Ordine degli Ingegneri della Provincia di Roma

6 giugno 2017

"Information Extraction" basata su *"Natural Language Processing"* della lingua italiana

roberto.gallerani@ordingbo.it
https://www.gallerani.it

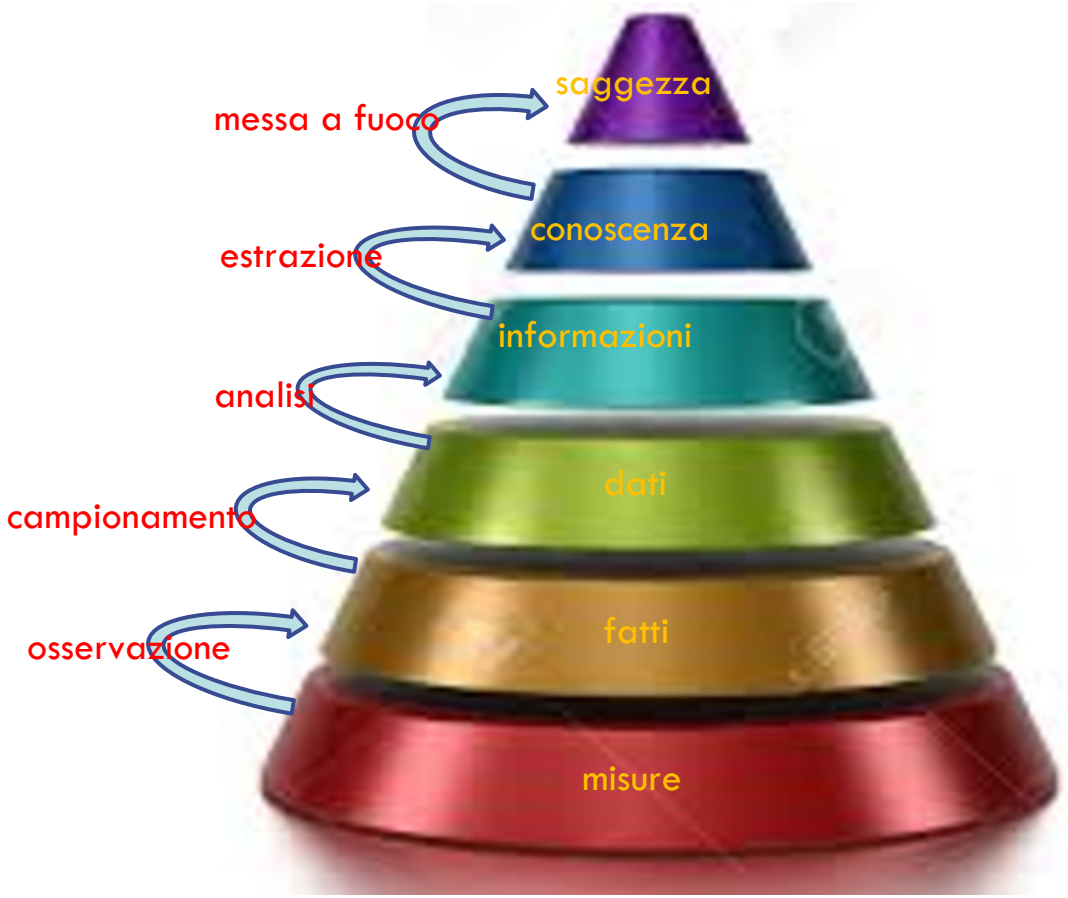
Sommario

- Dato vs Informazione: conoscenza
- *"Natural Language Processing"* e "*Information Extraction"*
- Tecniche e strumenti per l'estrazione di informazioni da testi
- L'impiego della piattaforma open source GATE
- Esempi
- Conclusioni

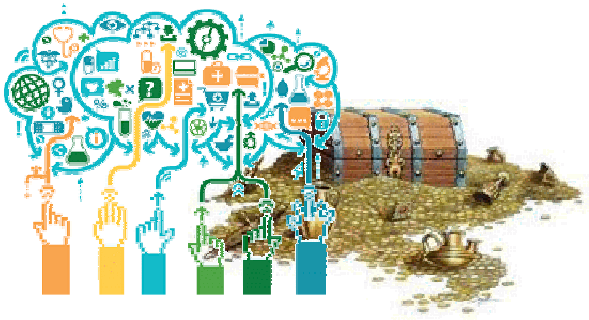
Dato vs Informazione

- Un dato è ciò che è immediatamente presente alla conoscenza prima di ogni elaborazione.

- I dati, nell'ambito informatico, si presentano sotto varie forme (numeri, lettere dell'alfabeto, immagini, suoni, simboli ecc.); a essi si deve attribuire un significato affinché rappresentino una realtà d'interesse.



… Ripetiamoci insieme dove va il mondo



Informazioni e **conoscenza** sono la **ricchezza** da valorizzare **e** sono il **motore** della **digitalizzazione**

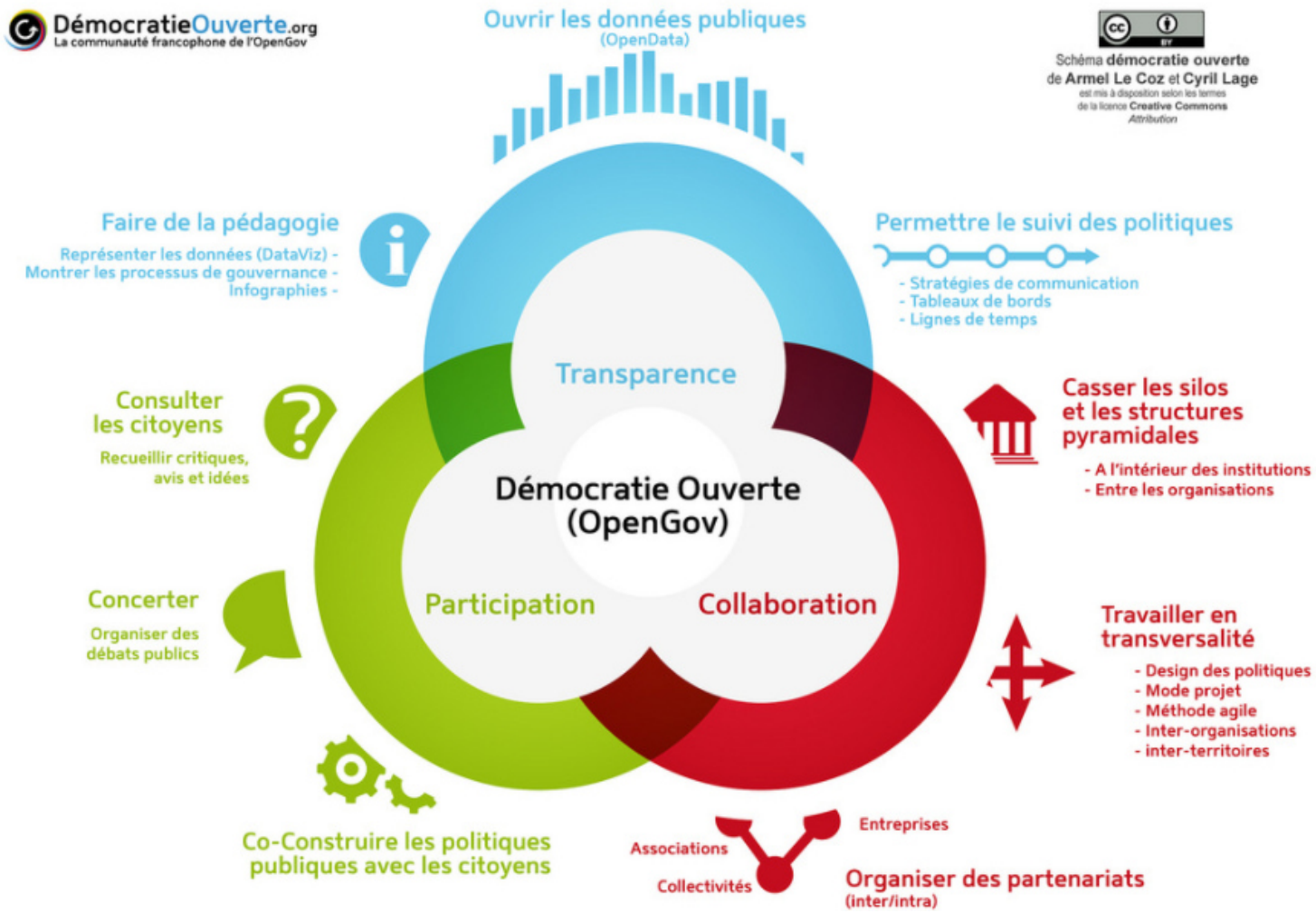
“Mondo” (privato e pubblico) tende a:
piattaforme capaci di esporre **servizi**, “**orchestrando**” comunicazioni, informazioni e processi sotto controllo di “sistemi” in grado di attuare e soddisfare automaticamente le regole e le policy prefissate, **rispondendo** automaticamente alle richieste degli utenti e **instradando** le **richieste** o **anticipandole**.

Informazioni
Disporre, analizzare, estrarre, sintetizzare

Orchestrazione
Workflow, autodattamento alle eccezioni

Capacità di interpretazione e anticipazione

… Ripetiamoci insieme dove va il mondo



… Ripetiamoci insieme dove va il mondo

"*Government as a Platform*" Tim O'Really

http://chimera.labs.oreilly.com/books/1234000000774/ch02.html

L'evoluzione naturale del ***government as a platform*** *appare sempre più quella di passare dall'esporre servizi e dati alla esposizione di servizi basati su regole e policy governati da un software "orchestratore": software defined government e software defined bank*.

Se veniamo "profilati" per venderci pubblicità o servizi perché non esserlo per venir serviti meglio dalla PA.

Es.:

Potremmo essere avvisati direttamente che abbiamo i requisiti per ottenere un beneficio.

La richiesta di una licenza edilizia potrebbe essere fatta attraverso dei servizi applicativi in grado di verificare le informazioni presenti nei diversi database,procedendo subito con il via libera o meno. Al cittadino spetterebbe solo di inserire la richiesta con le informazioni aggiuntive.

In una rete di servizi interconnessi potrebbero essere rilevate in tempo reale anche piccole trasgressioni, con "minimulte" o un richiamo, segnalando in automatico un comportamento sbagliato, incentivando comportamenti socialmente accettabili.

… Ripetiamoci insieme dove va il mondo

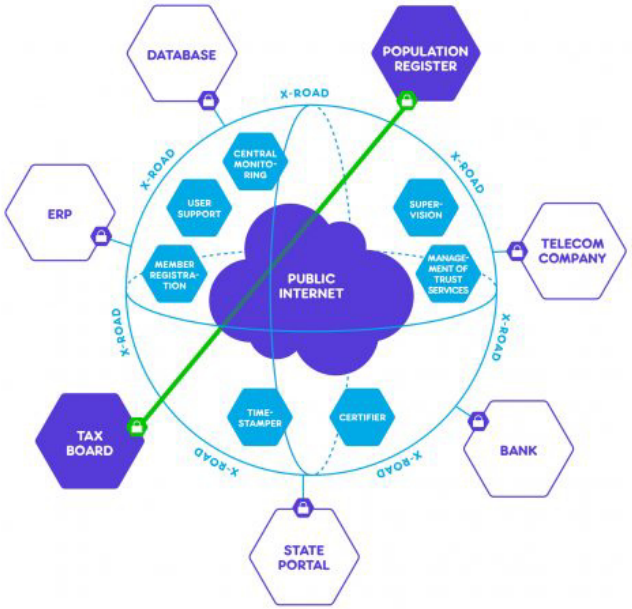
… altri lo stanno già facendo …

X-ROAD (Governo estone): piattaforma di integrazione che permette di scambiare servizi e informazioni. X-ROAD fornisce servizi di interrogazione a diversi database e rappresenta la base per un sistema articolato di servizi al cittadino.

In e-Estonia **database** sono **decentrati**:
✓ Non c'è un singolo *owner* o *controller*
✓ Ogni ente o agenzia governativo sceglie i servizi più idoenei tra quelli disponibili
✓ I servizi possono essere aggiunti in ogni momento

In 2013 e-Estonia:

✓Oltre **287 millioni** di interrogazioni attraverso X-road.
✓Oltre **170** banche dati integrate attraverso i servizi di X-Road.
✓Oltre **2.000** servizi disponibili attraverso X-Road.
✓Oltre **900** organizzazioni utilizzavano quotidianamente X-Road.
✓Più del **50%** dei cittadini dell'Estonia facevano uso di portali basati su X-ROAD



… Ripetiamoci insieme dove va il mondo

Da informazioni distribuite e non strutturate …

nei documenti non strutturati del e dal web

nei documenti aziendali e pubblici di varia natura

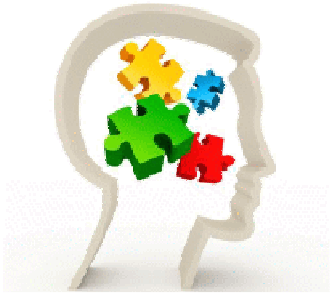
sui supporti di comunicazione (email, chat, …)

Riconoscere informazioni

non significa eseguire ricerche su testo indicizzato con al più la conoscenza delle relazioni di sinonimia e tecniche di previsione o sostituzione di parole

significa comprendere il significato della frase o almeno di un numero sufficiente di forme espressive

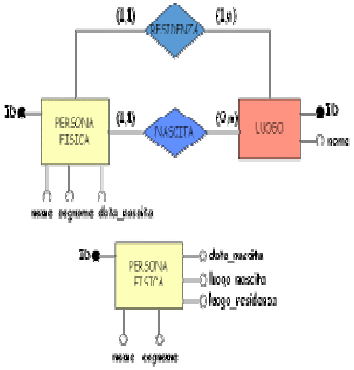
… Ripetiamoci insieme dove va il mondo



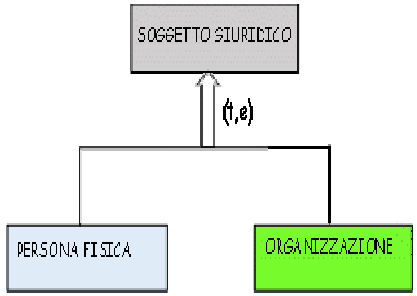
… a informazioni utilizzabili

… e, attraverso la comprensione, estrarre i concetti (entità) e le proprietà (informazioni)

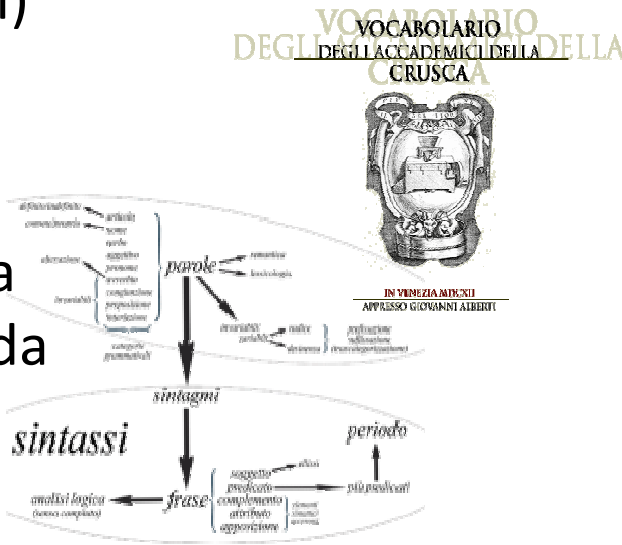
inoltre



significa comprendere le relazioni tra concetti diversi, le quali si possono rappresentare in proprietà (relazioni 1:1) o in collegamenti strutturati tra istanze di entità diverse (relazioni N:M)



La comprensione semantica è strettamente dipendente dalla lingua (ITA,ENG,…), dalla sua grammatica, da terminologie e forme espressive tipiche del contesto (dominio)



Alcune prime considerazioni su frasi di esempio…

"Mario Rossi è nato a Napoli il 22/07/1953 e vive a Roma. Egli lavora presso il Ministero dell'economia. Il suo recapito telefonico è 3291234567 8."

"ROSSI MARIO e ROSSI GIORGIO sono accusati di furto aggravato con scasso anche se è da tener conto che il più giovane dei due fu dichiarato incapace di intendere e di volere precedentemente alla data del fatto"

"… si presenta VERDI MARIO alle ore 12:40; posta la domanda "dove e quando è nato?" egli dichiara di essere nato a MILANO il 28.09.1962"

"Mario Rossi, amico di Carlo Verdi, è nato a Milano il 22/07/1953 dove risiede attualmente dopo un periodo di attività nella capitale negli anni 1990-2010. Lavorando presso il Ministero dell'economia ha maturato esperienze significative per l'incarico. Il recapito telefonico a cui contattarlo è 3291234567 8."

Scenario

✓ **Dominio**: documenti supporto alle autorità investigative (verbali di interrogatorio, trascrizioni di intercettazioni, sentenze)

✓ **Vincoli tecnologici**: open source

✓ **Impieghi/Obiettivi:** supporto all'analista, potenziamento sistemi gestione documentale (ricerca e classificazione)

… e (al solito) ultimo, ma non ultimo

✓ Limitatezza **budget/risorse** disponibili

Cosa significa: interpretare il linguaggio …

definito/indefinito ← *articolo*
comune/proprio ← *nome*
verbo
alterazione ← *aggettivo*
pronome
avverbio
invariabili: *congiunzione*, *preposizione*, *interiezione*
categorie grammaticali

parole → *semantica*, *lessicologia*

parole → *invariabili/variabili* → *radice*, *desinenza* (*prefissazione*, *suffissazione*, *(trascategorizzazione)*)

PP → Prep *sul*, NP
NP → Ag *suo*, N *tavolo*

sintagmi

sintassi

frase → *soggetto* (*ellissi*), *predicato*, *complemento*, *attributo*, *apposizione* (*elementi sintattici accessori*)

frase → *analisi logica* *(senso compiuto)*

frase → *più predicati* → *periodo*

… Cosa significa: interpretare il linguaggio …

Linguistica

- ✓**Fonetica e fonologia**: i "suoni" di una lingua;
- ✓**Morfologia** (morfemi): le strutture interne di una parola;
- ✓**Sintassi**; la struttura delle frasi nelle relazioni tra le diverse parole componenti;
- ✓**Semantica**: il significato delle parole e delle frasi nel complesso;
- ✓**Pragmatica**: come il contesto influisce sull'interpretazione dei significati di una frase;
- ✓**Analisi del discorso**: unità linguistiche lunghe più di un enunciato.

"Rapina al supermercato con rivoltella da mille euro" → ambiguità sintattica sul *complemento di stima* "da mille euro" (Rapina ? Rivoltella ?) … cosa ci aiuta ? Il contesto, il senso complessivo del discorso o, nella peggiore delle ipotesi, conoscenze più ampie anche di domini diversi.

