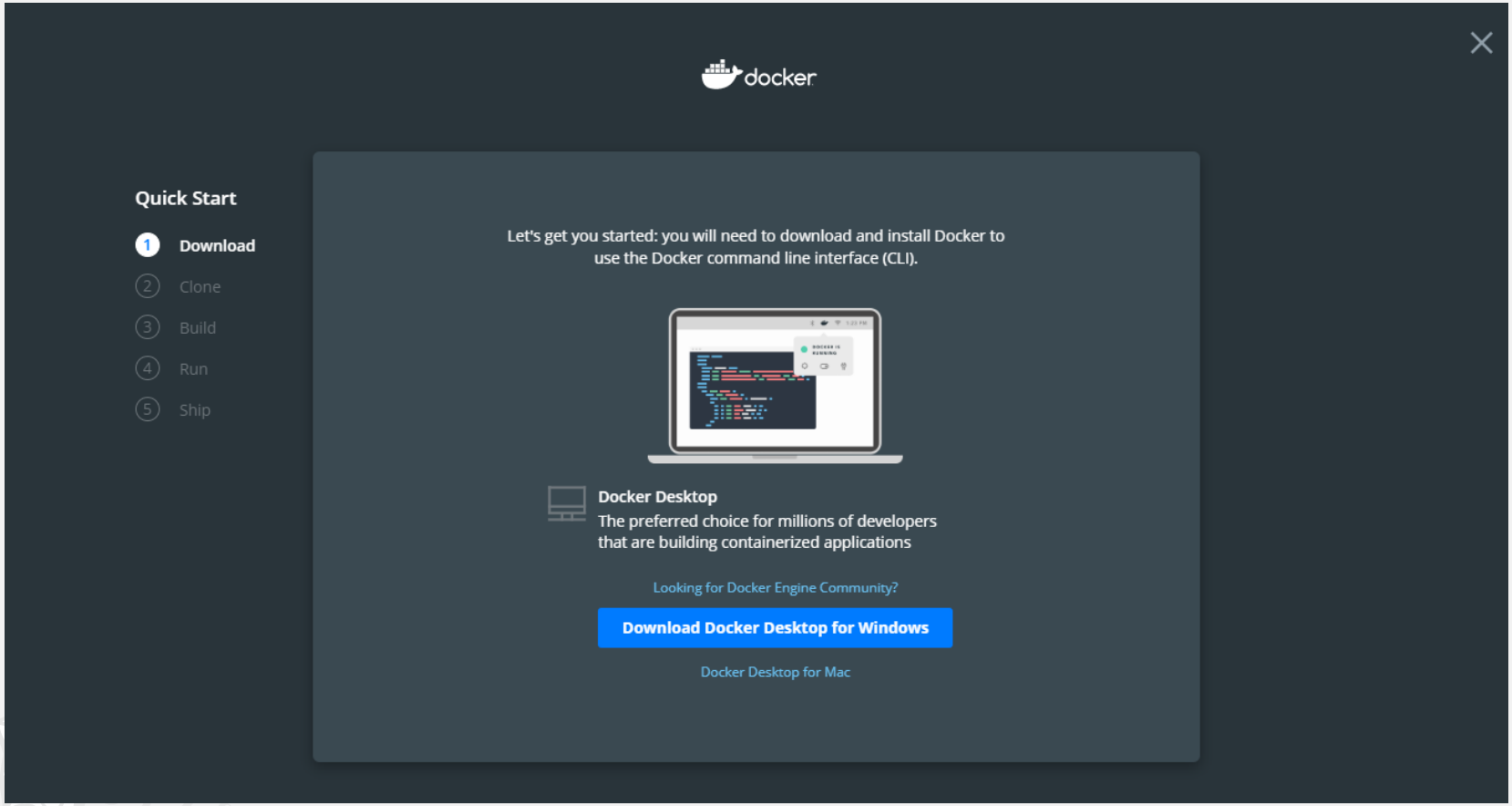
3 - Applicazioni: Portia Web Scraper

Per installare e avviare Portia Web Scraper è necessario installare Docker:

https://docs.docker.com/docker-for-windows/



- Portia Web Scraper – *Scrapinghub su Docker*



- Portia Web Scraper – *Scrapinghub su Docker*

Per avviare il programma su Docker:

- `docker run -i -t --rm -v <PROJECT_FOLDER>:/app/data/projects:rw -p 9001:9001 scrapinghub/portia`

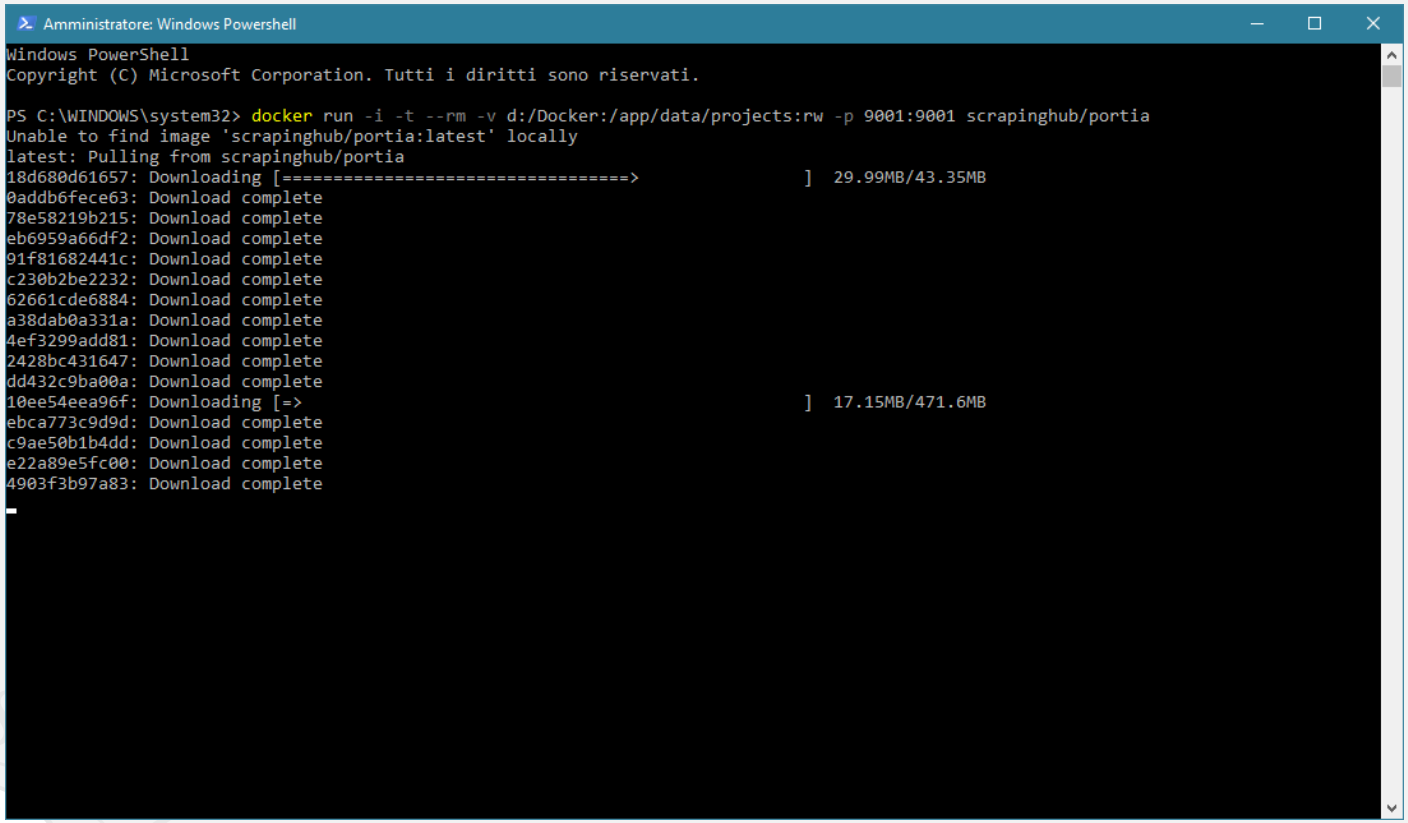
- `docker run -i -t --rm -v d:/Docker:/app/data/projects:rw -p 9001:9001 scrapinghub/portia`

Per estrarre i dati dal sito web di interesse, una volta configurato lo spider:

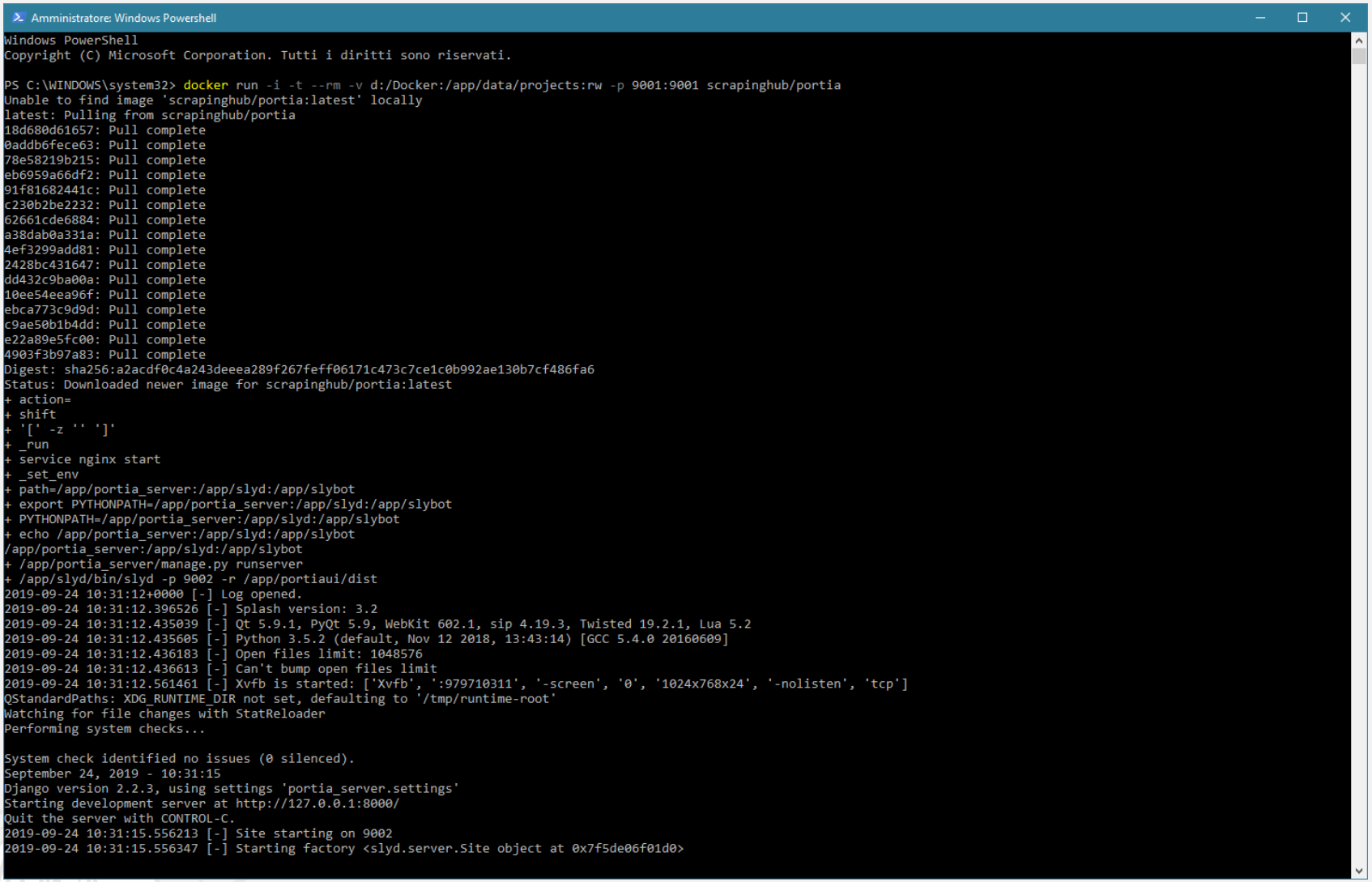
- `docker run -i -t --rm -v <PROJECT_FOLDER>:/app/data/projects:rw -v <OUTPUT_FOLDER>:/mnt:rw -p 9002:9001 scrapinghub/portia portiacrawl /app/data/projects/<PROJECT_NAME> <SPIDER_NAME> -t <OUTPUT_FORMAT> -o /mnt/<SPIDER_NAME>.<FILE_EXT_OUT_FORMAT>`

- Portia Web Scraper – *Scrapinghub su Docker*

- `docker run -i -t --rm -v d:/Docker:/app/data/projects:rw -p 9001:9001 scrapinghub/portia`