… Cosa significa: interpretare il linguaggio …

Processo: "*Natural Language Analysis*"

- ✓**pre-elaborazione del testo**
- ✓analisi lessicale
- ✓analisi sintattica
- ✓analisi semantica

Pre-elaborazione del testo

Document Triage
conversione di file digitali in un ***corpus*** di documenti testuali ben definiti: riconoscimento della *codifica dei caratteri*, identificazione della *lingua del testo*, esclusione di *elementi non rilevanti* nell'analisi testuale (es. header, footer, immagini, ecc), rimpiazzo di caratteri/sequenze.

Tokenization
sequenza di caratteri è divisa in ***token*** *(parole, numeri, punteggiatura, ecc).* Non banale con lingue come il tedesco, il cinese o il tailandese.

Sentence segmentation
identificazione dei confini delle diverse frasi che compongono il documento.

… Cosa significa: interpretare il linguaggio …

Analisi lessicale

- ✓pre-elaborazione del testo
- ✓**analisi lessicale**
- ✓analisi sintattica
- ✓analisi semantica

Tema e **lemma**. Il **tema** è la **radice di una parola**, ottenibile rimuovendo da una sua forma flessa (ad esempio la coniugazione di un verbo o il plurale di un sostantivo) la desinenza. Il **lemma** è invece la **forma canonica** della parola, ovvero quella che viene convenzionalmente scelta per rappresentarne tutte le forme flesse; in italiano ad esempio il lemma di un verbo è la sua coniugazione all'infinito presente, di un aggettivo il singolare maschile.

amava	→ tema: ama;	lemma: amare	VER:ind+impf+3+s
cane	→ tema: can	lemma: cane	NOUN-M:s
cani	→ tema: can	lemma: cane	NOUN-M:p

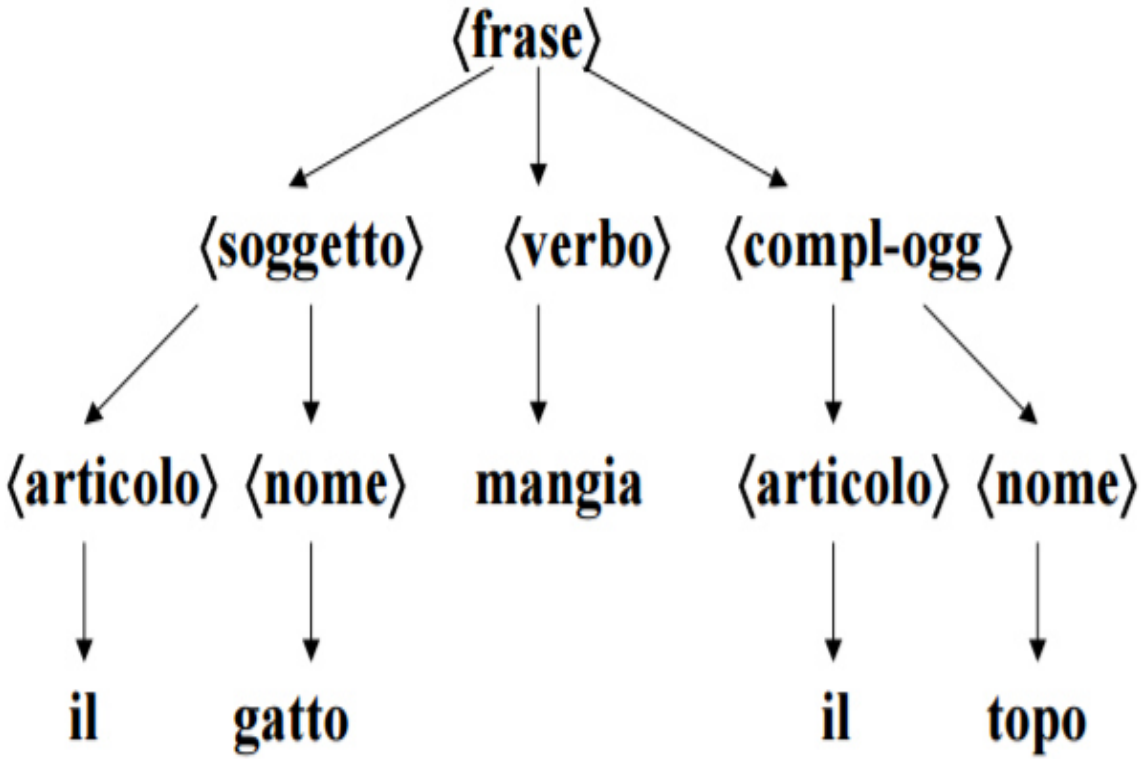
Cosa significa: interpretare il linguaggio …

Analisi sintattica

- ✓pre-elaborazione del testo
- ✓analisi lessicale
- ✓**analisi sintattica**
- ✓analisi semantica

Stabilire come le parole sono in relazione tra loro in un a frase.

Albero sintattico è composto da più livelli di raggruppamento delle parole.
Le foglie costituiscono il risultato dell'operazione di *Part-Of-Speech tagging* (POS *tagging*) I rami intermedi risultano invece dal passo di *chunking*, che raggruppa tra loro più parti del discorso sintatticamente correlate. Alla cima vi è la frase stessa.



Effetto *notazioni* morfologie parole e diversità algoritmi parsing in relazione a lingua !!!

… Cosa significa: interpretare il linguaggio …

Analisi semantica

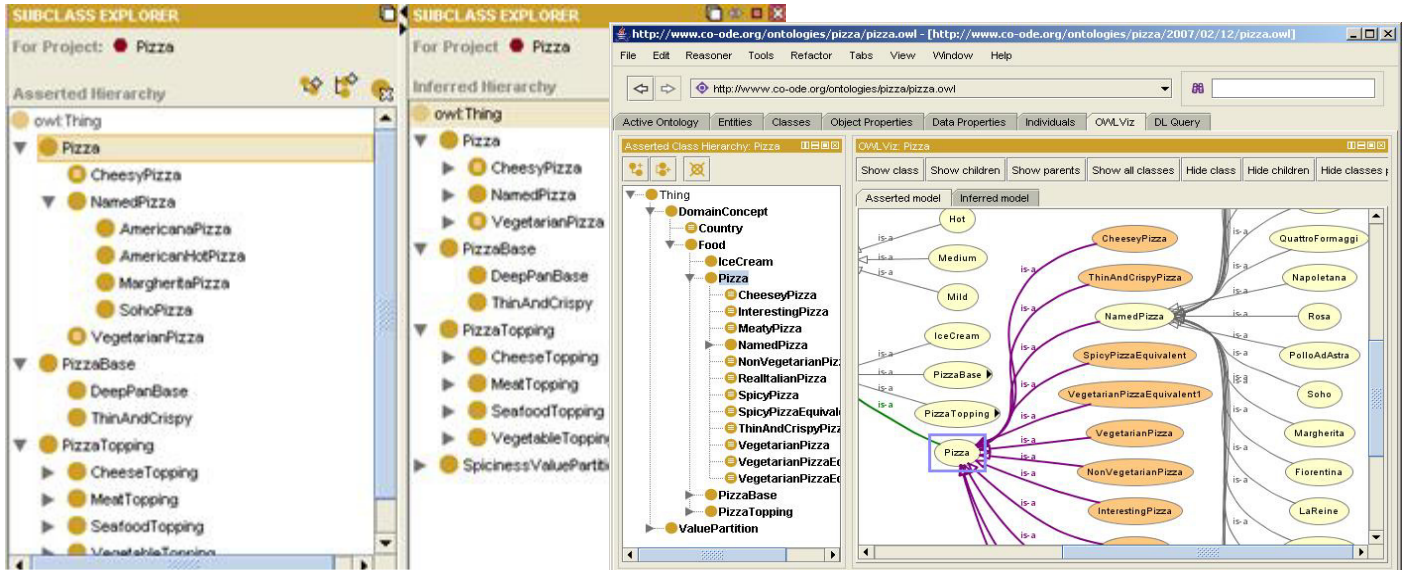
Sfruttare le informazioni ottenute dai precedenti passi per interpretare il significato della frase nel complesso.

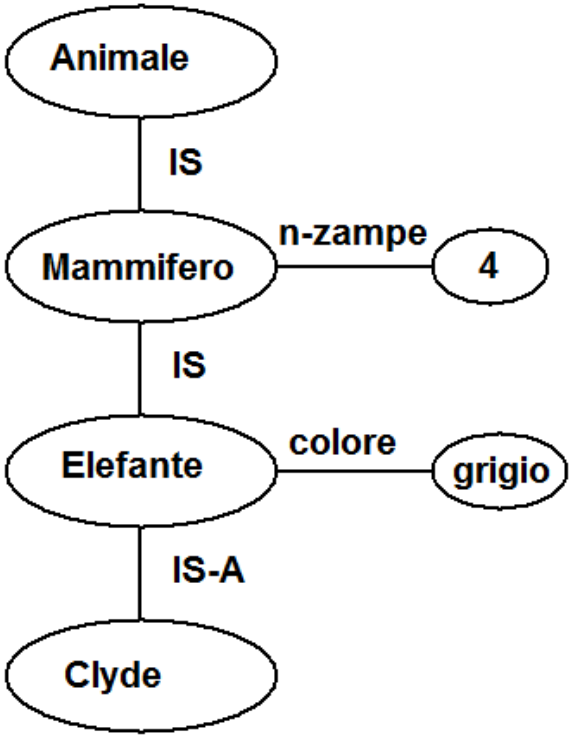
- ✓pre-elaborazione del testo
- ✓analisi lessicale
- ✓analisi sintattica
- ✓**analisi semantica**

Il diverso significato delle parole:

“polo” *maglietta, caramella, auto, sport, …. ???*
“golf” *capo di abbigliamento, auto, sport … ???*

L’esprimere concetti attraverso entità, proprietà e relazioni





… Cosa significa: interpretare il linguaggio …

"Word Sense Disambiguation": operazione che mira a determinare quale dei vari significati attribuibili ad una singola parola sia quello migliore nel contesto della frase in cui la parola è inserita.

Ambiguità sintattica

Luigi ha visto un uomo nel parco con il cannocchiale.

Vs

Ci scusiamo dei possibili fastidi causati porgendo cordiali saluti.

Ambiguità semantica

Mario prese un espresso contando di ricavarne un beneficio.

Carlo decise di acquistare una polo come regalo di compleano per Mario.

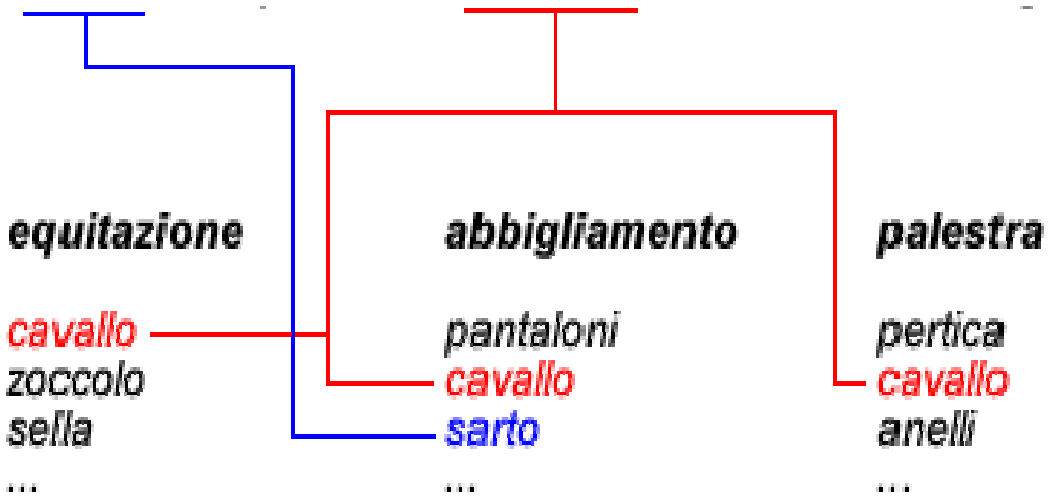
… Cosa significa: interpretare il linguaggio.

Analisi pragmatica

Si pone attenzione sull'intenzione dell'autore di una frase.
Ciò dipende da conoscenze esterne, legate al contesto in cui una frase è inserita.

Es.:
✓conoscenza del ruolo e dello status degli interlocutori,
✓collocazione spazio-temporale della situazione,
✓conoscenza dell'argomento trattato.

Il sarto ha riparato il cavallo e ora va meglio.



Estrarre le informazioni …

L' ***Information Extraction*** **(IE)** ha come obiettivo l'estrazione automatica di informazioni strutturate da documenti non strutturati o semi-strutturati.

I tre livelli di intervento

Named Entity Recognition: riconoscimento di nomi propri di entità (es.: persone, organizzazioni e luoghi), riconoscimento di espressioni temporali (es. date), di indirizzi, e di altre particolari informazioni (es. importi, voci di bilancio, ecc.).

Coreference Resolution: rilevamento di legami di coreferenza (ovvero l'insieme dei rinvii allo stesso referente: es. "Gianni disse che ***egli*** sarebbe andato a casa") e di anafora (ovvero riferimenti tra porzioni di testo più o meno distanti fra loro: es. *"Hanno preso l'operaio. L'uomo sembra abbia ucciso la moglie "*; *"Ho ascoltato il disco* di Sting *e non ho potuto fare a meno di apprezzarlo."*).

Relationship Extraction: riconoscimento di legami associativi tra entità; ad esempio dalla frase "Mario Rossi ***risiede*** a Bologna." si può dedurre la ***relazione di residenza*** che sussiste tra una persona (Mario Rossi) e una località (Bologna).

… estrarre le informazioni …

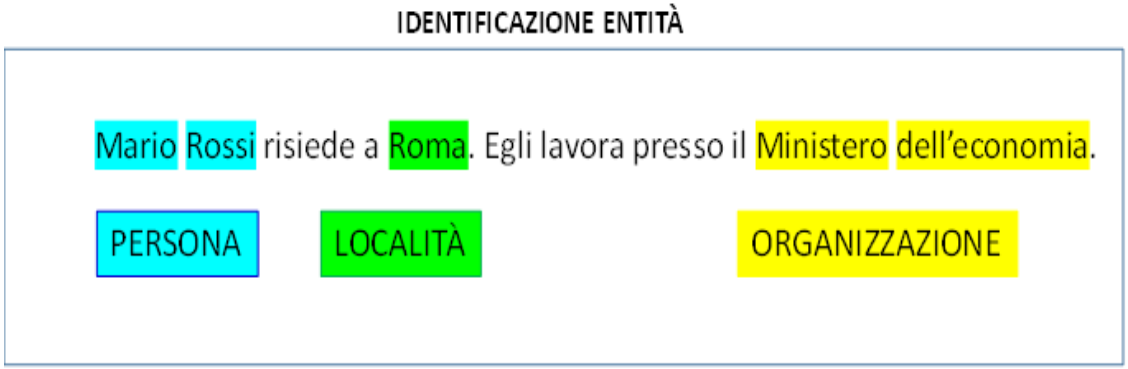
Named Entity Recognition - Approcci

Lookup list
ossia pattern suddivisi in categorie posti su “glossari/liste/dizionari”. E’ semplice ma poco preciso/dettagliato, difficile da stilare e manutenere.

Rule-based
il testo viene analizzato secondo un insieme di regole che determinano la presenza di entità all’interno di un testo e ne definiscono le relative informazioni semantiche. Le regole fanno uso, in genere, di informazioni di tipo morfologico e sintattico.
La complessità del linguaggio rende difficile la stesura del set di regole se il contesto è complesso, ma in ambiti circoscrivibili con un buon lavoro di studio e di *test* è possibile raggiungere un elevato livello di precisione.

Machine learning
Basato sulla generazione di un modello (statistico o neural network) di interpretazione, generato attraverso una fase di training/valutazione. Elevato carico computazionale in fase di training/valutazione, elevato carico di attività in fase di predisposizione dei pattern di apprendimento.

… estrarre le informazioni …



… estrarre le informazioni …

un esempio sugli indirizzi ….

77. VERDI VITO Antonio, detto Giuseppe, nato a Crotone il 20.03.1978, residente a Cadelbosco di Sopra (RE), via Dante Alighieri domiciliato a Scandiano (RE), frazione Arceto, via 2 Agosto 1980 Vittime di Bologna nr. 4, attualmente detenuto c/o Casa Circonda
e difeso di fid

78. SALEMI G
fiducia dall'Av

79. CALANDR
Elena nr. 16,
c/o lo studio

80. FOGGIA
assistito e dif

81. FAZZI Fra
Foro di Reggi

82. ZOLLA An

Indirizzo

CAP	
c/o	
civico	
denominazione	2 Agosto 1980 Vittime di Bologna
frazione	Arceto
indirizzo	via 2 Agosto 1980 Vittime di Bologna piano 4, Arceto Frazione di Scandiano (RE)
interno	
luogo	Scandiano
piano	4
provincia	RE
qualificatore	via
rule	Ind1Bis
scala	
subCiv	

Open Search & Annotate tool

Type	Set
Indirizzo	
Indirizzo	
Indirizzo	
Indirizzo	
Indirizzo	
Indirizzo	
Indirizzo	
Indirizzo	

[es.: elaborazione atto pubblico-dati personali modificati]

… estrarre le informazioni …

un esempio sugli indirizzi ….

77. VERDI VITO Antonio, detto Giuseppe, nato a Crotone il 20.03.1978, residente a Cadelbosco di Sopra (RE), via Dante Alighieri nr. 53/2, di fatto

SoggettoFisico

AIRE	
Codice Fiscale	
Data di nascita	20.03.1978
Indirizzo domicilio	via 2 Agosto 1980 Vittime di Bologna piano 4, Arceto Frazione di Scandiano (RE)
Indirizzo residenza	via Dante Alighieri n. 53/2, Cadelbosco di Sopra (RE)
Luogo di nascita	Crotone
MATERNITA	
PATERNITA	
Partita IVA	
Soprannome	Giuseppe
cognome	VERDI
nome	VITO Antonio
rule	SF4,SFDN1
titoli	

Open Search & Annotate tool

[es.: elaborazione atto pubblico-dati personali modificati]

… estrarre le informazioni …

un esempio sugli indirizzi ….

Qualificatore Indirizzo	(**Viale**)
Denominazione Propria	(**Risorgimento**)
Qualificatore Numero Civico	(**n.**)
Numero Civico	(**83**)
Qualificatore SubCivico	(**palazzina**)
SubCivico	(**B**)
Qualificatore Scala	(**scala**)
Scala	(**A**)
Qualificatore Piano	(**piano**)
Numero Piano	(**3°**)
Qualificatore Interno	(**int**)
Interno	(**12**)
Qualificatore Presso	(**c/o**)
Denominazione relativa a Presso	(**Famiglia Mariani**)
Qualificatore CAP	(**CAP**)
CAP	**(40069)**
Frazione	(**Ponte Ronca**)
Qualificatore Frazione	(**di**)
Luogo	**Zola Predosa**)
Provincia	(**BO**)

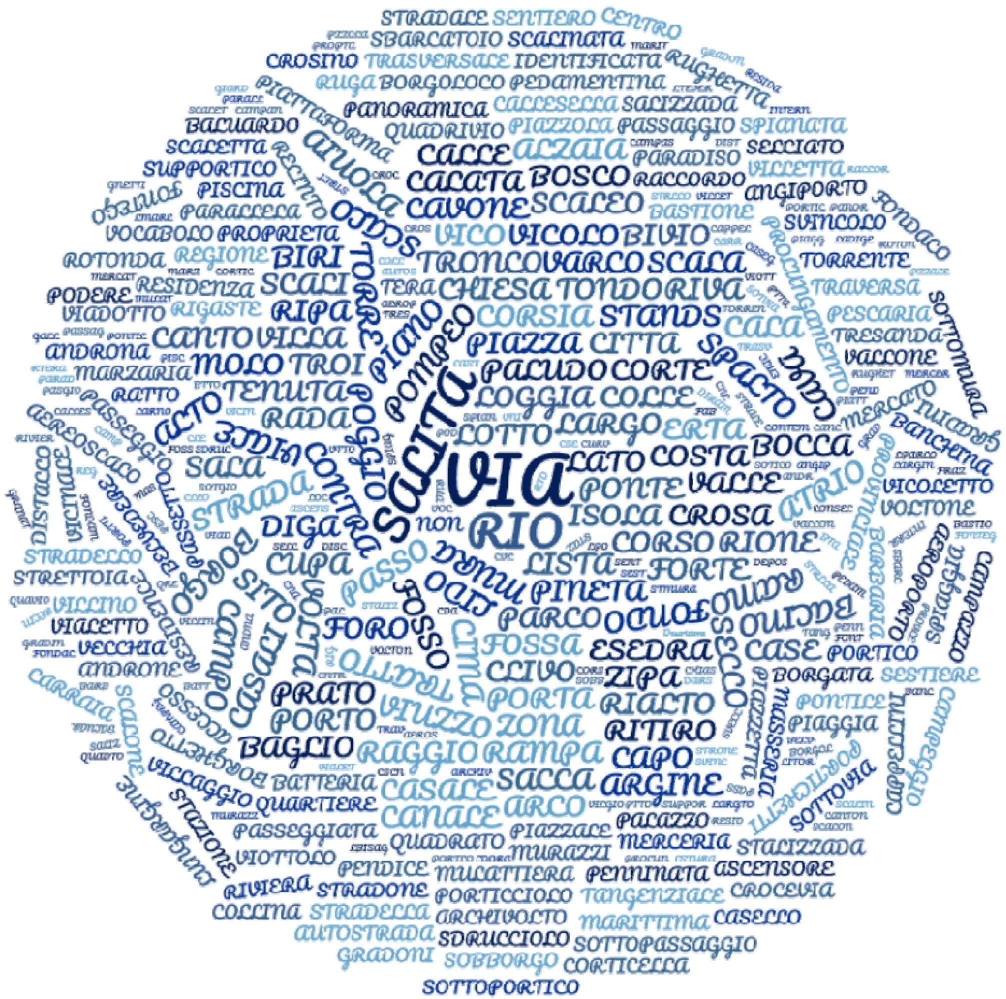
… estrarre le informazioni …

un esempio sugli indirizzi ….

Qualificatore di indirizzo *("qualificatore di toponimo"): es. via, viale, …. Sono **327 !***

http://www.laputa.it/denominazioni-urbanistiche-generiche/#15/45.6686/13.1041

arco, ratto, rettifilo, riello, baluardo, banchi, fondo, regione, riva, bacino, bastione, calata , corso, campo, canale, scalette, scalone, salita inferiore/superiore, case sparse, spiaggia, spiazzo, strada bassa/inferiore/superiore/di bonifica/militare/…, distacco, fossato, lungo*, passeggiata, vicolo cieco, raggio, quadrato, volta, ….



… estrarre le informazioni

un esempio sugli indirizzi ….

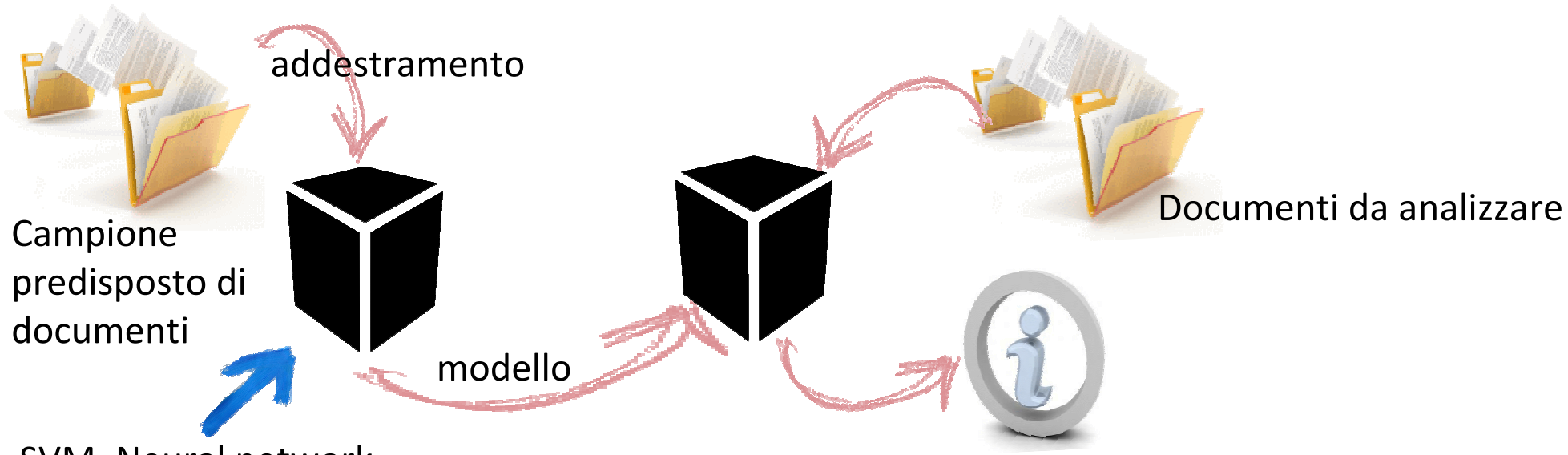
*…. durante il **corso** delle indagini 3 intercettazioni attirarono l'attenzione degli inquirenti ……*

*… esaminammo in **via** preliminare 10 atti …*
*…. nel **fondo** del fiume ….*
*…. quella **volta** a Milano...*
*…. I **volti** di Carmine e Pasquale tradirono una reazione….*
*… in **corso** d'istruttoria 3 persone furono interrogate…."*
*... nel **corso** dell'udienza ….";*

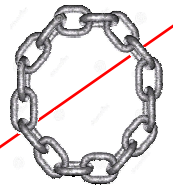
*"lo incontrai per la prima volta in **via Garibaldi**, nel mese di gennaio, una volta arrivato a **Napoli**;".*

…. Via 2 agosto…..
…..Via 2 agosto 1980….. ?

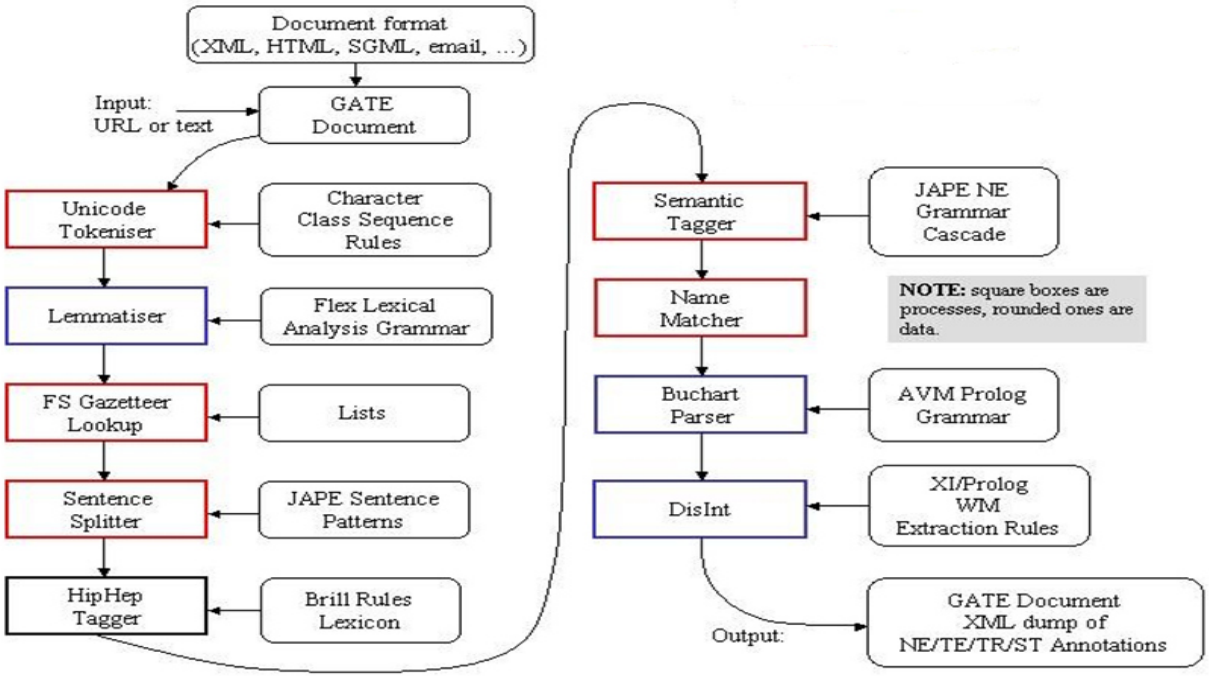
"Machine learning" e "Rule based"



Machine learning



Rule based



"Machine learning"

Dato insieme **T** *(training set)* composto dalle coppie (***input,output***) determina un modello che associa ad ogni elemento ***input,*** detto **istanza**, un valore ***output***, tale che applicato ad un $\boldsymbol{input}_i$ restituisca $\boldsymbol{output}_j \in \mathbf{T}$.
Un'istanza ***input*** è un vettore di elementi/attributi (**feature**).

Supervised learning, in cui il sistema viene istruito tramite esempi, ovvero dati in input dei quali è specificato il corretto valore di output;

Unsupervised learning, se al contrario i dati in input non sono etichettati per dare un riferimento nell'impostazione del modello, in questo caso spetta all'algoritmo stesso trovare il modello dei dati attraverso il riconoscimento di pattern;

Reinforced learning, altro caso in cui non è fornito un set di esempi, in questo tipo di algoritmi tuttavia viene fornito un *feedback* sulla correttezza dei risultati ottenuti dai vari passi di elaborazione.

"Machine learning"

NER via Machine Learning

è un problema di *sequence labeling* (o *sequence tagging*), attività di *pattern recognition* per la categorizzazione, attraverso l'assegnamento di etichette, degli elementi che compongono una sequenza appartenente alla Named Entity.
Gli algoritmi *sequence labeling* considerano un'istanza anche in relazione ad elementi adiacenti.

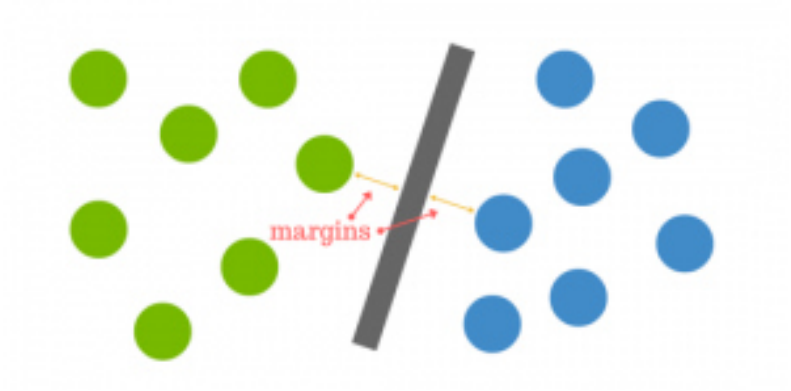
Support Vector Machine (SVM)

Training set

Presente all'interrogatorio il Dr. Mario Rossi, difensore dell'indagato.

"Machine learning"

Support Vector Machine (SVM)

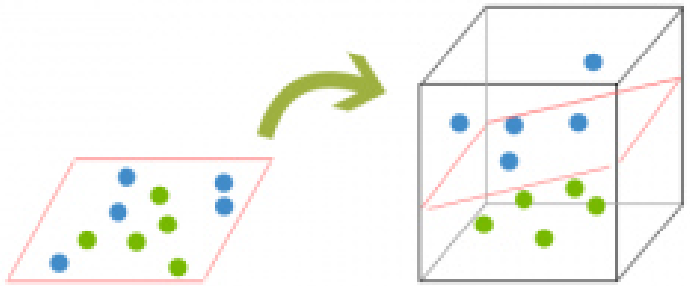


SVM sono basate sull'idea di trovare un "iperpiano" in grado di dividere al "meglio" un dataset di attributi in due 2 classi.

In 2D l'"iperpiano è una retta. I due "punti attributo" più vicini alla retta sono i "support vector".

… e se in 2D non è possibile ?

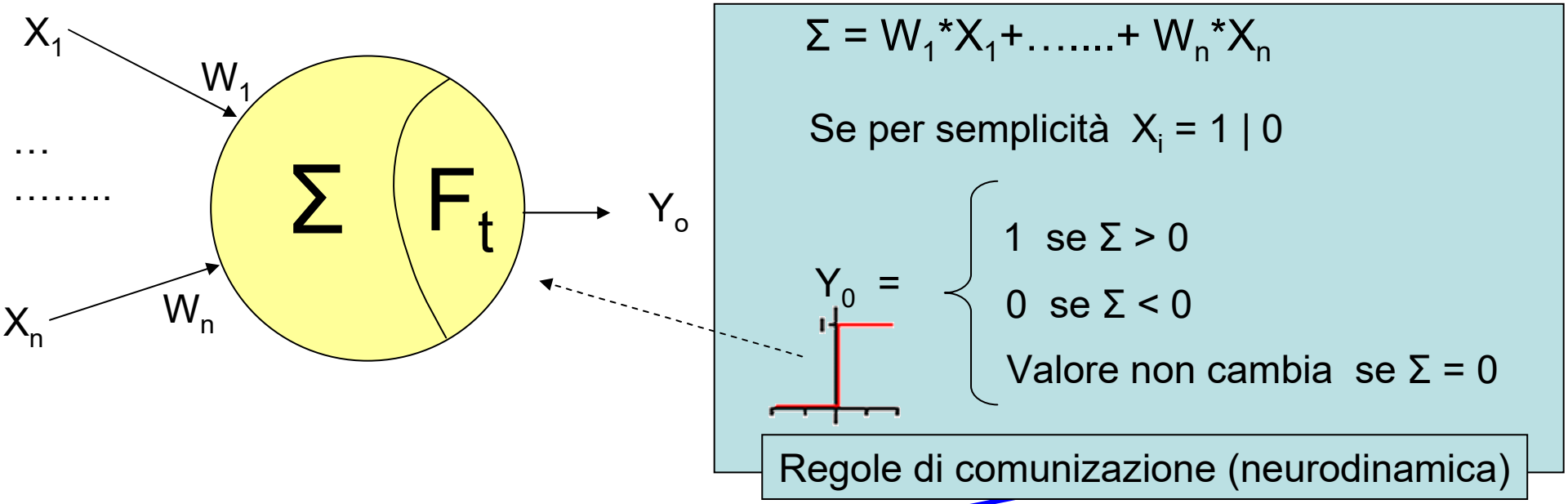
… si passa in 3D

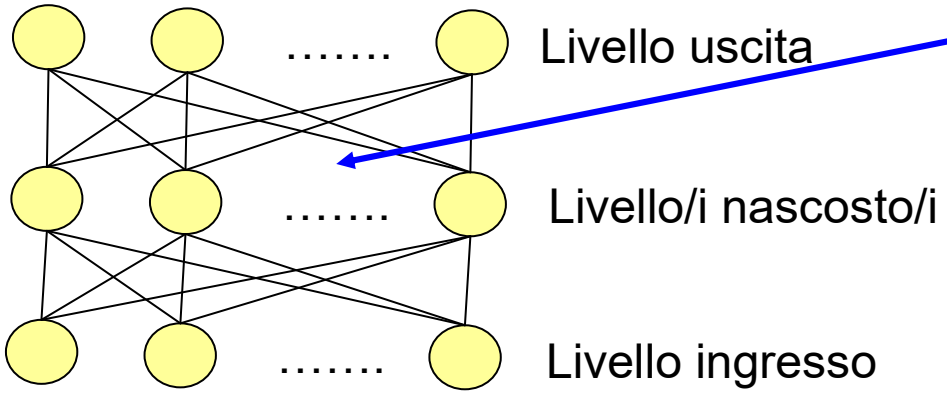


… e se in 3D non è possibile ? I dati sono mappati a dimensione N+1 procedendo fino a individuare un "iperpiano" e i relativi support vector nella dimensione N+J in grado di segregare in due classi.

"Machine learning"

Neural network (*perceptron model*)

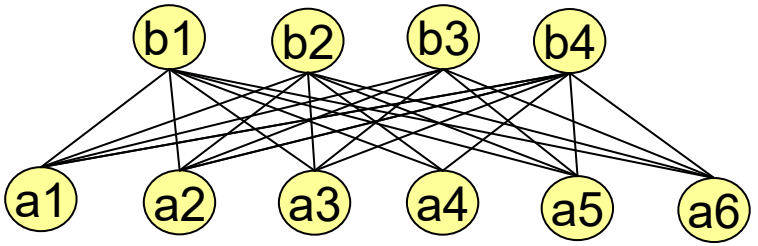




Le informazioni fluiscono sulla rete sulla base delle connessioni ossia sulla base dei loro pesi.