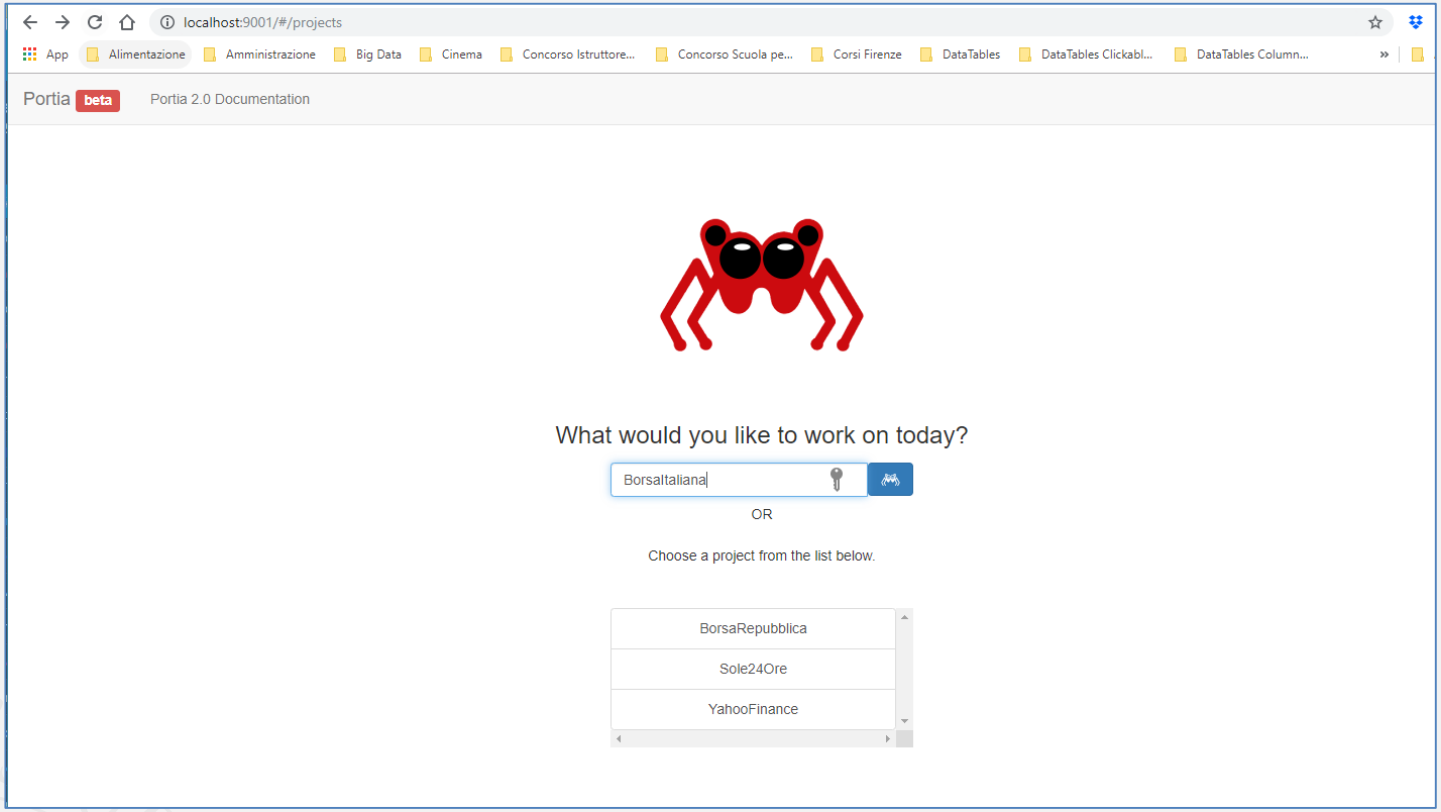


- Portia Web Scraper – *Scrapinghub su Docker*



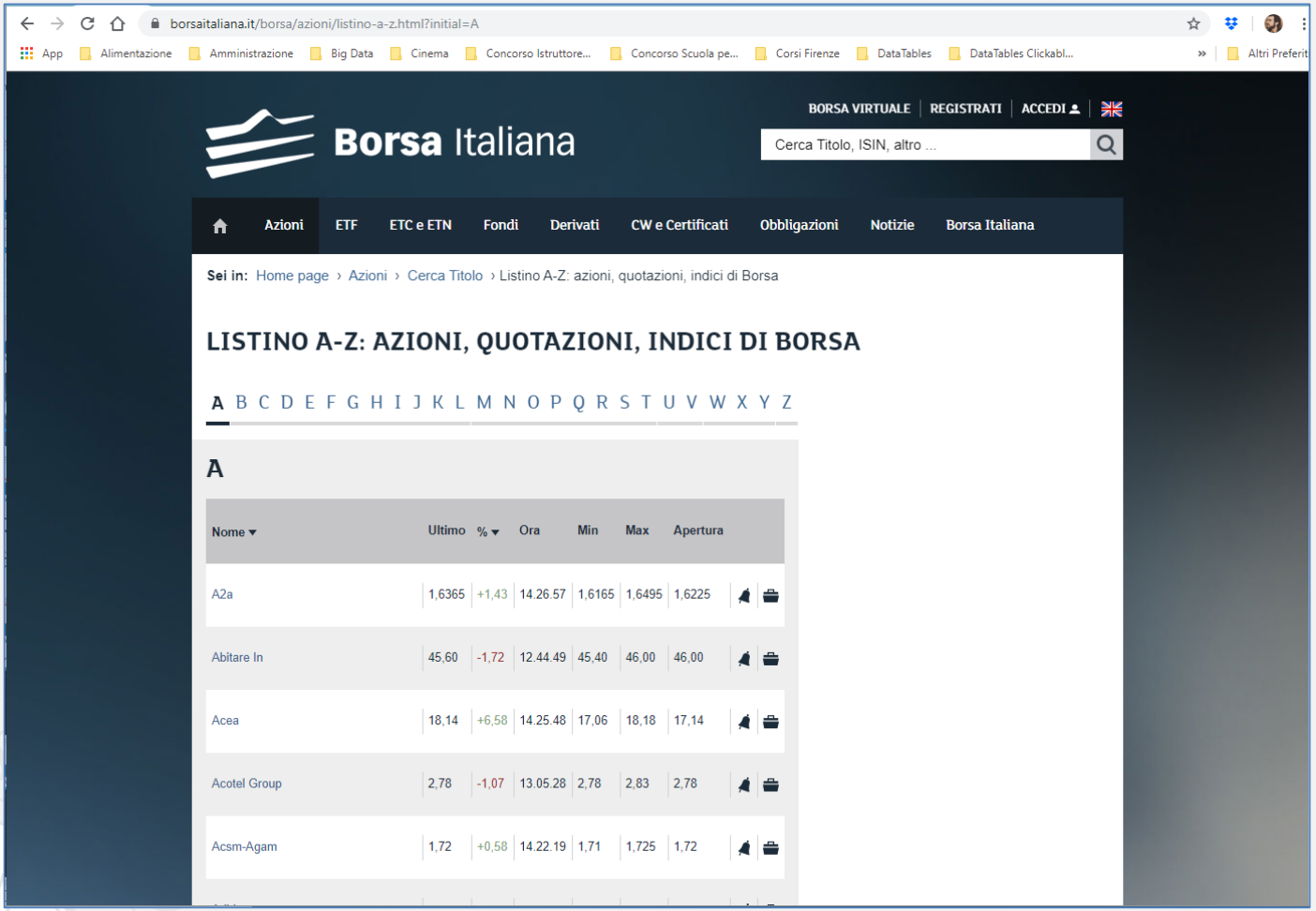
- Portia Web Scraper – *Creare un progetto*

- Portia sarà quindi accessibile localmente aprendo il browser web al seguente indirizzo: `http://localhost:9001`



- Portia Web Scraper – *Creare un progetto*

- Obiettivo: estrarre dati finanziari da pagine web

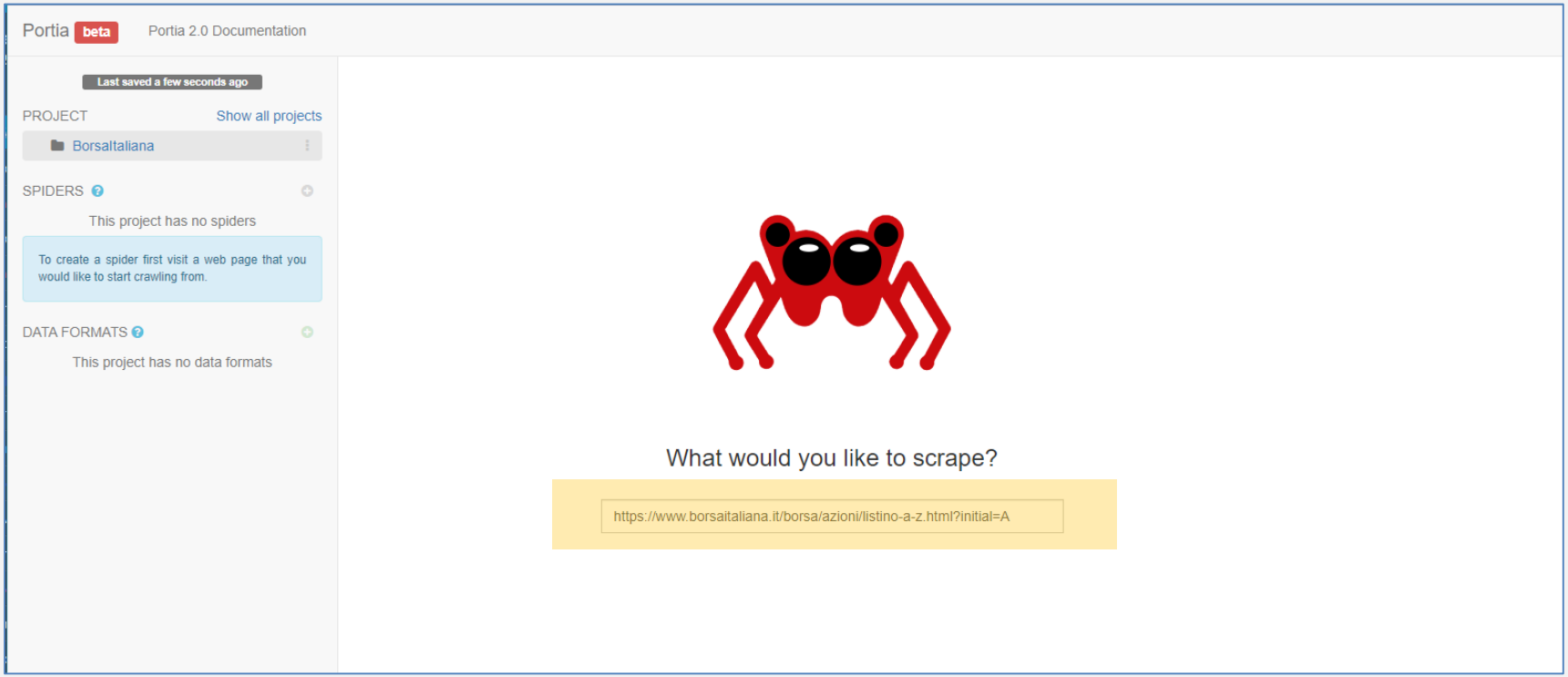


- Portia Web Scraper – *Creare un progetto*

- Verificare sempre eventuali disclaimer del sito sorgente dei dati!

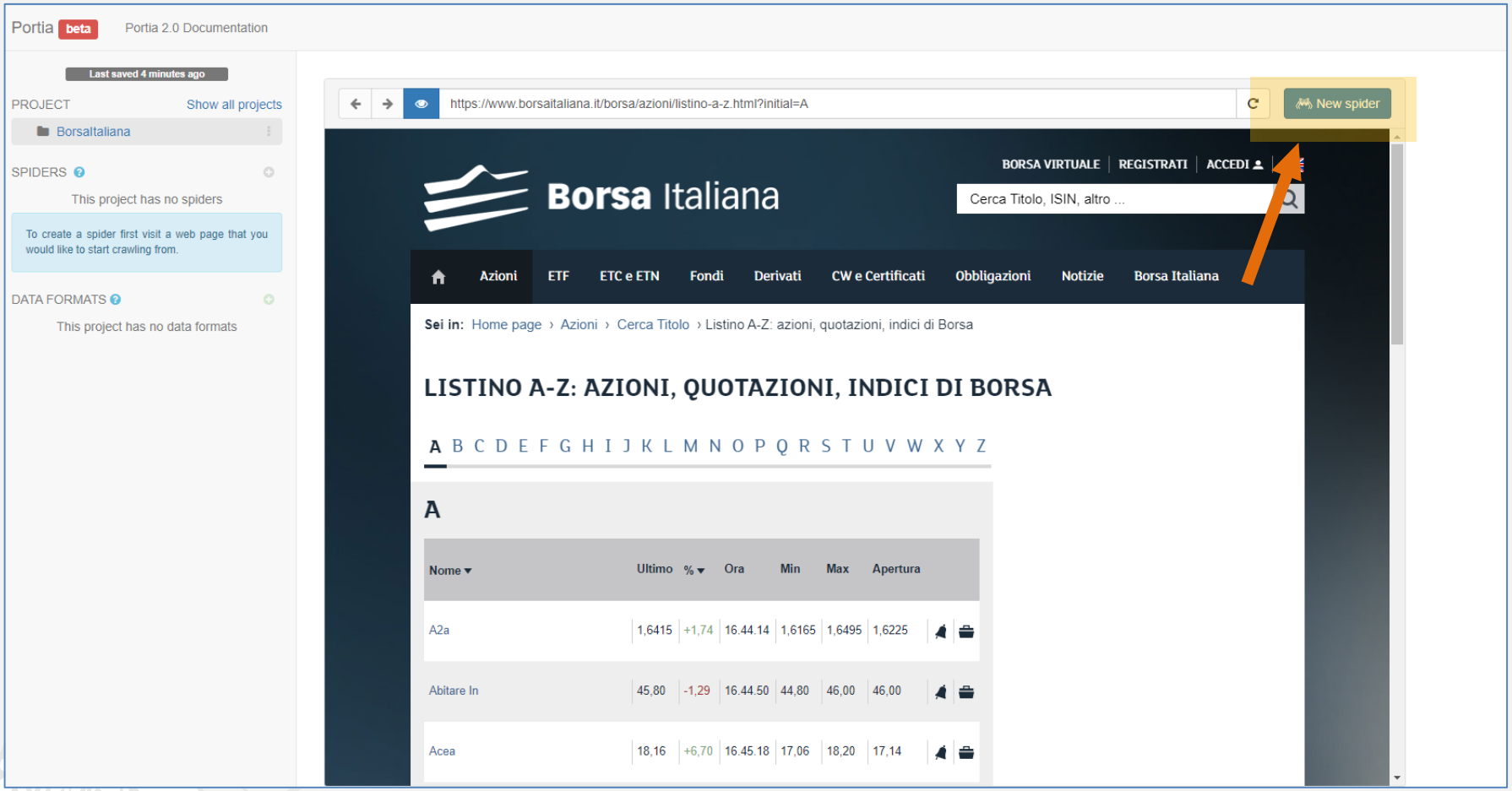
- Portia Web Scraper – *Configurare lo Spider*