"Machine learning"

Neural network



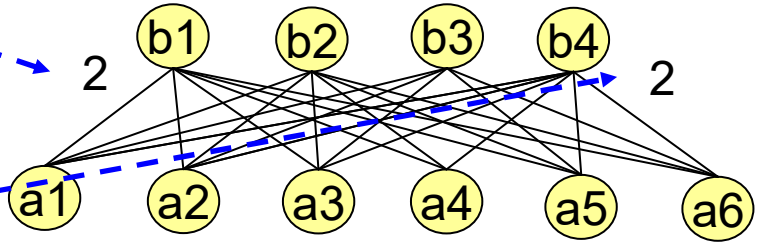
Addestramento:
m sequenze ([a1,a2,a3,a4,a5,a6],[b1,b2,b3,b4])
con m < min(6,4) => 4

A1 = (1,0,1,0,1,0) B1 = (1,1,0,0) → forma bipolare → X1 = (1,-1,1,-1,1,-1) Y1 = (1,1,-1,-1)
A2 = (1,1,1,0,0,0) B1 = (1,0,1,0) → forma bipolare → X2 = (1,1,1,-1,-1,-1) Y2 = (1,-1,1,-1)
…. m = 4 (4 pattern di apprendimento)

Matrice delle correlazioni (dei pesi W) → M = $X_1^T*Y_1+\ldots+X_4^T*Y_4$ → matrice 6 x 4

Elemento m_{ij} rappresenta il peso della connessione del neurone b_j in corrispondenza dell'ingresso a_i

2	0	0	-2
0	-2	2	0
2	0	0	-2
-2	0	0	2
0	2	-2	0
-2	0	0	2



"Machine learning"

Neural network

Esecuzione

Dall'ingresso all'uscita:

A * M = [vettore riga 6 elem.] * [matrice 6 x 4] = [vettore riga 4 elem.] → Ft → [vettore finale]

Es.:

[1,0,1,0,1,0] *

2	0	0	-2
0	-2	2	0
2	0	0	-2
-2	0	0	2
0	2	-2	0
-2	0	0	2

= [4,2,-2,-4] → [1,1,0,0]

Dall'uscita all'ingresso:

B * M^T = [vettore riga 4 elem.] * [matrice 4 x 6] = [vettore riga 6 elem.] → Ft → [vettore finale]

Es.:

[1,0,1,0] *

2	0	2	-2	0	-2
0	-2	0	0	2	0
0	2	0	0	-2	0
-2	0	-2	2	0	2

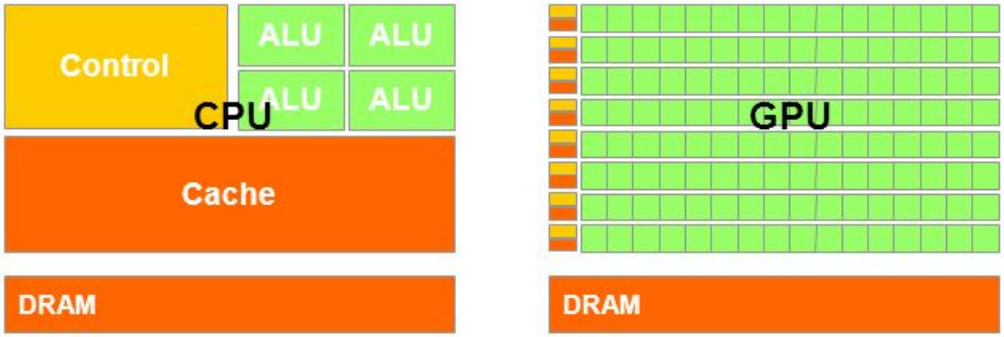
= [2,2,2,-2,-2,-2] → [1,1,1,0,0,0]

"Machine learning"

Neural network, "deep learning" e il ruolo crescente dell'uso di GPU/parallel programming

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, b = \begin{bmatrix} b_0 \\ \vdots \\ b_p \end{bmatrix}, e = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & \vdots & \ddots & x_{2p} \\ \vdots & \vdots & \ddots & \cdots \\ 1 & x_{N1} & \cdots & x_{Np} \end{bmatrix}$$

… è necessaria capacità di calcolo e capacità di calcolo e….



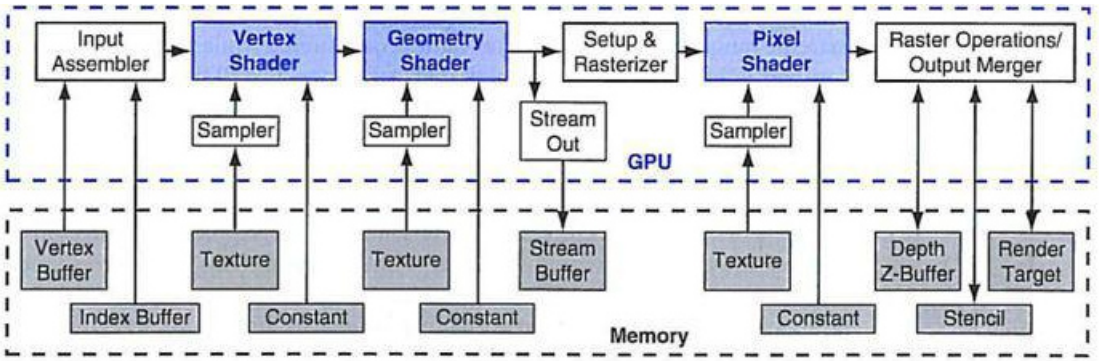
… le GPU sono fatte per fare calcoli in pipeline e in parallelo in thread sui numerosi core della GPU….

Hw a costi accessibili

Ambienti di programmazione parallela e neural processing

Snapdragon Neural Processing Engine SDK

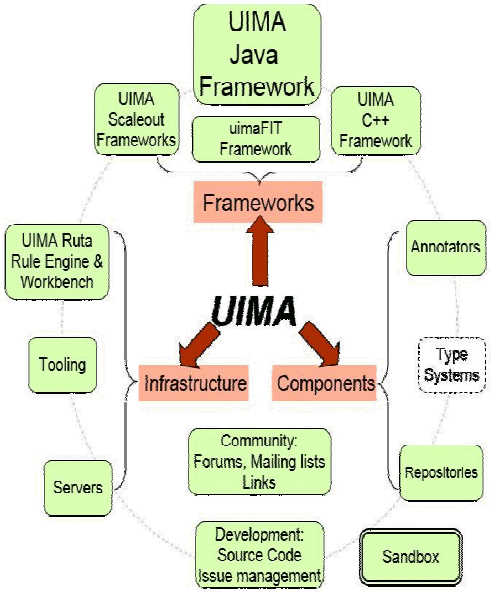
… e le pipeline sono programmabili ….



Strumenti

Stanford CoreNLP (http://nlp.stanford.edu/software/corenlp.shtml)
Il gruppo di ricerca di Stanford University mette a disposizione una serie di strumenti di NLP che sono complessivamente denominati Stanford CoreNLP. Stanford CoreNLP è rilasciato con GNU General Public License.

UIMA (http://uima.apache.org/index.html)
Unstructured Information Management Architecture



http://opennlp.apache.org/

è una libreria Apache per analisi di testi che supporta vari tipi di azione

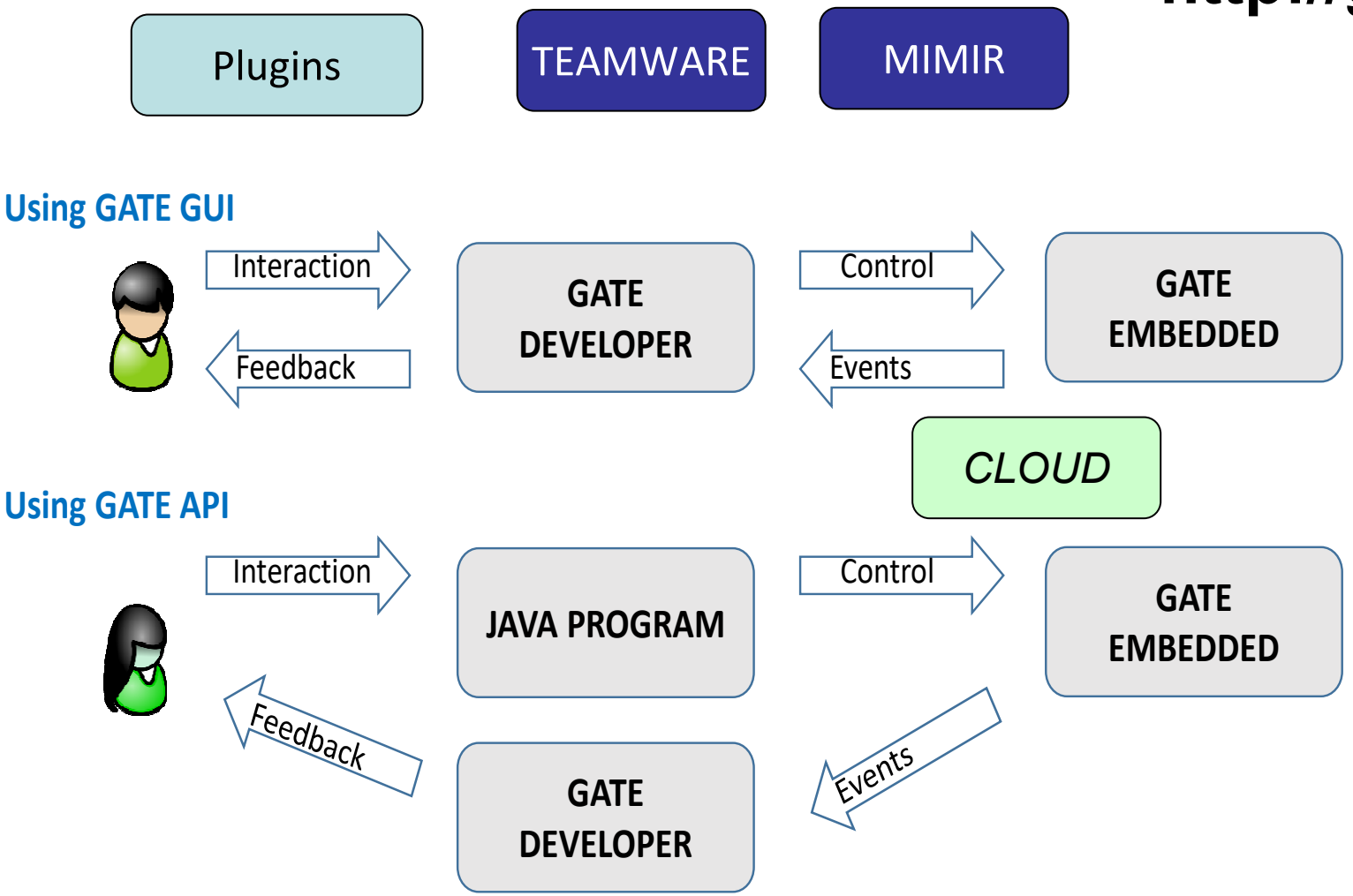
general architecture for text engineering https://gate.ac.uk/

è un'infrastruttura open source per lo sviluppo di sistemi di IE proposta da University of Sheffield.

GATE

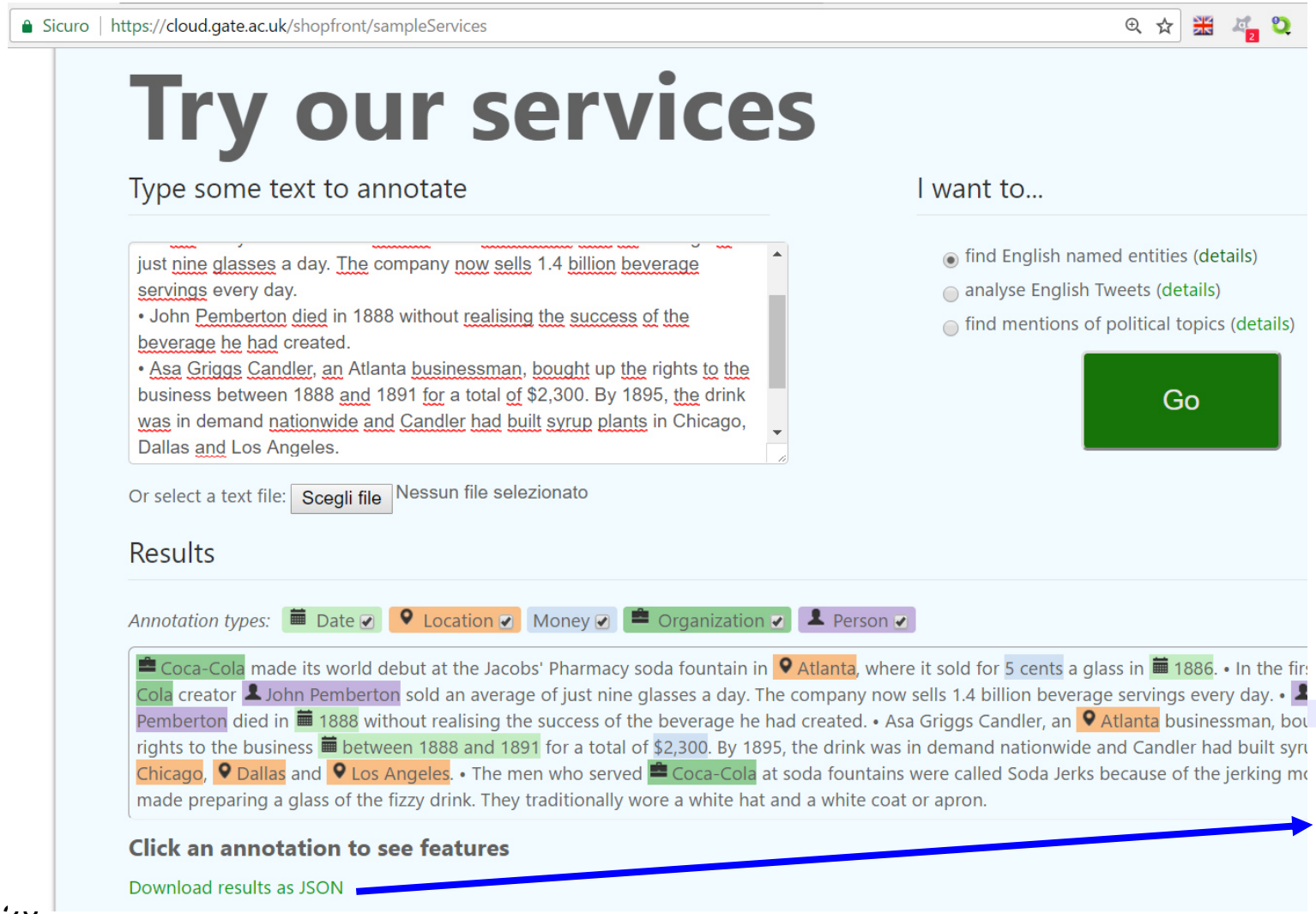
General Architecture for Text Engineering

http://gate.ac.uk/



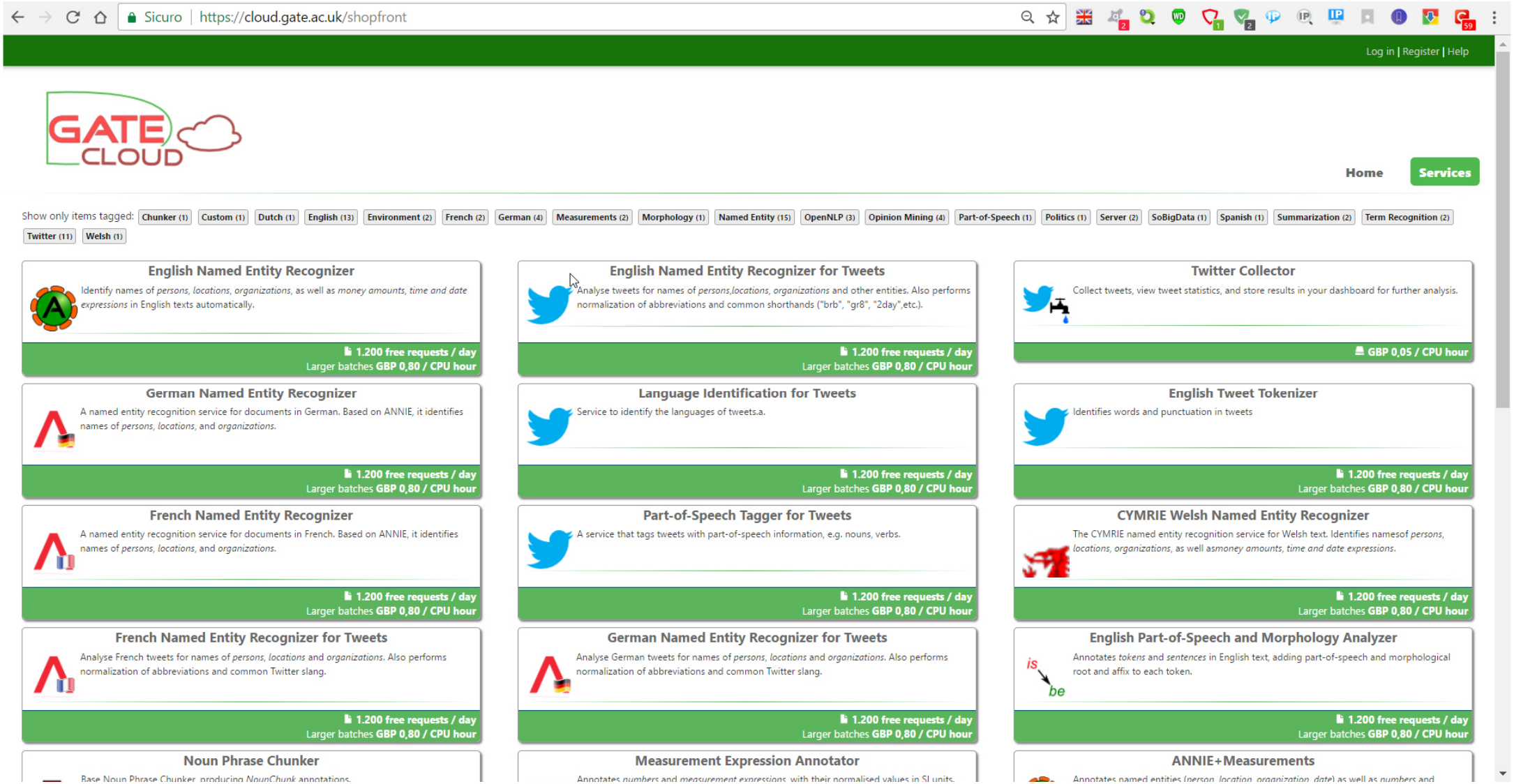
GATE Cloud

Uso a pagamento; utile preview di esempi (non italiano)




```
"Person": [{
        "indices": [162, 176],
        "firstName": "John",
        "gender": "male",
        "surname": "Pemberton",
        "kind": "fullName",
        "rule": "PersonFull",
        "ruleFinal": "PersonFinal",
        "matches": [475, 476]
    }, {
        "indices": [286, 300],
        "firstName": "John",
        "gender": "male",
        "surname": "Pemberton",
        "kind": "fullName",
        "rule": "PersonFull",
        "ruleFinal": "PersonFinal",
        "matches": [475, 476]
    }
],
"Location": [{
        "indices": [72, 79],
        "kind": "locName",
        "rule": "InLoc1",
        "locType": "city",
        "ruleFinal": "LocFinal",
        "matches": [472, 478]
    }, {
        "indices": [401, 408],
        "locType": "city",
        "rule": "Location1",
        "ruleFinal": "LocFinal",
        "matches": [472, 478]
    }, {
        "indices": [586, 593],
        "kind": "locName",
        "rule": "InLoc1",
        "locType": "city",
```