- Inserire l'URL di partenza;



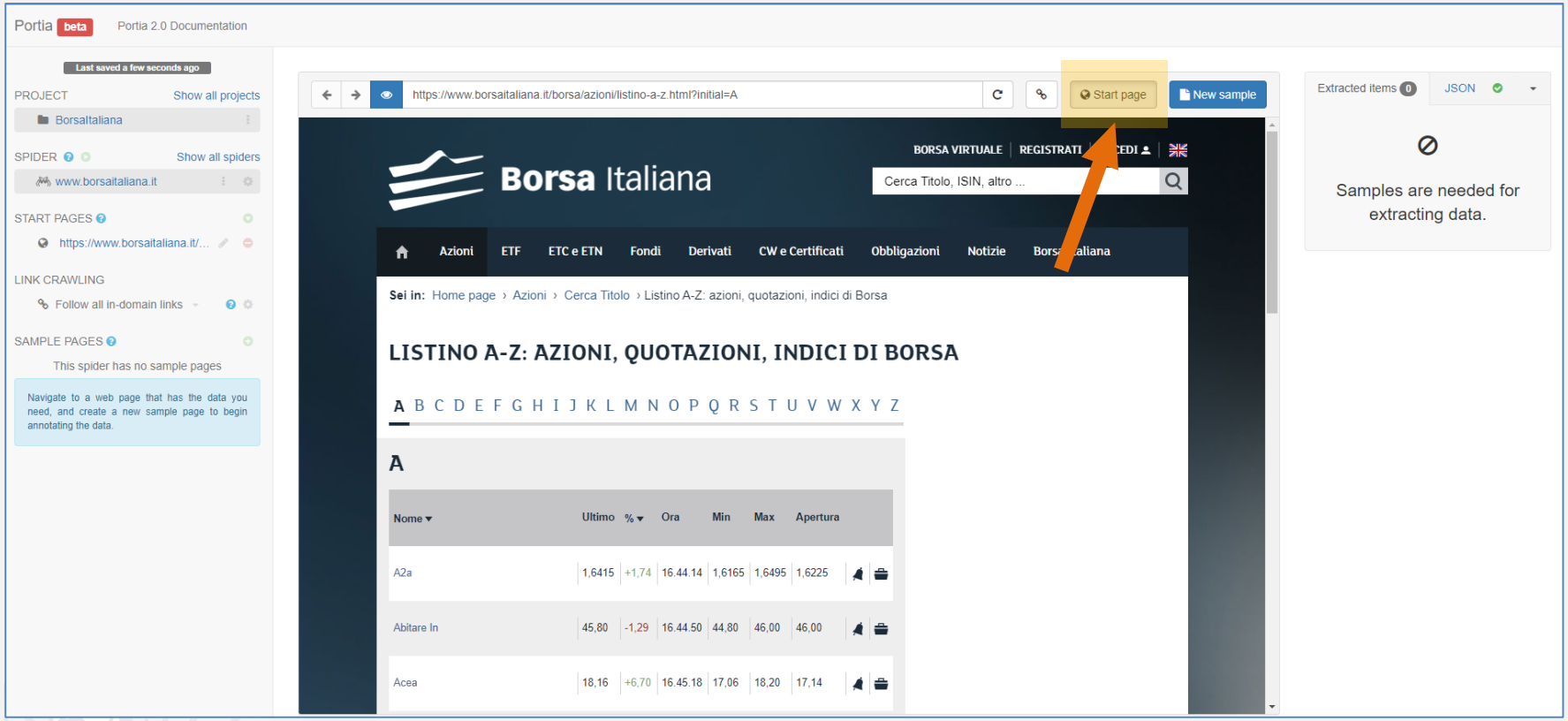
- Portia Web Scraper – *Configurare lo Spider*

- Portia offre una modalità di visualizzazione/navigazione della pagina corrispondente all'URL inserito.



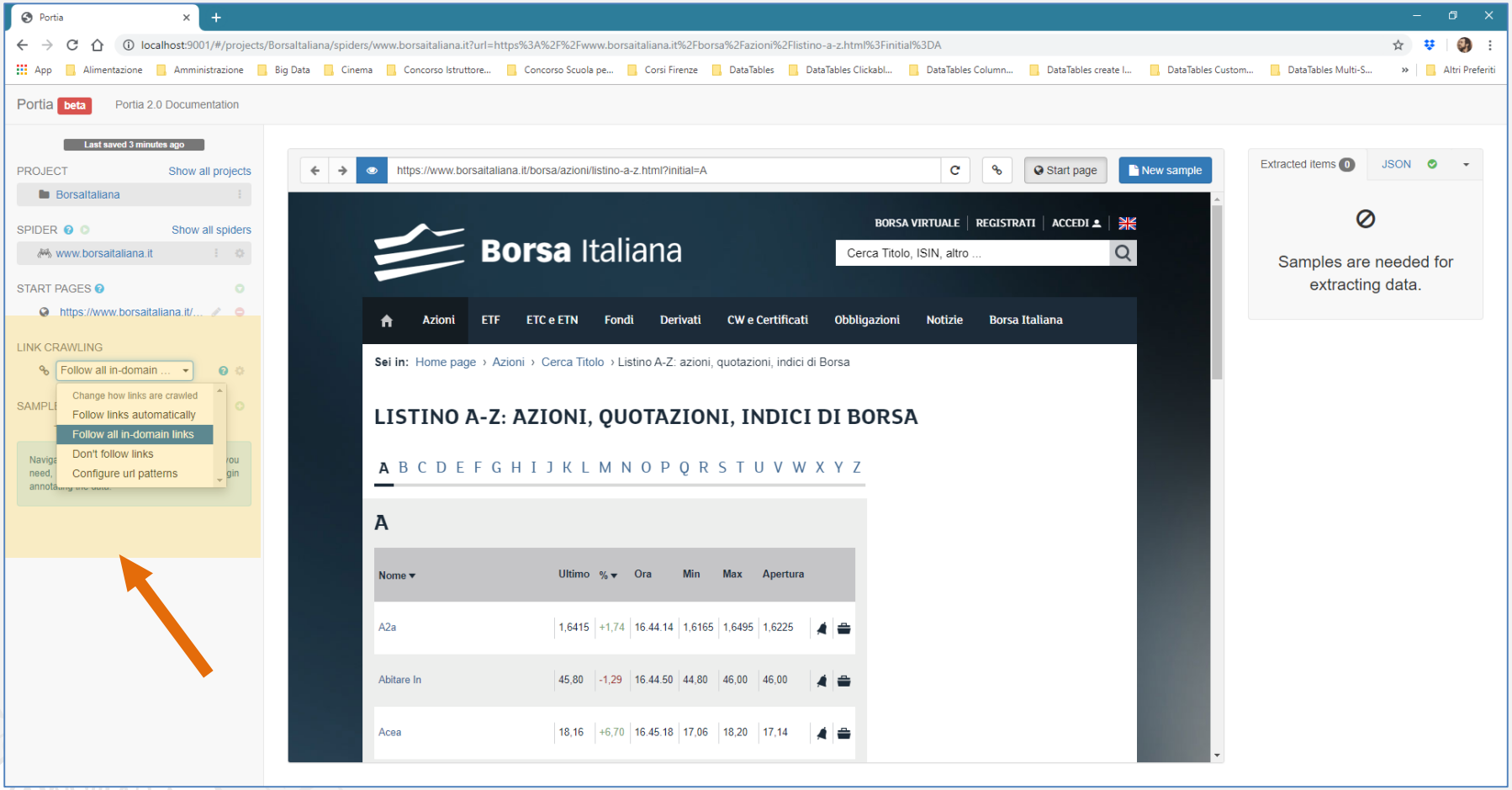
- Portia Web Scraper – *Configurare lo Spider*