GATE Cloud


Sicuro | https://cloud.gate.ac.uk/shopfront
Log in | Register | Help
GATE CLOUD
Home
Services
Show only items tagged: Chunker (1) Custom (1) Dutch (1) English (13) Environment (2) French (2) German (4) Measurements (2) Morphology (1) Named Entity (15) OpenNLP (3) Opinion Mining (4) Part-of-Speech (1) Politics (1) Server (2) SoBigData (1) Spanish (1) Summarization (2) Term Recognition (2) Twitter (11) Welsh (1)
English Named Entity Recognizer
Identify names of persons, locations, organizations, as well as money amounts, time and date expressions in English texts automatically.
1.200 free requests / day
Larger batches GBP 0,80 / CPU hour
English Named Entity Recognizer for Tweets
Analyse tweets for names of persons,locations, organizations and other entities. Also performs normalization of abbreviations and common shorthands ("brb", "gr8", "2day",etc.).
1.200 free requests / day
Larger batches GBP 0,80 / CPU hour
Twitter Collector
Collect tweets, view tweet statistics, and store results in your dashboard for further analysis.
GBP 0,05 / CPU hour
German Named Entity Recognizer
A named entity recognition service for documents in German. Based on ANNIE, it identifies names of persons, locations, and organizations.
1.200 free requests / day
Larger batches GBP 0,80 / CPU hour
Language Identification for Tweets
Service to identify the languages of tweets.a.
1.200 free requests / day
Larger batches GBP 0,80 / CPU hour
English Tweet Tokenizer
Identifies words and punctuation in tweets
1.200 free requests / day
Larger batches GBP 0,80 / CPU hour
French Named Entity Recognizer
A named entity recognition service for documents in French. Based on ANNIE, it identifies names of persons, locations, and organizations.
1.200 free requests / day
Larger batches GBP 0,80 / CPU hour
Part-of-Speech Tagger for Tweets
A service that tags tweets with part-of-speech information, e.g. nouns, verbs.
1.200 free requests / day
Larger batches GBP 0,80 / CPU hour
CYMRIE Welsh Named Entity Recognizer
The CYMRIE named entity recognition service for Welsh text. Identifies namesof persons, locations, organizations, as well asmoney amounts, time and date expressions.
1.200 free requests / day
Larger batches GBP 0,80 / CPU hour
French Named Entity Recognizer for Tweets
Analyse French tweets for names of persons, locations and organizations. Also performs normalization of abbreviations and common Twitter slang.
1.200 free requests / day
Larger batches GBP 0,80 / CPU hour
German Named Entity Recognizer for Tweets
Analyse German tweets for names of persons, locations and organizations. Also performs normalization of abbreviations and common Twitter slang.
1.200 free requests / day
Larger batches GBP 0,80 / CPU hour
English Part-of-Speech and Morphology Analyzer
Annotates tokens and sentences in English text, adding part-of-speech and morphological root and affix to each token.
is
be
1.200 free requests / day
Larger batches GBP 0,80 / CPU hour
Noun Phrase Chunker
Measurement Expression Annotator
ANNIE+Measurements

GATE

GATE Cloud

English Named Entity Recognizer

Identify names of *persons, locations, organizations,* as well as *money amounts, time and date expressions* in English texts automatically.

1.200 free requests / day
Larger batches **GBP 0,80 / CPU hour**

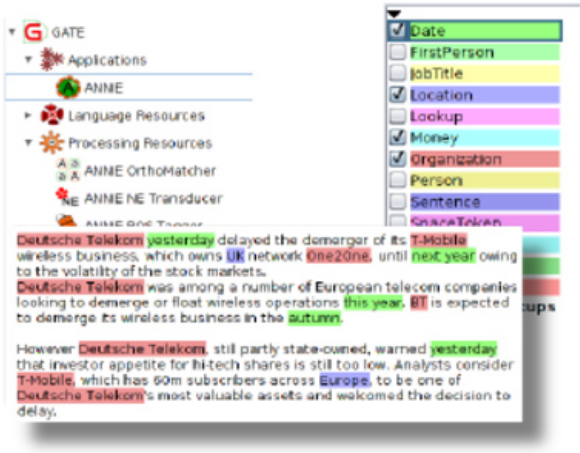
Sicuro | https://cloud.gate.ac.uk/shopfront/displayItem/annie-named-entity-recognizer

English Named Entity Recognizer

ANNIE is a named entity recognition pipeline that identifies basic entity types, such as *Person, Location, Organization, Money* amounts, *Time* and *Date* expressions.

It is the prototypical information extraction pipeline distributed with the GATE framework and forms the base of many more complex GATE-based IE applications.

Annotation details



Test this pipeline

Type the content to annotate:

Coca-Cola made its world debut at the Jacobs' Pharmacy soda fountain in Atlanta, where it sold for 5 cents a glass in 1886.
• In the first year Coca-Cola creator John Pemberton sold an average of just nine glasses a day. The company now sells 1.4 billion beverage servings every day.
• John Pemberton died in 1888 without realising the success of the beverage he had created.
• Asa Griggs Candler, an Atlanta businessman, bought up the rights to the business between 1888 and 1891 for a total of $2,300. By 1895, the drink was in demand nationwide and Candler had built syrup plants in Chicago, Dallas and Los Angeles.

GATE

GATE Cloud

English Named Entity Recognizer

Identify names of *persons, locations, organizations,* as well as *money amounts, time and date expressions* in English texts automatically.

1.200 free requests / day
Larger batches **GBP 0,80 / CPU hour**

Or select a text file: Scegli file Nessun file selezionato

Output type: JSON

Document format: plain text

Restore defaults — Address, Date, Location, Organization, Person, Money, Percent, Token, SpaceToken, Sentence

Test Pipeline — download

Annotation types: Date, Location, Money, Organization, Person

Coca-Cola made its world debut at the Jacobs' Pharmacy soda fountain in Atlanta, where it sold for 5 cents a glass in 1886. • In the first year Coca-Cola creator John Pemberton sold an average of just nine glasses a day. The company now sells 1.4 billion beverage servings every day. • John Pemberton died in 1888 without realising the success of the beverage he had created. • Asa Griggs Candler, an Atlanta businessman, bought up the rights to the business between 1888 and 1891 for a total of $2,300. By 1895, the drink was in demand nationwide and Candler had built syrup plants in Chicago, Dallas and Los Angeles. • The men who served Coca-Cola at soda fountains were called Soda Jerks because of the jerking motion they made preparing a glass of the fizzy drink. They traditionally wore a white hat and a white coat or apron.

Click an annotation to see features

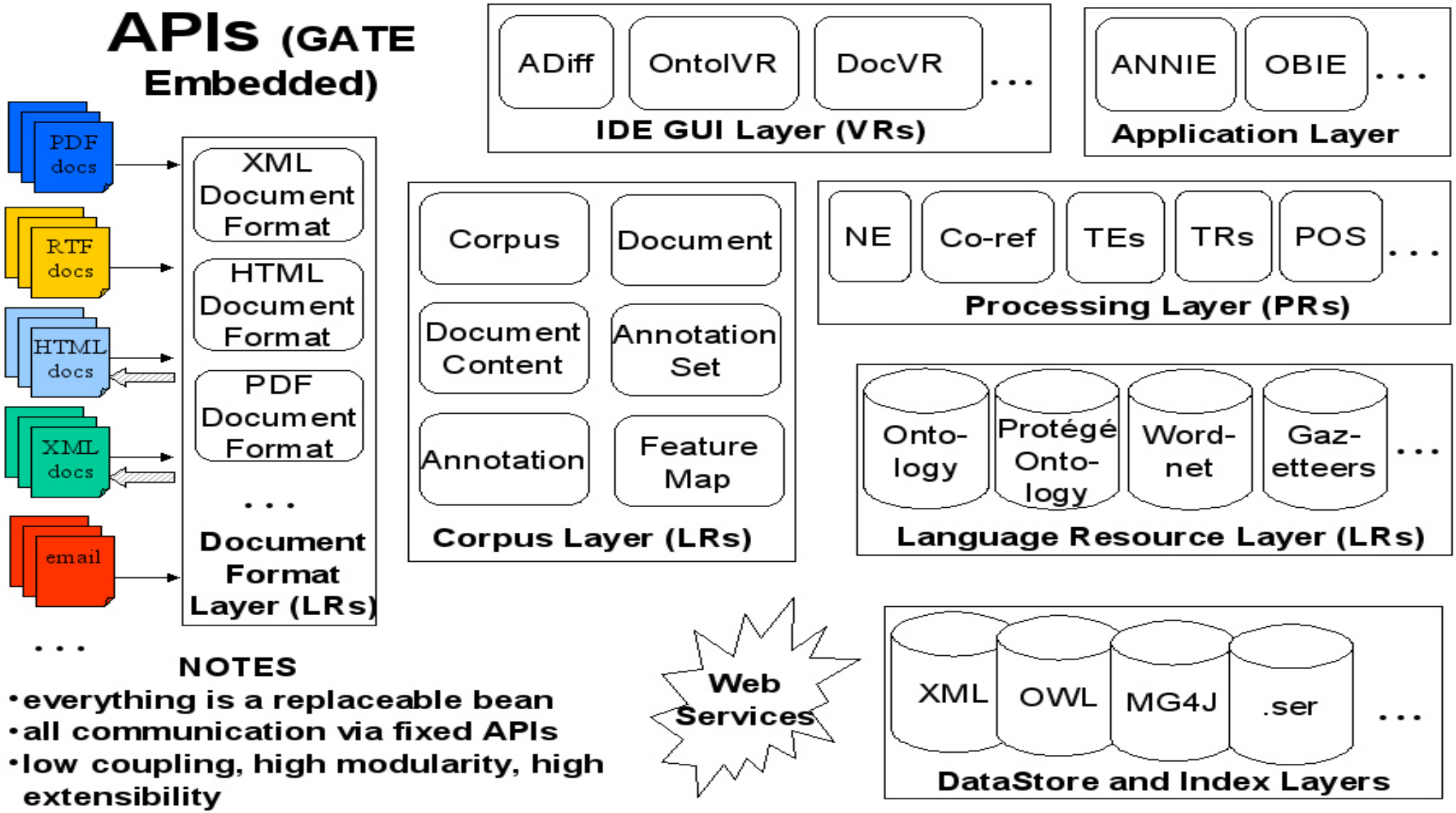
JSON
- JSON
- XML

plain text
- plain text
- HTML
- XML
- Cochrane Library
- Pubmed
- MediaWiki
- Twitter JSON
- DataSift JSON

Pratica / Esempi interattivi

GATE CLOUD.

GATE


APIs (GATE Embedded)
PDF docs
RTF docs
HTML docs
XML docs
email
. . .
XML Document Format
HTML Document Format
PDF Document Format
. . .
Document Format Layer (LRs)
ADiff
OntoIVR
DocVR
. . .
IDE GUI Layer (VRs)
ANNIE
OBIE
. . .
Application Layer
Corpus
Document
Document Content
Annotation Set
Annotation
Feature Map
Corpus Layer (LRs)
NE
Co-ref
TEs
TRs
POS
. . .
Processing Layer (PRs)
Onto-logy
Protégé Onto-logy
Word-net
Gaz-etteers
. . .
Language Resource Layer (LRs)
NOTES
•everything is a replaceable bean
•all communication via fixed APIs
•low coupling, high modularity, high extensibility
Web Services
XML
OWL
MG4J
.ser
. . .
DataStore and Index Layers

GATE

Product	Description	License
GATE Developer	an integrated development environment for language processing components bundled with the most widely used Information Extraction system and a comprehensive set of other plugins	LGPL
GATE Embedded	an object library optimised for inclusion in diverse applications giving access to all the services used by *GATE developer* and more	LGPL

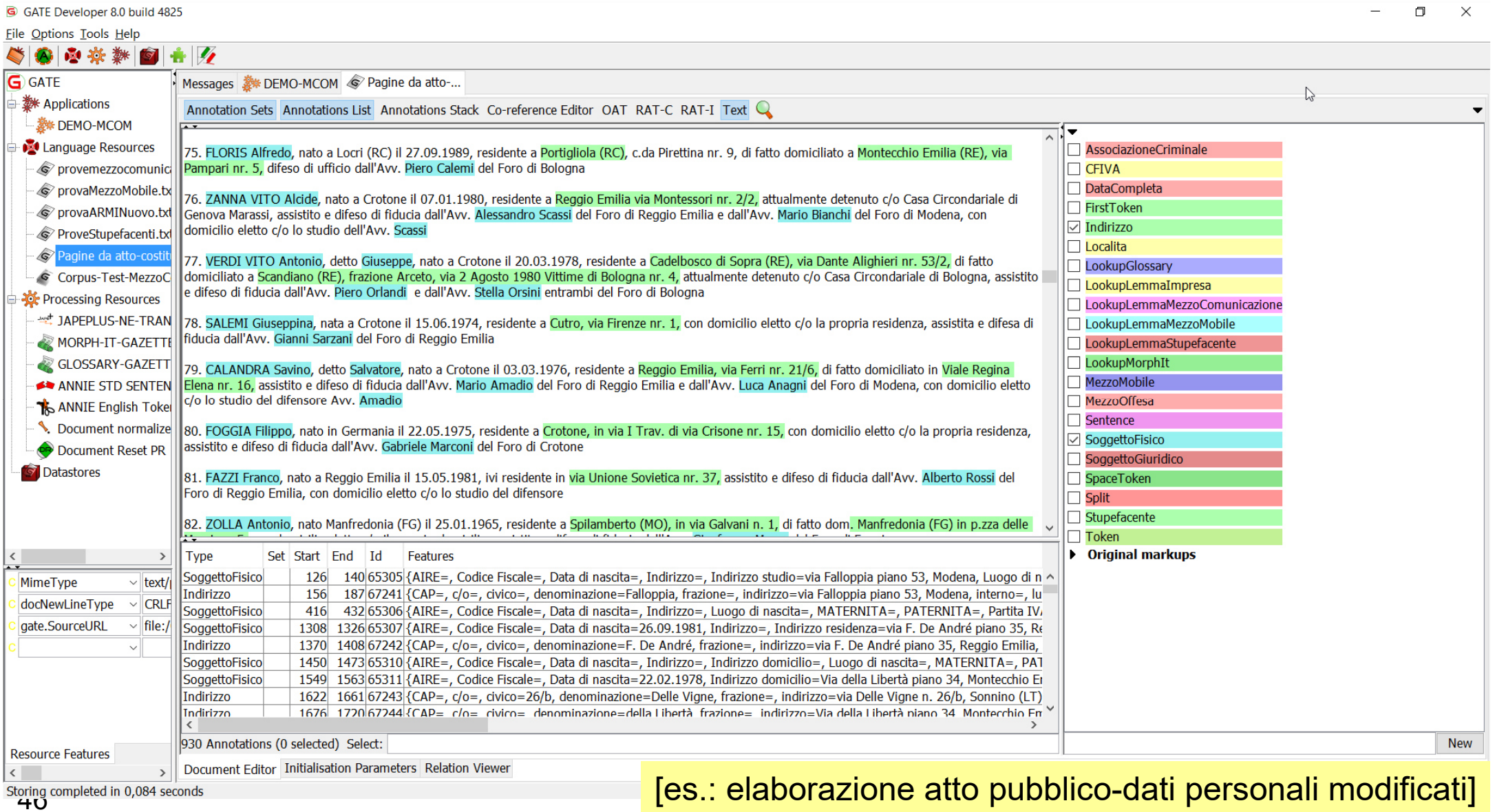
GNU Lesser General Public License (**GNU LGPL** o solo **LGPL**)

La LGPL è una licenza di tipo *copyleft* ma, a differenza della licenza *GNU GPL*, ***non richiede che eventuale software "linkato" al programma sia pubblicato sotto la medesima licenza***.

« *Un programma che non contenga alcun derivato di nessuna porzione della Libreria, ma è progettato per lavorare con la Libreria attraverso compilazione o collegamento con questa, viene definito "**un'opera che usa la Libreria**". Tale opera, isolata, non è derivata dalla Libreria, e pertanto ricade al di fuori dell'ambito di questa Licenza.* »

GATE – Developer …

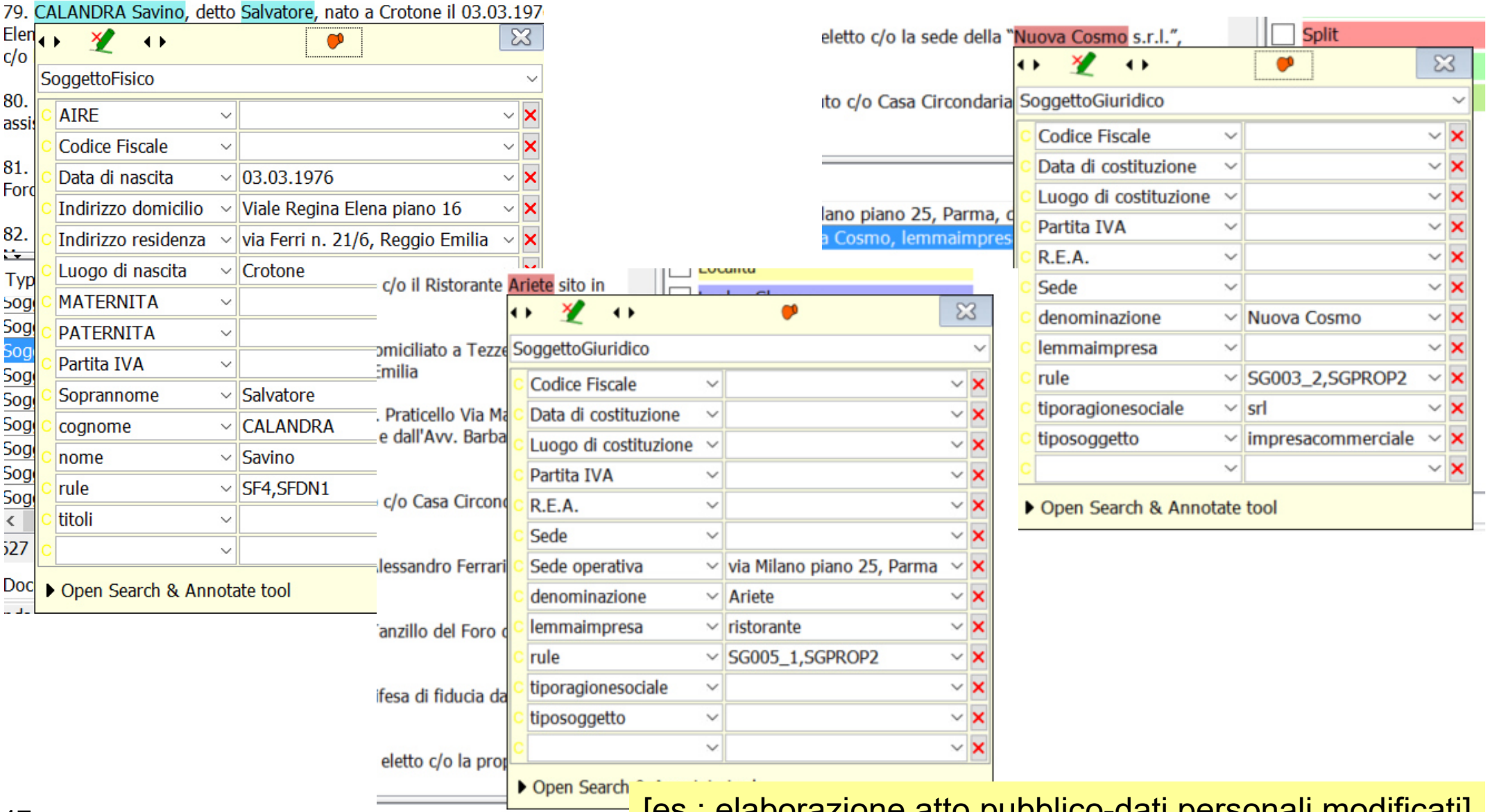
E' l'IDE di gestione di progetti per: processi, glossari, grammatiche, corpus documenti, ecc.