- E' ora possibile associare inizializzare un nuovo Spider e associare la pagina inserita come pagina iniziale (Start Page).



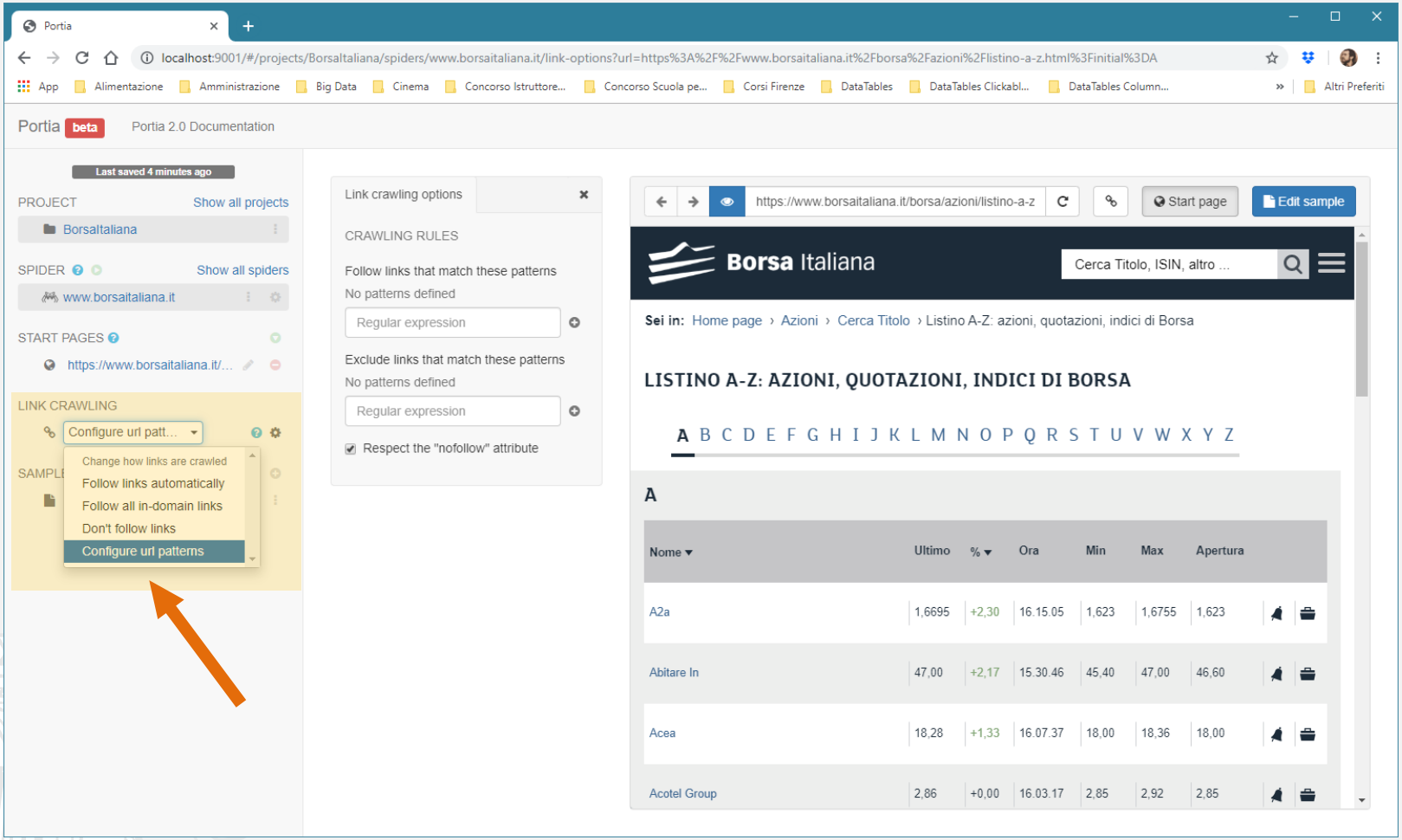
- Portia Web Scraper – *Configurare lo Spider*

- Si possono definire varie impostazioni, tra cui la politica di navigazione delle pagine:



- Portia Web Scraper – *Configurare lo Spider*

- Si possono definire varie impostazioni, tra cui la politica di navigazione delle pagine:

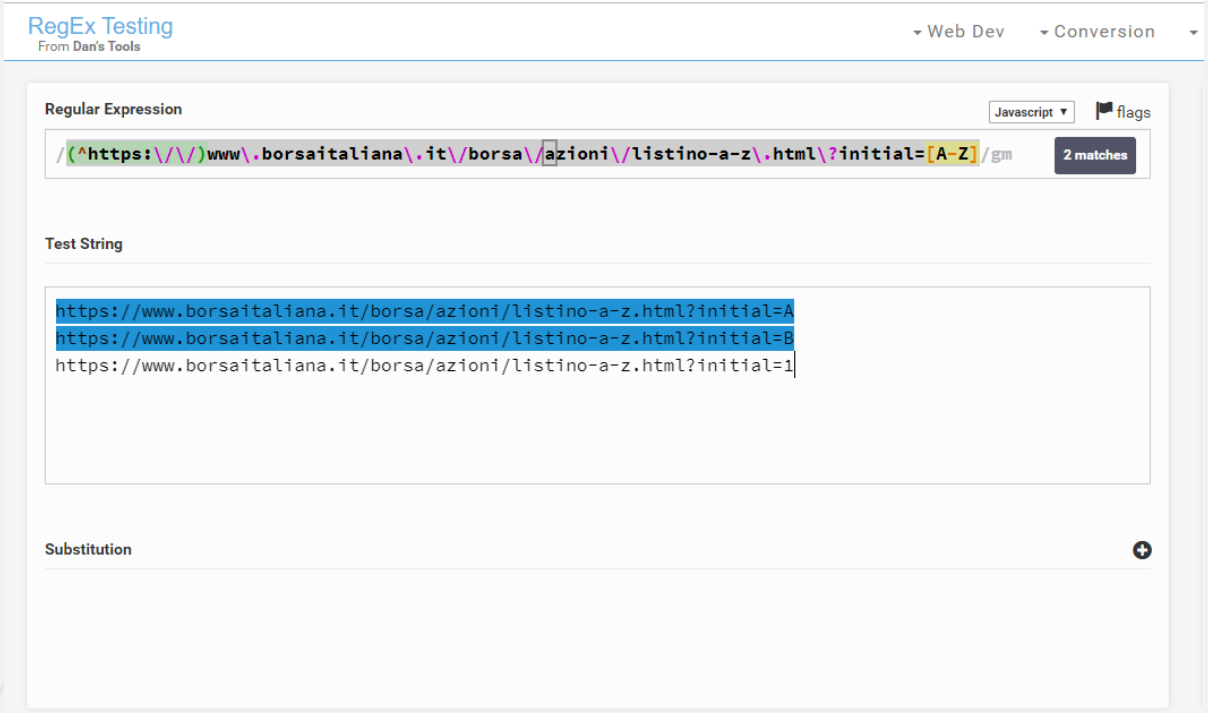


- Portia Web Scraper – *Configurare lo Spider*

- Molto comodo l'utilizzo delle ***Espressioni Regolari*** per identificare pattern di URLs:

https://www.borsaitaliana.it/borsa/azioni/listino-a-z.html?initial=A

(^https:\/\/)www\.borsaitaliana\.it\/borsa\/azioni\/listino-a-z\.html\?initial=[A-Z]

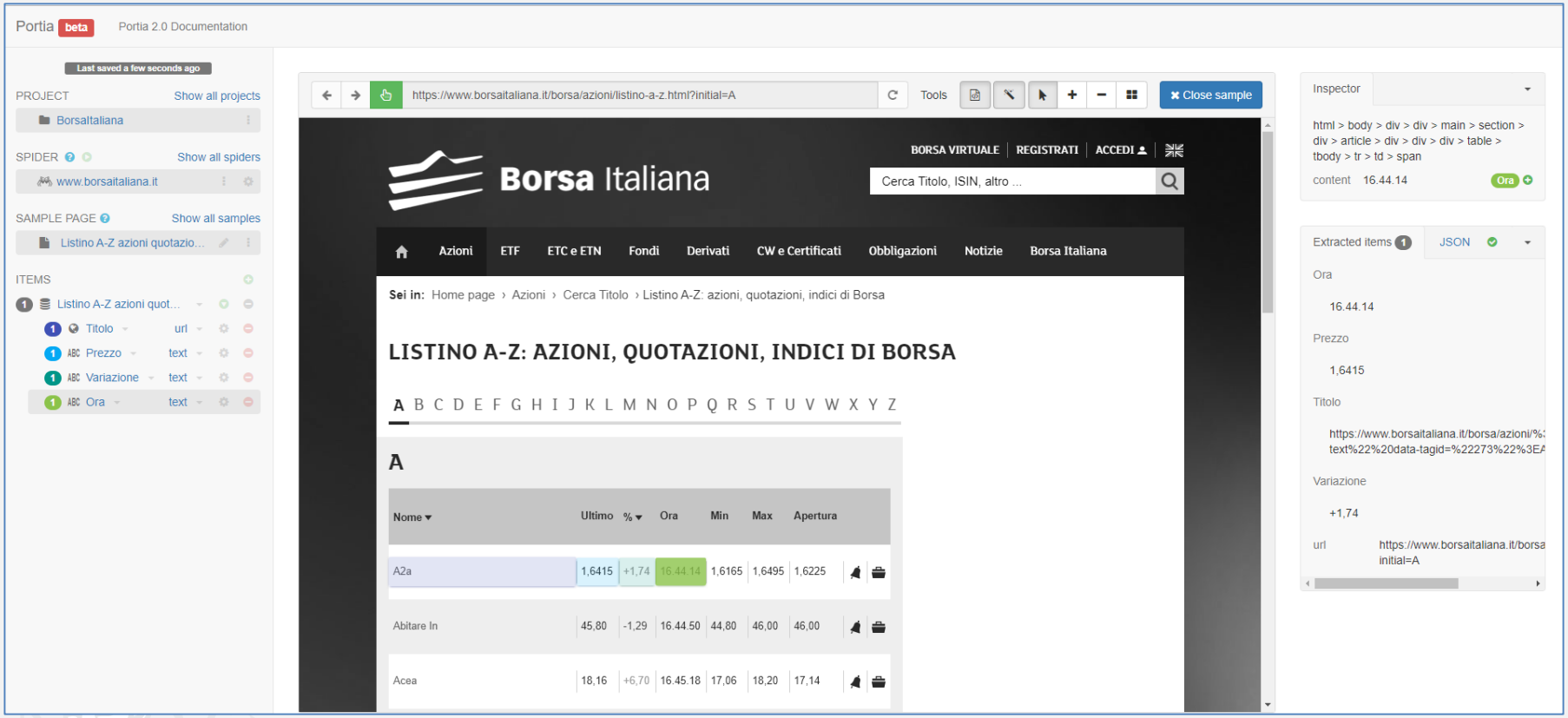


- Portia Web Scraper – *Configurare lo Spider*

- Creare una ***Sample Page*** per istruire lo spider all'estrazione dei dati di interesse;