[es.: elaborazione atto pubblico-dati personali modificati]

… GATE – Developer

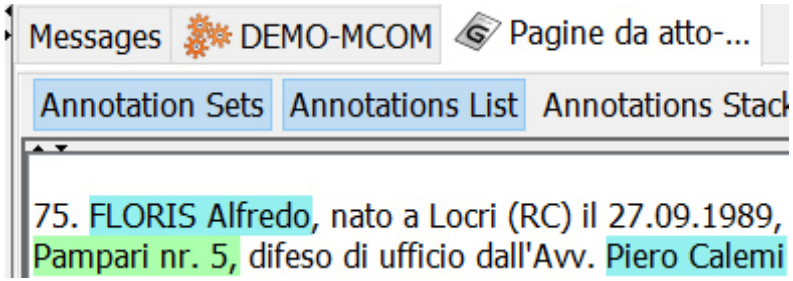
Scheda di dettaglio annotazioni dall'IDE

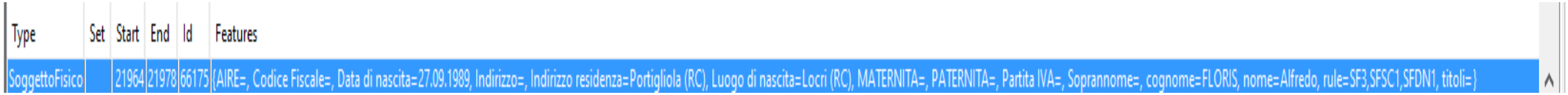


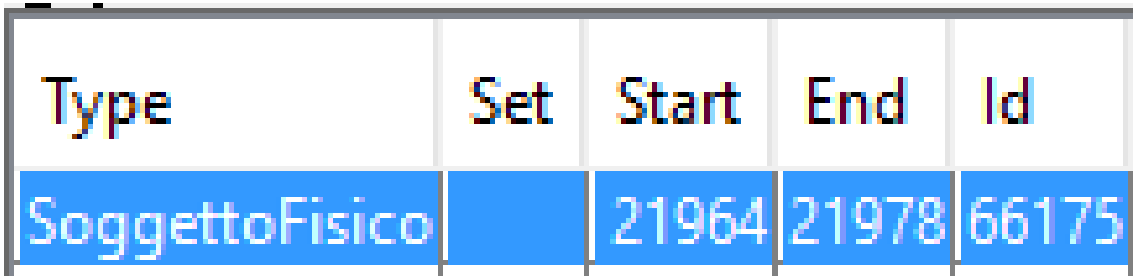
[es.: elaborazione atto pubblico-dati personali modificati]

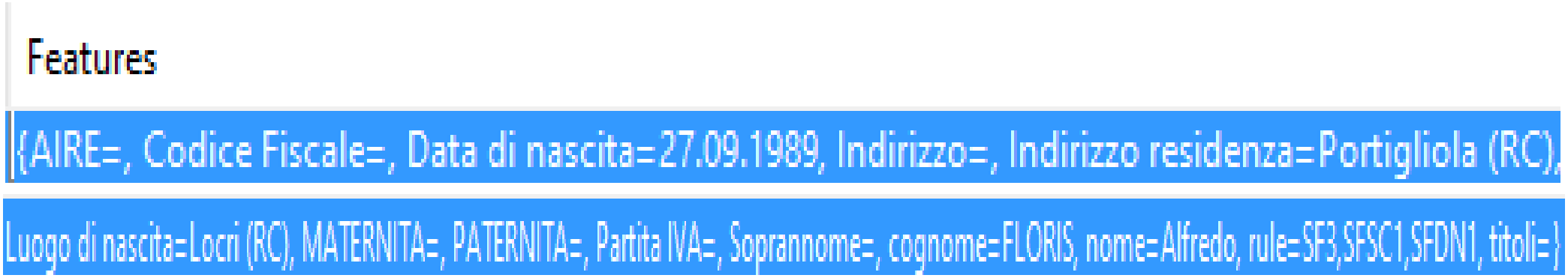
… GATE – Developer …

L'annotazione nell'IDE









[es.: elaborazione atto pubblico-dati personali modificati]

… GATE – Developer …

L'annotazione nel documento XML generato

<?xml version='1.0' encoding='UTF-8'?><GateDocument version="3"><!-- The document's features--><GateDocumentFeatures><Feature> <Name className="java.lang.String">gate.SourceURL</Name> <Value className="java.lang.String">file:/C:/roberto/progetti/Gate/MyProjectsStore/ANNIE-TEST-00005-IN-ITINERE/CorpusDocumenti/ProveDemo/Pagine%20da%20atto-costituz-parte-civile_libera.txt</Value></Feature><Feature> <Name className="java.lang.String">MimeType</Name> <Value className="java.lang.String">text/plain</Value></Feature><Feature> <Name className="java.lang.String">docNewLineType</Name> <Value className="java.lang.String">CRLF</Value></Feature></GateDocumentFeatures><!-- The document content area with serialized nodes -->

……

……..

<Node id="21964"/>**FLORIS**<Node id="21970"/> <Node id="21971"/>**Alfredo**<Node id="21978"/>,<Node id="21979"/> <Node id="21980"/>nato<Node id="21984"/> <Node id="21985"/>a<Node id="21986"/> <Node id="21987"/>**Locri**<Node id="21992"/> <Node id="21993"/>(<Node id="21994"/>**RC**<Node id="21996"/>)<Node id="21997"/> <Node id="21998"/>il<Node id="22000"/> <Node id="22001"/>**27**<Node id="22003"/>.<Node id="22004"/>**09**<Node id="22006"/>.<Node id="22007"/>**1989**<Node id="22011"/>,<Node id="22012"/> <Node id="22013"/>residente<Node id="22022"/> <Node id="22023"/>a<Node id="22024"/> <Node id="22025"/>**Portigliola**<Node id="22036"/> <Node id="22037"/>(<Node id="22038"/>**RC**<Node id="22040"/>)<Node id="22041"/>,<Node id="22042"/> <Node id="22043"/>c<Node id="22044"/>.<Node id="22045"/>da<Node id="22047"/> <Node id="22048"/>Pirettina<Node id="22057"/> <Node id="22058"/>nr<Node id="22060"/>.<Node id="22061"/> <Node id="22062"/>9<Node id="22063"/>,<Node id="22064"/> <Node id="22065"/>di<Node id="22067"/> <Node id="22068"/>fatto<Node id="22073"/> <Node id="22074"/>domiciliato<Node id="22085"/> <Node id="22086"/>a<Node id="22087"/> <Node id="22088"/>Montecchio<Node id="22098"/> <Node id="22099"/>Emilia<Node id="22105"/> <Node id="22106"/>(<Node id="22107"/>RE<Node id="22109"/>)<Node id="22110"/>,<Node id="22111"/> <Node id="22112"/>via<Node id="22115"/> <Node id="22116"/>Pampari<Node id="22123"/> <Node id="22124"/>nr<Node id="22126"/>.<Node id="22127"/> <Node id="22128"/>5<Node id="22129"/>,<Node id="22130"/> <Node id="22131"/>difeso<Node id="22137"/> <Node id="22138"/>di<Node id="22140"/> <Node id="22141"/>ufficio<Node id="22148"/> <Node id="22149"/>dall<Node id="22153"/>'<Node id="22154"/>Avv<Node id="22157"/>.<Node id="22158"/> <Node id="22159"/>Piero<Node id="22164"/> <Node id="22165"/>Calemi<Node id="22171"/> <Node id="22172"/>del<Node id="22175"/> <Node id="22176"/>Foro<Node id="22180"/> <Node id="22181"/>di<Node id="22183"/> <Node id="22184"/>Bologna<Node id="22191"/>

[es.: elaborazione atto pubblico-dati personali modificati]

… GATE – Developer …

L'annotazione nel documento XML generato

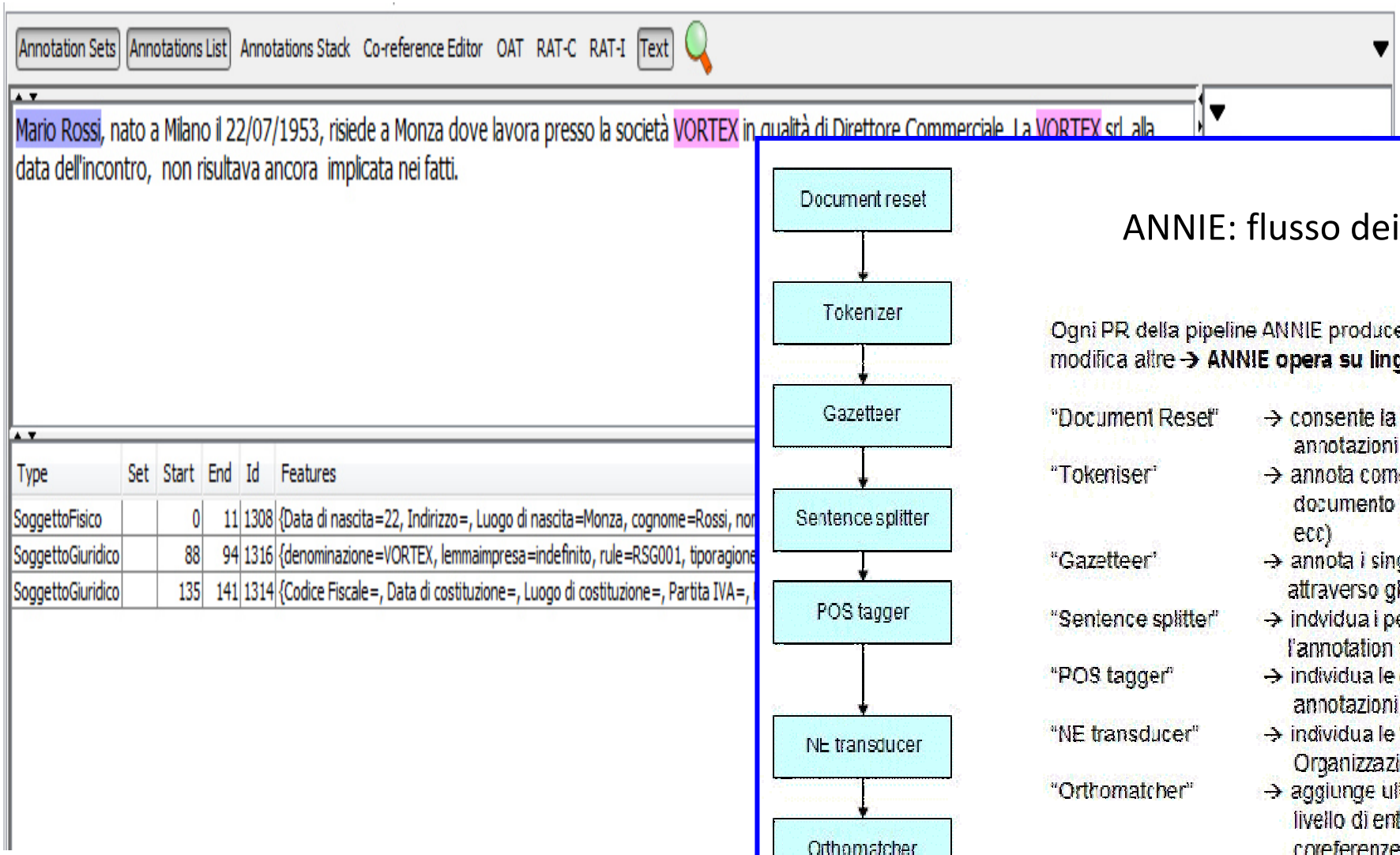
<Annotation Id="66175" Type="SoggettoFisico" StartNode="21964" EndNode="21978">
<Feature> <Name className="java.lang.String">Partita IVA</Name> <Value
className="java.lang.String"></Value></Feature><Feature> <Name
className="java.lang.String">cognome</Name> <Value
className="java.lang.String">FLORIS</Value></Feature><Feature> <Name
className="java.lang.String">AIRE</Name> <Value className="java.lang.String"></Value></Feature><Feature>
<Name className="java.lang.String">MATERNITA</Name> <Value
className="java.lang.String"></Value></Feature><Feature> <Name className="java.lang.String">Codice
Fiscale</Name> <Value className="java.lang.String"></Value></Feature><Feature> <Name
className="java.lang.String">rule</Name> <Value
className="java.lang.String">SF3,SFSC1,SFDN1</Value></Feature><Feature> <Name
className="java.lang.String">Soprannome</Name> <Value
className="java.lang.String"></Value></Feature><Feature> <Name className="java.lang.String">nome</Name>
<Value className="java.lang.String">Alfredo</Value></Feature><Feature> <Name
className="java.lang.String">Indirizzo residenza</Name> <Value className="java.lang.String">Portigliola
(RC)</Value></Feature><Feature> <Name className="java.lang.String">Indirizzo</Name> <Value
className="java.lang.String"></Value></Feature><Feature> <Name className="java.lang.String">Luogo di
nascita</Name> <Value className="java.lang.String">Locri (RC)</Value></Feature><Feature> <Name
className="java.lang.String">PATERNITA</Name> <Value
className="java.lang.String"></Value></Feature><Feature> <Name className="java.lang.String">Data di
nascita</Name> <Value className="java.lang.String">27.09.1989</Value></Feature><Feature> <Name
className="java.lang.String">titoli</Name> <Value className="java.lang.String"></Value>
</Feature>
</Annotation>

[es.: elaborazione atto pubblico-dati personali modificati]

GATE IDE

- ✓Organizzazione di IDE.
- ✓LR: Language Resources (documenti, corpus).
- ✓PR: Process Resources (la pipeline, plugin/PR).
- ✓Esecuzione di una pipeline.
- ✓Navigazione di un documento elaborato.

GATE – *Rule based …*



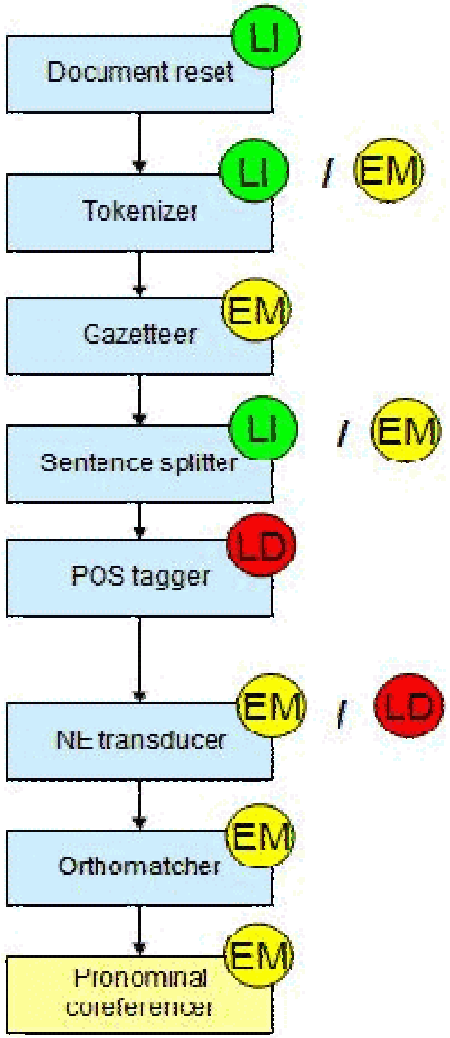
ANNIE: flusso dei processi

Ogni PR della pipeline ANNIE produce nuove annotazioni oppure ne modifica altre → **ANNIE opera su lingua INGLESE** ←

"Document Reset"	→ consente la rimozione controllata/filtrata di annotazioni
"Tokeniser"	→ annota come "token" i singoli elementi del documento (parole, numeri, punteggiatura, ecc)
"Gazetteer"	→ annota i singoli elementi associabili a liste attraverso glossari e dizionari
"Sentence splitter"	→ individua i periodi e li evidenzia con l'annotation type "sentence"
"POS tagger"	→ individua le categorie grammaticali aggiungo annotazioni specifiche alle precedenti
"NE transducer"	→ individua le "entità" (Date, Persone, Località, Organizzazioni, etc)
"Orthomatcher"	→ aggiunge ulteriori annotazioni concettuali a livello di entità nella risoluzione di coreferenze

… GATE – *Rule based* …

ANNIE:
flusso dei processi ed effetti della lingua



Ogni PR della pipeline ANNIE produce nuove annotazioni oppure ne modifica altre → **ANNIE opera su lingua INGLESE** ←

"Document Reset"	→ consente la rimozione controllata/filtrata di annotazioni
"Tokeniser"	→ annota come "token" i singoli elementi del documento (parole, numeri, punteggiatura, ecc)
"Gazetteer"	→ annota i singoli elementi associabili a liste attraverso glossari e dizionari
"Sentence splitter"	→ individua i periodi e li evidenzia con l'annotation type "sentence"
"POS tagger"	→ individua le categorie grammaticali aggiundo annotazioni specifiche alle precedenti
"NE transducer"	→ individua le "entità" (Date, Persone, Località, Organizzazioni, etc)
"Orthomatcher"	→ aggiunge ulteriori annotazioni concettuali a livello di entità nella risoluzione di coreferenze

Scenario modifiche al cambio lingua (lingue latine)

"language independent"	→ piccole modifiche
"easily modifiable"	→ deve essere modificata; abbastanza semplicemente ma con possibili interventi di programmazione
"language dependent"	→ richiede di essere interamente rivista o sostituita con altra integrabile in architettura GATE

LI EM LD

… GATE – *Rule based* …

Un flusso di processi

!	Name	Type
	Document Reset PR	Document Reset PR
	ANNIE English Tokeniser_00141	ANNIE English Tokeniser
	ANNIE STD SENTENCE SPLITTER	ANNIE Sentence Splitter
	GLOSSARY-GAZETTEER	ANNIE Gazetteer
	MORPH-IT-GAZETTEER	ANNIE Gazetteer
	JAPEPLUS-NE-TRANSDUCER	JAPE-Plus Transducer

Effetto del Tokeniser

Mario Rossi, nato a Milano il 22/07/1953, risiede a Monza dove lavora presso la società VORTEX di Vimercate.

Type	Set	Start	End	Id	Features
Token		1	6	2874	{kind=word, length=5, orth=upperInitial, string=Mario}
Token		7	12	2876	{kind=word, length=5, orth=upperInitial, string=Rossi}
Token		12	13	2877	{kind=punctuation, length=1, string=,}
Token		14	18	2879	{kind=word, length=4, orth=lowercase, string=nato}
Token		19	20	2881	{kind=word, length=1, orth=lowercase, string=a}
Token		21	27	2883	{kind=word, length=6, orth=upperInitial, string=Milano}
Token		29	31	2886	{kind=word, length=2, orth=lowercase, string=il}
Token		32	34	2888	{kind=number, length=2, string=22}
Token		34	35	2889	{kind=punctuation, length=1, string=/}
Token		35	37	2890	{kind=number, length=2, string=07}
Token		37	38	2891	{kind=punctuation, length=1, string=/}
Token		38	42	2892	{kind=number, length=4, string=1953}
Token		42	43	2893	{kind=punctuation, length=1, string=,}
Token		45	52	2896	{kind=word, length=7, orth=lowercase, string=risiede}
Token		54	55	2899	{kind=word, length=1, orth=lowercase, string=a}
Token		56	61	2901	{kind=word, length=5, orth=upperInitial, string=Monza}
Token		63	67	2904	{kind=word, length=4, orth=lowercase, string=dove}
Token		68	74	2906	{kind=word, length=6, orth=lowercase, string=lavora}
Token		76	82	2909	{kind=word, length=6, orth=lowercase, string=presso}
Token		83	85	2911	{kind=word, length=2, orth=lowercase, string=la}
Token		86	93	2913	{kind=word, length=7, orth=lowercase, string=società}
Token		94	100	2915	{kind=word, length=6, orth=allCaps, string=VORTEX}
Token		102	104	2918	{kind=word, length=2, orth=lowercase, string=di}
Token		105	114	2920	{kind=word, length=9, orth=upperInitial, string=Vimercate}
Token		114	115	2921	{kind=punctuation, length=1, string=.}

… GATE – *Rule based* …

Tokeniser

AnnotationType ***Token*** *con* ***feature*****:**

kind → word → orth → upperInitial, allCaps, lowerCase, mixedCaps)

kind → number

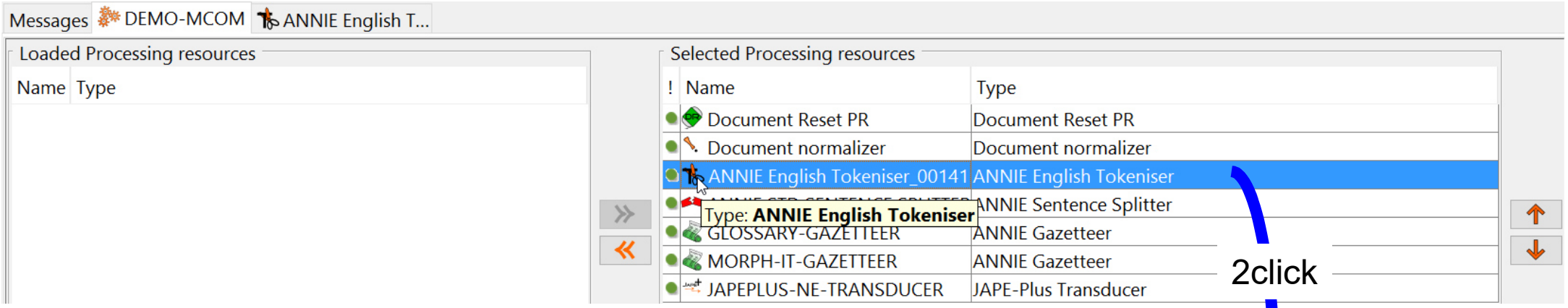
kind → punctuation → es.: **. , ; : ( )** ecc.

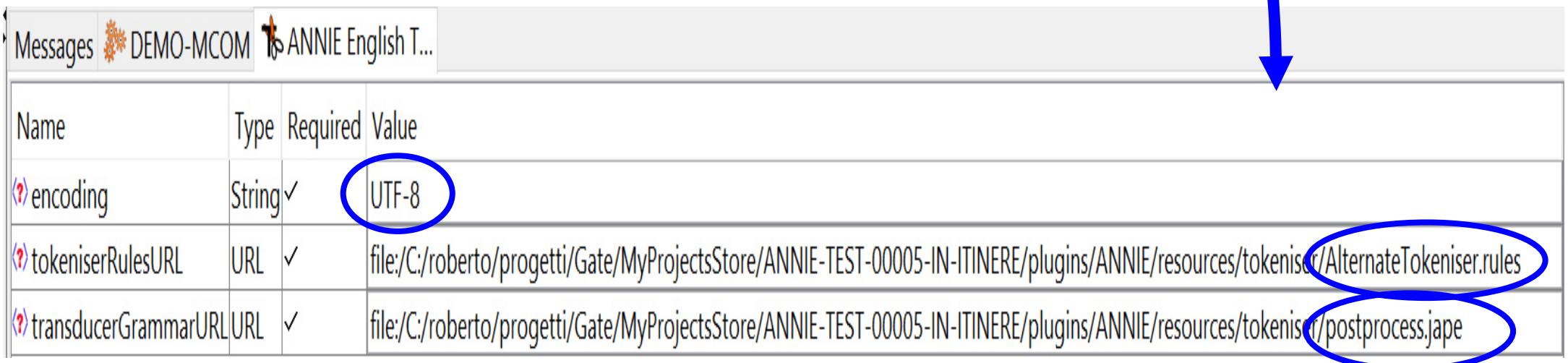
AnnotationType ***SpaceToken***

```
// Es.: TOKEN => WORD

"UPPERCASE_LETTER" (LOWERCASE_LETTER)* > Token;orth=upperInitial;kind=word;
"UPPERCASE_LETTER" (UPPERCASE_LETTER)+ > Token;orth=allCaps;kind=word;
"LOWERCASE_LETTER" (LOWERCASE_LETTER)* > Token;orth=lowercase;kind=word;

("LOWERCASE_LETTER" "LOWERCASE_LETTER"+"UPPERCASE_LETTER"+ \
 (UPPERCASE_LETTER|LOWERCASE_LETTER)*)|\
("LOWERCASE_LETTER" "LOWERCASE_LETTER"*"UPPERCASE_LETTER"+\
 (UPPERCASE_LETTER|LOWERCASE_LETTER)*)|\
("UPPERCASE_LETTER" "UPPERCASE_LETTER" (UPPERCASE_LETTER|LOWERCASE_LETTER)*\
 ("LOWERCASE_LETTER")+ (UPPERCASE_LETTER|LOWERCASE_LETTER)*)|\
("UPPERCASE_LETTER" "LOWERCASE_LETTER"+ ("UPPERCASE_LETTER"+
"LOWERCASE_LETTER"+))+\
> Token;orth=mixedCaps;kind=word;
```

… GATE – *Rule based* …

Tokeniser (alcuni elementi di configurazione)





PR Tokenizer

- ✓Esecuzione del PR.
- ✓Navigazione dei risultati sul documento.
- ✓La configurazione e le regole che governano il *tokenizer.*

… GATE – *Rule based* …

Un flusso di processi

!	Name	Type
	Document Reset PR	Document Reset PR
	ANNIE English Tokeniser_00141	ANNIE English Tokeniser
	ANNIE STD SENTENCE SPLITTER	ANNIE Sentence Splitter
	GLOSSARY-GAZETTEER	ANNIE Gazetteer
	MORPH-IT-GAZETTEER	ANNIE Gazetteer
	JAPEPLUS-NE-TRANSDUCER	JAPE-Plus Transduc

Effetto del PR "GLOSSARY-GAZETTEER"

Mario Rossi, nato a Milano il 22/07/1953, risiede a Monza dove lavora presso la società VORTEX di Vimercate.

Type	Set	Start	End	Id	Features	
LookupGlossary		1	6	2938	{language=IT, majorType=nome, minorType=maschio}	→ "Mario"
LookupGlossary		21	27	2941	{language=IT, majorType=localita, minorType=comune}	
LookupGlossary		21	27	2942	{language=IT, majorType=localita, minorType=provincia}	→ "Milano"
LookupGlossary		56	61	2946	{language=IT, majorType=localita, minorType=comune}	→ "Monza"
LookupGlossary		105	114	2950	{language=IT, majorType=localita, minorType=comune}	→ "Vimercate"

… GATE – *Rule based* …

Glossari del PR "GLOSSARY-GAZETTEER"

Messages | APPVERIFICAF4 | GLOSSARY-GAZETT…

Cognomi.lst | Add

List name	Major	Minor	Language	Annotation type
NonParteIndFraz.lst	nonparteindfraz	undef	IT	LookupGlossary
Numeri.lst	numero	numerolett	IT	LookupGlossary
NumeriRomani.lst	numero	numeroromano	IT	LookupGlossary
Ordinali.lst	numero	numeroord	IT	LookupGlossary
PrefissiCognome.lst	cognome	prefisso	IT	LookupGlossary
PrefissiSFAmbigui.lst	prefissosf	ambiguo	IT	LookupGlossary
PrefissoCostituzione.lst	prefissocost	undef	IT	LookupGlossary
PrefissoDataNascita.lst	prefdatanascita	undef	IT	LookupGlossary
PrefissoImplicitoResDom.lst	implicitoind	undef	IT	LookupGlossary
PrefissoLuogoNascita.lst	prefluogonascita	undef	IT	LookupGlossary
PrefissoMaternita.lst	prefmaternita	undef	IT	LookupGlossary
PrefissoNickame.lst	prefissonick	undef	IT	LookupGlossary
PrefissoPaternita.lst	prefpaternita	undef	IT	LookupGlossary
PrefissoReg.lst	prefissoreg	undef	IT	LookupGlossary
PrefissoResidenza.lst	residenza	prefisso	IT	LookupGlossary
ProvinceItaliane.lst	localita	provincia	IT	LookupGlossary
QualificatoreCAP.lst	qualificatorecap	undef	IT	LookupGlossary
QualificatoreCodFis.lst	qualificatoreCF	undef	IT	LookupGlossary
QualificatoreFrazione.lst	frazione	qualificatore	IT	LookupGlossary
QualificatoreIndirizzo.lst	indirizzo	qualificatore	IT	LookupGlossary
QualificatoreNumCiv.lst	qualificatorecivico	undef	IT	LookupGlossary
QualificatoreNumero.lst	qualificatorenumerogenerico	undef	IT	LookupGlossary
QualificatoreNumInt.lst	qualificatoreinterno	undef	IT	LookupGlossary
QualificatoreParIVA.lst	qualificatorePIVA	undef	IT	LookupGlossary
QualificatorePiano.lst	qualificatorepiano	undef	IT	LookupGlossary
QualificatorePresso.lst	qualificatorepresso	undef	IT	LookupGlossary
QualificatoreScala.lst	qualificatorescala	undef	IT	LookupGlossary
QualificatoreSubCiv.lst	qualificatoresubciv	undef	IT	LookupGlossary
QualificatoreVicinanza.lst	qualificatorevicinanza	undef	IT	LookupGlossary
RegioniItaliane.lst	localita	regione	IT	LookupGlossary
SenzaNumCiv.lst	senzanumcivico	undef	IT	LookupGlossary
SfFormeDubbie.lst	sfformedubbie	undef	IT	LookupGlossary
SigleProvinceItaliane.lst	localita	siglaprovincia	IT	LookupGlossary
SoggettiGiuridici.lst	organizzazione	sogggiuridico	IT	LookupGlossary
SoggettiGiuridiciAmbigui.lst	organizzazione	sogggiuridicoambiguo	IT	LookupGlossary
Tempo.lst	tempo	misuratempo	IT	LookupGlossary
TipiAppartenenzaAssociazioneCriminale.lst	tipoappasscrim	undef	IT	LookupGlossary
TipiRagioneSociale.lst	organizzazione	tiposoggiur	IT	LookupGlossary
TipiRapportoPersone.lst	rapportopersona	undef	IT	LookupGlossary
TipiRapportoPersoneSoggGiur.lst	rapportopersonasgiur	undef	IT	LookupGlossary
TipoAssociazioneCriminale.lst	orgcrim	tipoasscrim	IT	LookupGlossary
TipoOrganizzazioneCriminale.lst	orgcrim	tipoorgcrim	IT	LookupGlossary

Value
d
da
dagli
dai
dal
dall
dalla
dalle
dallo
de
degli
dei
del
dell
della
delle
dello
di
le
li
lo

Glossari del PR "GLOSSARY-GAZETTEER"

… GATE – *Rule based* …

lists.def

```
1  AnagrafiNaz_Loc.lst:anagrafe:undef:IT:LookupGlossary
2  Anni.lst:data:anno:IT:LookupGlossary
3  AssociazioniCriminali.lst:orgcrim:asscriminali:IT:LookupGlossary
4  AttivitaIllecita.lst:orgcrim:attivitallecita:IT:LookupGlossary
5  AttivitaLavorativa.lst:attivitalav:undef:IT:LookupGlossary
6  AvverbiNumeraliRomani.lst:numero:avvnumromano:IT:LookupGlossary
7  Classe.lst:classeanno:undef:IT:LookupGlossary
8  CognomiAmbigui.lst:cognome:ambiguo:IT:LookupGlossary
9  CognomiNonAmbigui.lst:cognome:nonAmbiguo:IT:LookupGlossary
```

CognomiAmbigui.lst

```
1  rossi
2  russo
3  ferrari
4  bianchi
5  romano
6  colombo
7  ricci
8  marino
9  greco
10 bruno
11 gallo
12 conti
13 mancini
14 costa
15 lombardi
16 barbieri
17 fontana
```

TipiRagioneSociale.lst

```
1  societa' cooperativa a r.l.~tiposoggetto=impresacommerciale~tiporagionesociale=scarl
2  società cooperativa a r.l.~tiposoggetto=impresacommerciale~tiporagionesociale=scarl
3  scarl~tiposoggetto=impresacommerciale~tiporagionesociale=scarl
4  s.c.a r.l~tiposoggetto=impresacommerciale~tiporagionesociale=scarl
5  soc. coop.~tiposoggetto=impresacommerciale~tiporagionesociale=soc.coop
6  soc.coop.~tiposoggetto=impresacommerciale~tiporagionesociale=soc.coop
7  soc coop~tiposoggetto=impresacommerciale~tiporagionesociale=soc.coop
8  soc. coop~tiposoggetto=impresacommerciale~tiporagionesociale=soc.coop
9  soc.coop~tiposoggetto=impresacommerciale~tiporagionesociale=soc.coop
10 s.c.~tiposoggetto=impresacommerciale~tiporagionesociale=soc.coop
11 sarl~tiposoggetto=impresacommerciale~tiporagionesociale=srl
12 srl~tiposoggetto=impresacommerciale~tiporagionesociale=srl
13 s.r.l.~tiposoggetto=impresacommerciale~tiporagionesociale=srl
14 s.r.l~tiposoggetto=impresacommerciale~tiporagionesociale=srl
15 spa~tiposoggetto=impresacommerciale~tiporagionesociale=spa
16 s.p.a~tiposoggetto=impresacommerciale~tiporagionesociale=spa
17 s.p.a.~tiposoggetto=impresacommerciale~tiporagionesociale=spa
18 snc~tiposoggetto=impresapersonale~tiporagionesociale=snc
19 s.n.c~tiposoggetto=impresapersonale~tiporagionesociale=snc
20 s.n.c.~tiposoggetto=impresapersonale~tiporagionesociale=snc
```

… GATE – *Rule based* …

Glossari del PR "GLOSSARY-GAZETTEER"

Messages | DEMO-MCOM | Pagine da atto-... | GLOSSARY-GAZETT...