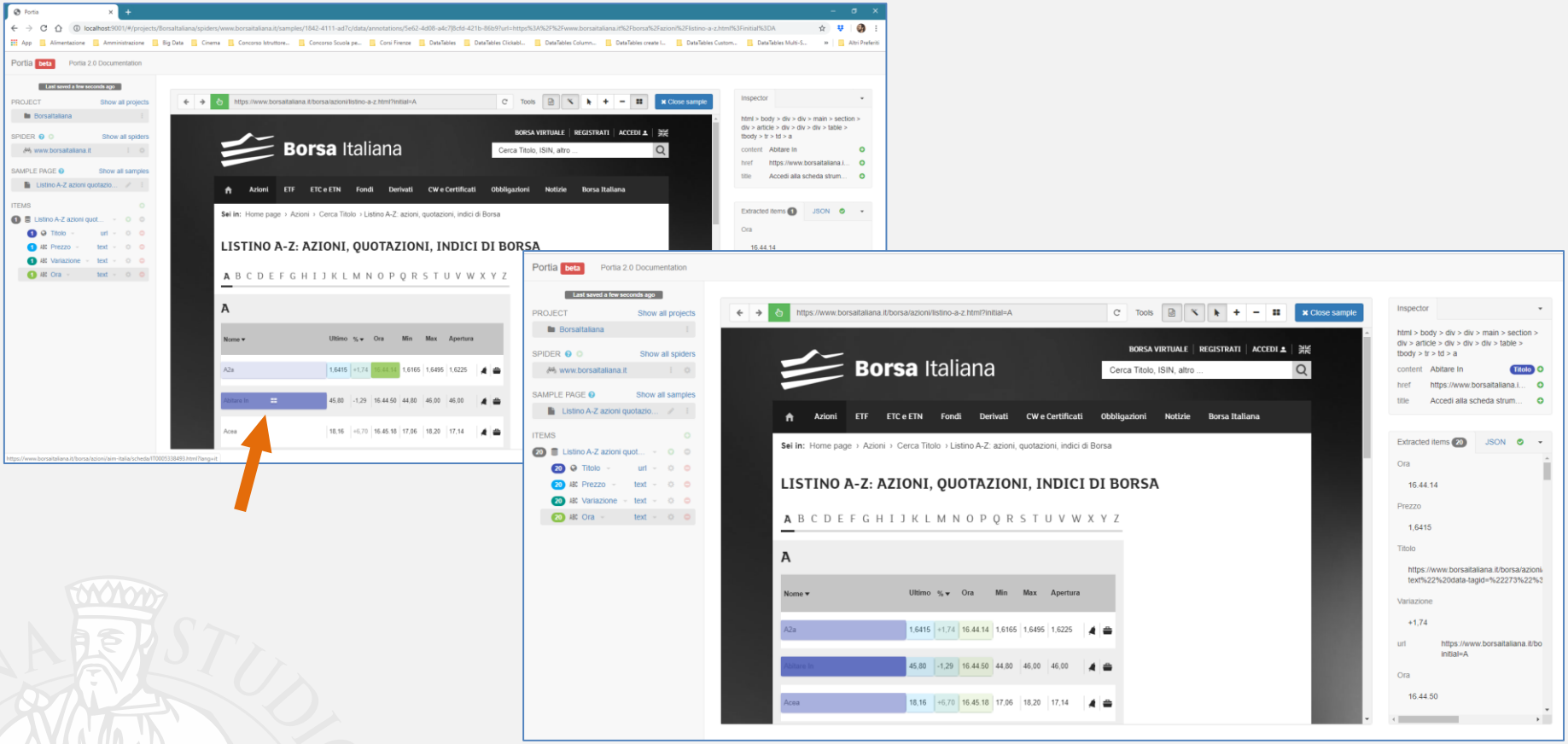
- Portia Web Scraper – *Configurare lo Spider*

- Creare una ***Sample Page*** per istruire lo spider all'estrazione dei dati di interesse;



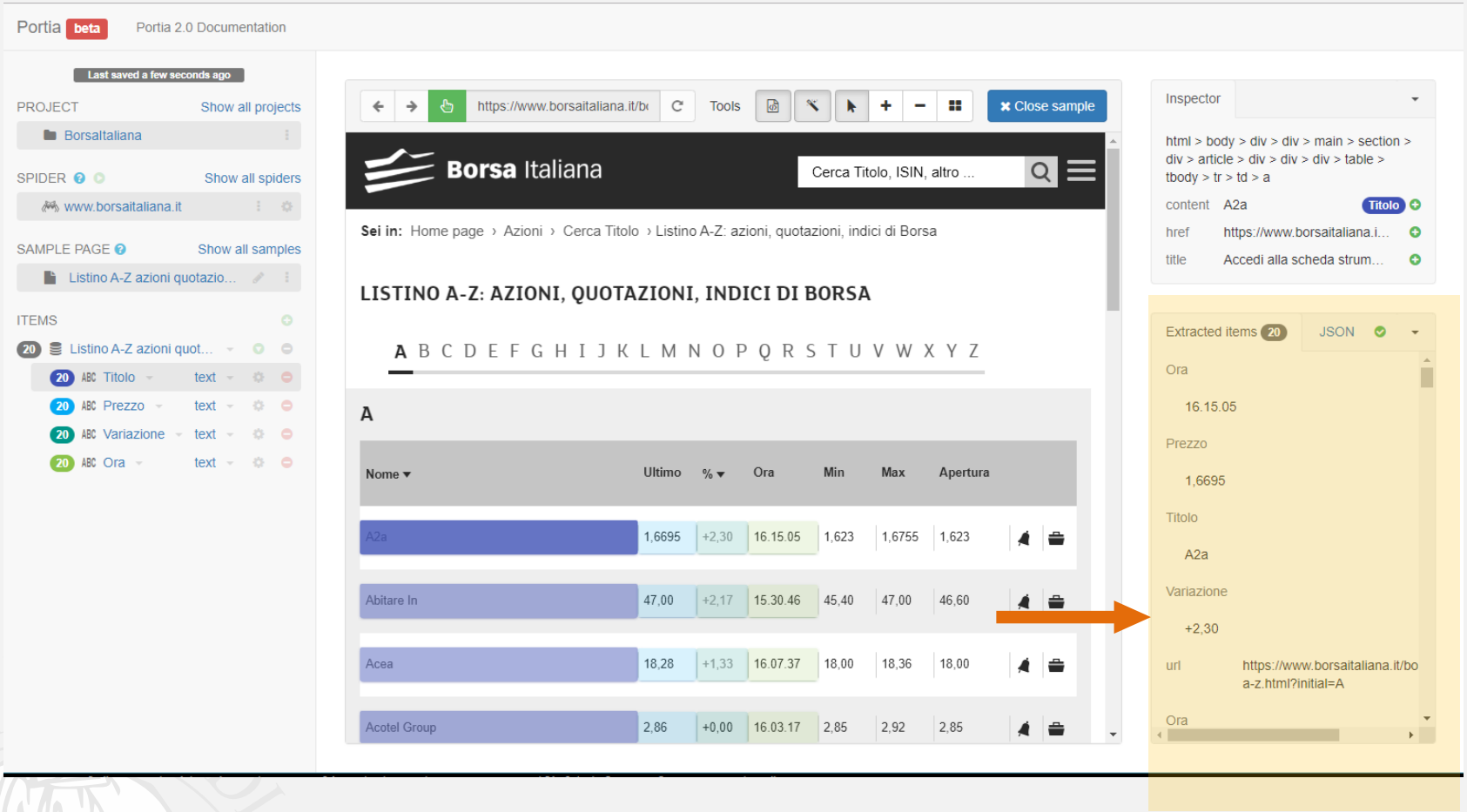
- Portia Web Scraper – *Configurare lo Spider*

- E' possibile definire semplicemente attraverso l'interfaccia grafica gli elementi di interesse da estrarre, nonché identificare pattern che si ripetono all'interno della pagina.



- Portia Web Scraper – *Avviare lo Spider*

- Una preview dei risultati dell'estrazione è mostrata nel riquadro a destra dell'interfaccia



- Portia Web Scraper ~ References

- Documentazione dal sito ufficiale:
  https://portia.readthedocs.io/en/latest/installation.html#installation

- Guida PDF:
  https://buildmedia.readthedocs.org/media/pdf/portia/latest/portia.pdf

- References & 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

- 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.

[Massai et al., 2019] Massai, L., Nesi, P., Pantaleo, G, "*PAVAL: A location-aware virtual personal assistant for retrieving geolocated points of interest and location-based services*", in Engineering Applications of Artificial Intelligence, vol. 77, pp. 70-85, 2019.