Cognomi.lst | Add

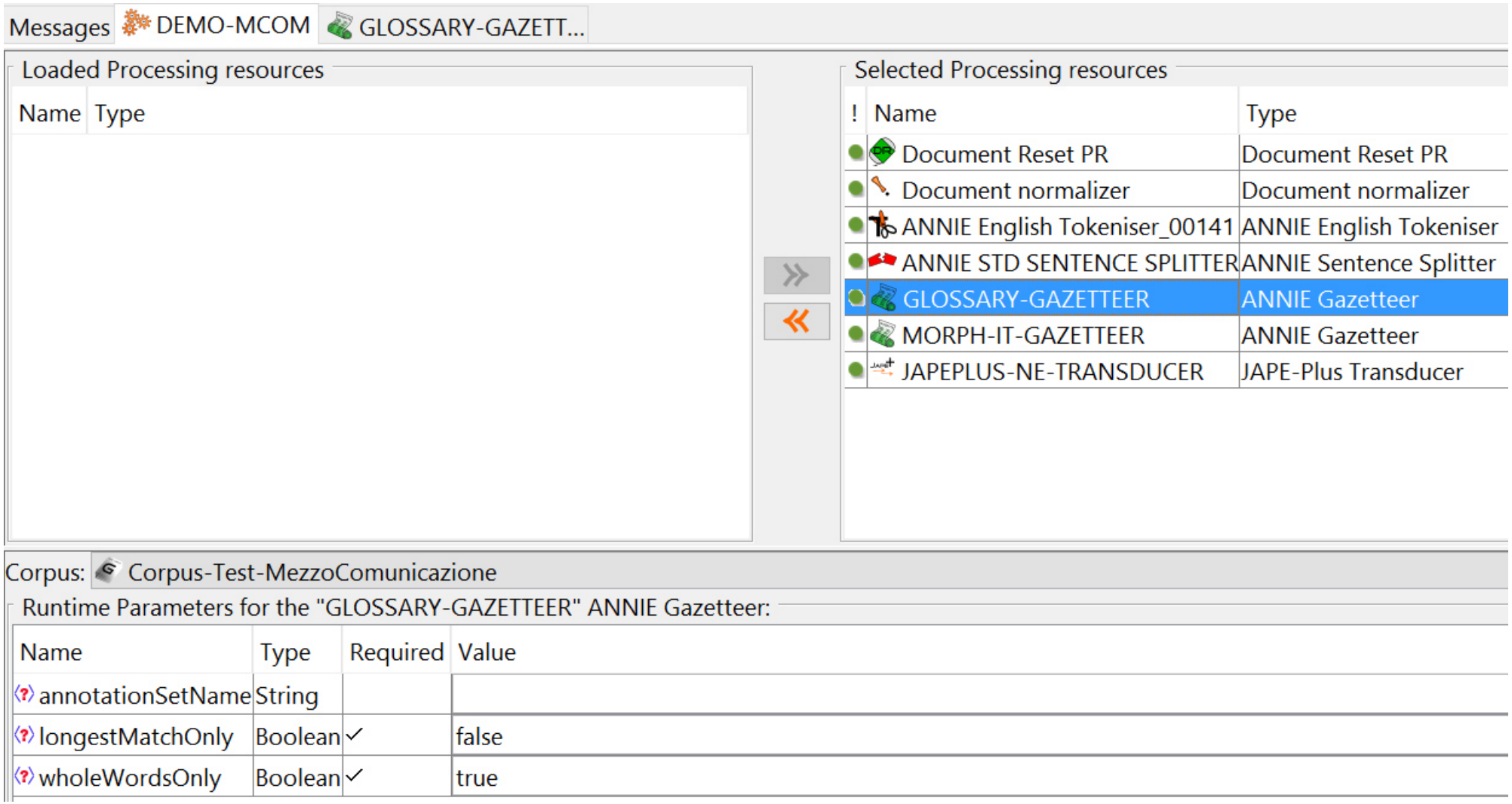
List name	Major	Minor	Lan...	Annotation ty...
QualificatoreFrazione.lst	frazione	qualificatore	IT	LookupGlossary
QualificatoreGestoreTel.l...	qualificatoreges...	undef	IT	LookupGlossary
QualificatoreIMEI.lst	qualificatoreIMEI	undef	IT	LookupGlossary
QualificatoreIndirizzo.lst	indirizzo	qualificatore	IT	LookupGlossary
QualificatoreNumCiv.lst	qualificatorecivi...	undef	IT	LookupGlossary
QualificatoreNumInt.lst	qualificatoreinte...	undef	IT	LookupGlossary
QualificatoreNumTel.lst	qualificatorenu...	undef	IT	LookupGlossary
QualificatoreNumTelInt.l...	qualificatoreinte...	undef	IT	LookupGlossary
QualificatoreNumero.lst	qualificatorenu...	undef	IT	LookupGlossary
QualificatorePIN.lst	qualificatorePIN	undef	IT	LookupGlossary
QualificatoreParIVA.lst	qualificatorePIVA	undef	IT	LookupGlossary
QualificatorePiano.lst	qualificatorepia...	undef	IT	LookupGlossary
QualificatorePresso.lst	qualificatorepre...	undef	IT	LookupGlossary
QualificatoreQuantita.lst	qualificatorequa...	undef	IT	LookupGlossary
QualificatoreScala.lst	qualificatorescala	undef	IT	LookupGlossary
QualificatoreSubCiv.lst	qualificatoresub...	undef	IT	LookupGlossary
QualificatoreVicinanza.lst	qualificatorevici...	undef	IT	LookupGlossary
RegioniItaliane.lst	localita	regione	IT	LookupGlossary
SenzaNumCiv.lst	senzanumcivico	undef	IT	LookupGlossary
SfFormeDubbie.lst	sfformedubbie	undef	IT	LookupGlossary
SigleProvinceItaliane.lst	localita	siglaprovincia	IT	LookupGlossary
SoggettiGiuridici.lst	organizzazione	soggiuridi...	IT	LookupGlossary
SoggettiGiuridiciAmbia...	organizzazione	soggiuridi...	IT	LookupGlossary
Tempo.lst	tempo	misuratempo	IT	LookupGlossary
TipiAppartenenzaAssoci...	tipoappasscrim	undef	IT	LookupGlossary
TipiRagioneSociale.lst	organizzazione	tiposoggiur	IT	LookupGlossary
TipiRapportoPersone.lst	rapportopersona	undef	IT	LookupGlossary
TipiRapportoPersoneSo...	rapportoperson...	undef	IT	LookupGlossary
TipoArma.lst	arma	undef	IT	LookupGlossary
TipoAssociazioneCrimin...	orgcrim	tipoasscrim	IT	LookupGlossary
TipoMezzoMobile.lst	tipomezzomobile	undef	IT	LookupGlossary
TipoOrganizzazioneCri...	orgcrim	tipoorgcrim	IT	LookupGlossary
TipoStupefacente.lst	stupefacente	undef	IT	LookupGlossary
TitoliPersonaF.lst	titpersonale	femmina	IT	LookupGlossary
TitoliPersonaM.lst	titpersonale	maschio	IT	LookupGlossary
TitoliStudioPersona.lst	titstpersona	undef	IT	LookupGlossary
UnitaMisura.lst	um	undef	IT	LookupGlossary
UnitaMisuraStupefacent...	umstupefacenti	undef	IT	LookupGlossary
UnitaOrdinePlurale.lst	ordineunita	plurale	IT	LookupGlossary

Filter | Add | +Cols | 46 entries | Case Ins. | Regex | Value

Value	Feature 1	Value 1	Feature 2	Value 2
aq	tiposoggetto	impresacommerciale	tiporagionesociale	aq
aq.	tiposoggetto	impresacommerciale	tiporagionesociale	aq
co	tiposoggetto	impresacommerciale	tiporagionesociale	co
co.	tiposoggetto	impresacommerciale	tiporagionesociale	co
inc	tiposoggetto	impresacommerciale	tiporagionesociale	inc
inc.	tiposoggetto	impresacommerciale	tiporagionesociale	inc
llc	tiposoggetto	impresacommerciale	tiporagionesociale	llc
llc.	tiposoggetto	impresacommerciale	tiporagionesociale	llc
ltd	tiposoggetto	impresacommerciale	tiporagionesociale	ltd
ltd.	tiposoggetto	impresacommerciale	tiporagionesociale	ltd
m.b.h.	tiposoggetto	impresacommerciale	tiporagionesociale	mbh
mbh	tiposoggetto	impresacommerciale	tiporagionesociale	mbh
nl	tiposoggetto	impresacommerciale	tiporagionesociale	nl
nl.	tiposoggetto	impresacommerciale	tiporagionesociale	nl
oq	tiposoggetto	impresacommerciale	tiporagionesociale	oq
oq.	tiposoggetto	impresacommerciale	tiporagionesociale	oq
plc	tiposoggetto	impresacommerciale	tiporagionesociale	plc
plc.	tiposoggetto	impresacommerciale	tiporagionesociale	plc
pty	tiposoggetto	impresacommerciale	tiporagionesociale	pty
pty.	tiposoggetto	impresacommerciale	tiporagionesociale	pty
s.a.	tiposoggetto	impresapersonale	tiporagionesociale	sacc
s.a.s	tiposoggetto	impresapersonale	tiporagionesociale	sas
s.a.s.	tiposoggetto	impresapersonale	tiporagionesociale	sas
s.acc.	tiposoggetto	impresapersonale	tiporagionesociale	sacc
s. acc.	tiposoggetto	impresapersonale	tiporagionesociale	sacc
s.c.	tiposoggetto	impresacommerciale	tiporagionesociale	soc.coop
s.c.a r.l	tiposoggetto	impresacommerciale	tiporagionesociale	scarl
s.n.c	tiposoggetto	impresapersonale	tiporagionesociale	snc
s.n.c.	tiposoggetto	impresapersonale	tiporagionesociale	snc

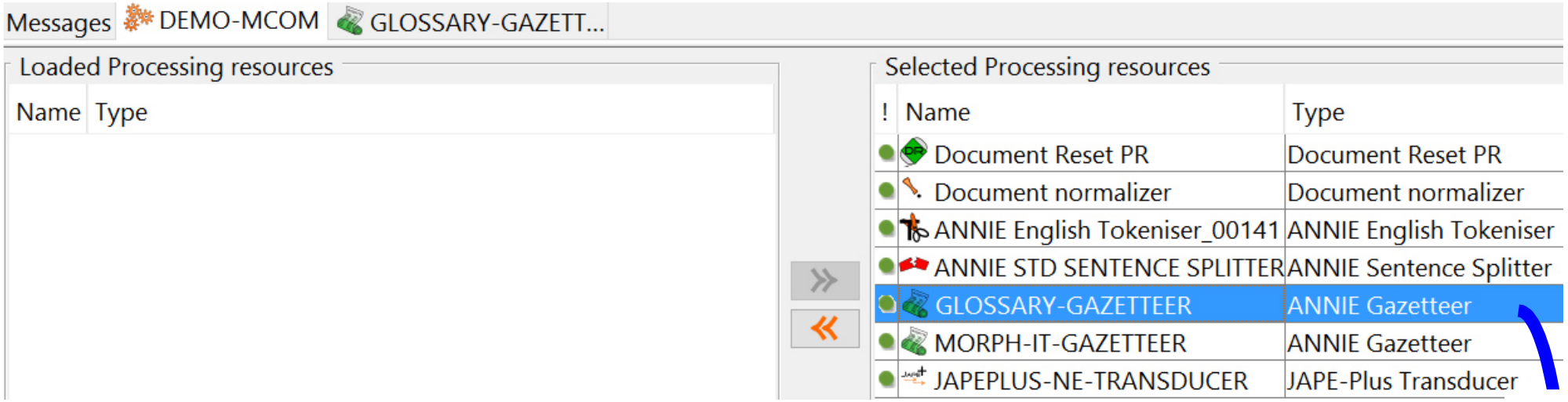
… GATE – *Rule based* …

Glossari del PR “GLOSSARY-GAZETTEER” (alcuni elementi di configurazione)

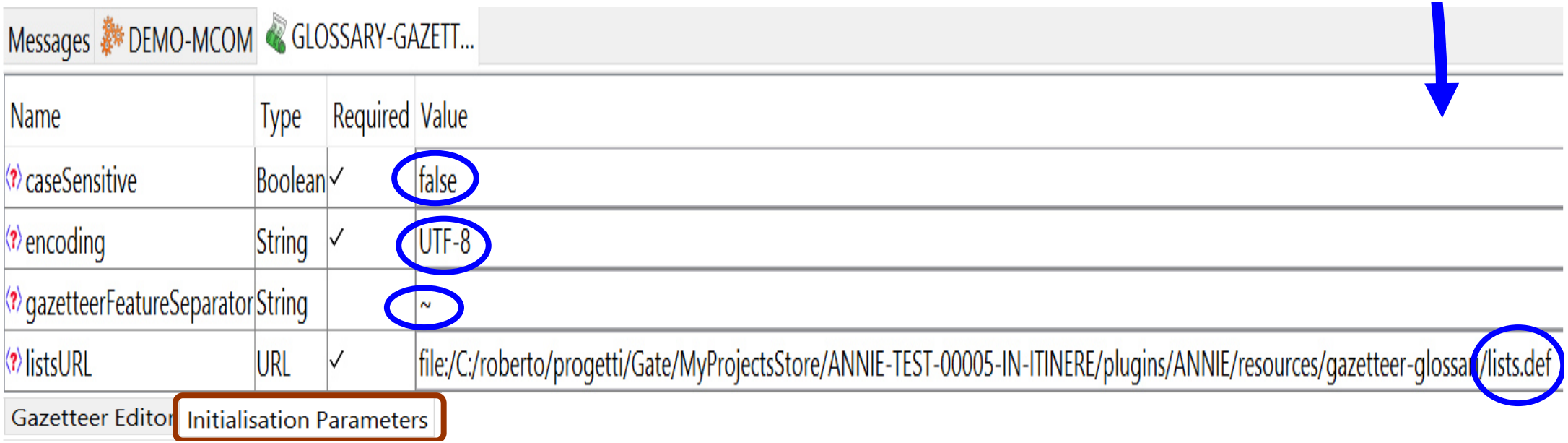


… GATE – *Rule based* …

Glossari del PR "GLOSSARY-GAZETTEER" (alcuni elementi di configurazione)



2click



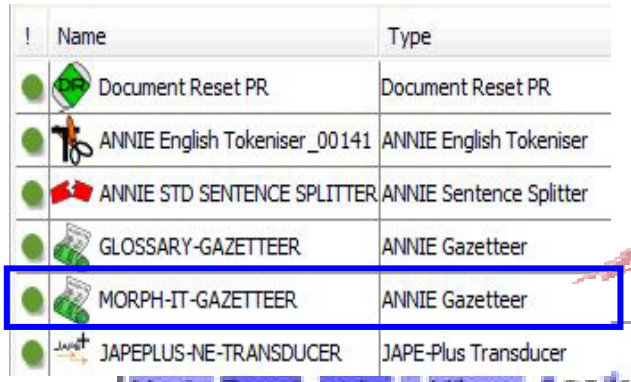
Pratica / Esempi interattivi

PR “GLOSSARY-GAZETTEER”

- ✓Esecuzione del PR.
- ✓Navigazione dei risultati sul documento.
- ✓La navigazione/gestione da IDE dei glossari.
- ✓ La configurazione e la forma dei file di glossario.

… GATE – *Rule based* …

Un flusso di processi



Effetto del PR "MORPH-IT-GAZETTEER"

Mario Rossi, nato a Milano il 22/07/1953, risiede a Monza dove lavora presso la società VORTEX di Vimercate.

Type	Set	Start	End	Id	Features
LookupMorphIt		1	6	2951	{language=IT, lemma=Mario, majorType=MajorUndef, minorType=MinorUndef, morfologia=NPR}
LookupMorphIt		7	12	2952	{language=IT, lemma=rosso, majorType=MajorUndef, minorType=MinorUndef, morfologia=NOUN-M:p}
LookupMorphIt		7	12	2953	{language=IT, lemma=Rossi, majorType=MajorUndef, minorType=MinorUndef, morfologia=NPR}
LookupMorphIt		7	12	2954	{language=IT, lemma=rosso, majorType=MajorUndef, minorType=MinorUndef, morfologia=ADJ:pos+m+p}
LookupMorphIt		14	18	2955	{language=IT, lemma=nascere, majorType=MajorUndef, minorType=MinorUndef, morfologia=VER:part+past+s+m}
LookupMorphIt		14	18	2956	{language=IT, lemma=Nato, majorType=MajorUndef, minorType=MinorUndef, morfologia=NPR}
LookupMorphIt		14	18	2957	{language=IT, lemma=nato, majorType=MajorUndef, minorType=MinorUndef, morfologia=ADJ:pos+m+s}
LookupMorphIt		14	18	2958	{language=IT, lemma=nato, majorType=MajorUndef, minorType=MinorUndef, morfologia=NOUN-M:s}
LookupMorphIt		19	20	2959	{language=IT, lemma=a, majorType=MajorUndef, minorType=MinorUndef, morfologia=PRE}

Morph-it → 504.906 forme e 34.968 lemma

Marco Baroni e Eros Zanchetta

http://sslmitdev-online.sslmit.unibo.it/linguistics/morph-it.php

Form	*Lemma*	*Features*
rimpinzeremmo	*rimpinzare*	*VER:cond+pre+1+p*
abominevole	*abominevole*	*ADJ:pos+m+s*
dabbenaggine	*dabbenaggine*	*NOUN-F:s*
ostensibilmente	*ostensibilmente*	*ADV*

… GATE – *Rule based* …

Effetto del PR "MORPH-IT-GAZETTEER"

Morph-it originale (txt)

```
aizzo   aizzare VER:ind+pres+1+s
aizzò   aizzare VER:ind+past+3+s
al  al  ARTPRE-M:s
ala ala NOUN-F:s
ala alare   VER:impr+pres+2+s
ala alare   VER:ind+pres+3+s
alacre  alacre  ADJ:pos+f+s
alacre  alacre  ADJ:pos+m+s
alacremente alacremente ADV
alacri  alacre  ADJ:pos+f+p
alacri  alacre  ADJ:pos+m+p
alacrissima alacre  ADJ:sup+f+s
alacrissime alacre  ADJ:sup+f+p
alacrissimi alacre  ADJ:sup+m+p
alacrissimo alacre  ADJ:sup+m+s
alai    alare   VER:ind+past+1+s
```

readme-morph-it.txt

```
================================================================
                        Morph-it!

      A free morphological lexicon for the Italian Language
================================================================

                       version 0.4.8
                       February 23 2009

****************************************************************
                THIS README IS NOT REALLY UP TO DATE
                     A NEW VERSION OF THIS
                      README FILE WILL BE
                   RELEASED (HOPEFULLY) SOON
                 (BUT I WOULDN'T COUNT ON THAT...)
****************************************************************

                      Copyright (c) 2004-2009
                Marco Baroni  (marco.baroni@unitn.it)
                Eros Zanchetta (eros@sslmit.unibo.it)

                   http://sslmit.unibo.it/morphit


Morph-it! is a free (as in free speech and in free beer) morphological
resource for the Italian language.

Morph-it! is a lexicon of inflected forms with their lemma and
morphological features. For example:

gattini     gattino     NOUN-M:p
andarono    andare      VER:ind+past+3+p
fastidiosetto   fastidioso  ADJ:dim+m+s
```

Effetto del PR "MORPH-IT-GAZETTEER"

… GATE – *Rule based* …

Morph-it Gazetteer

Come "sorgente" (txt) →

```
aizzo~lemma=aizzare~morfologia=VER:ind+pres+1+s
aizzò~lemma=aizzare~morfologia=VER:ind+past+3+s
al~lemma=al~morfologia=ARTPRE-M:s
ala~lemma=ala~morfologia=NOUN-F:s
ala~lemma=alare~morfologia=VER:impr+pres+2+s
ala~lemma=alare~morfologia=VER:ind+pres+3+s
alacre~lemma=alacre~morfologia=ADJ:pos+f+s
alacre~lemma=alacre~morfologia=ADJ:pos+m+s
alacremente~lemma=alacremente~morfologia=ADV
alacri~lemma=alacre~morfologia=ADJ:pos+f+p
alacri~lemma=alacre~morfologia=ADJ:pos+m+p
alacrissima~lemma=alacre~morfologia=ADJ:sup+f+s
alacrissime~lemma=alacre~morfologia=ADJ:sup+f+p
alacrissimi~lemma=alacre~morfologia=ADJ:sup+m+p
alacrissimo~lemma=alacre~morfologia=ADJ:sup+m+s
alai~lemma=alare~morfologia=VER:ind+past+1+s
```

Da IDE ↓

List name	Major	Minor	Language	Annotation type
DBmorph-it.lst	MajorUndef	MinorUndef	IT	LookupMorphIt
Lemmaimpresa.lst	lemmaimpresa	lemmaimpresa	IT	LookupLemmaImpresa
Lemmamezzocomunicazione.lst	lemmamezzocomunicazione	lemmamezzocomunicazione	IT	LookupLemmaMezzoC
Lemmamezzomobile.lst	lemmamezzomobile	lemmamezzomobile	IT	LookupLemmaMezzoM
Lemmastupefacente.lst	lemmastupefacente	lemmastupefacente	IT	LookupLemmaStupefa

Value	Feature 1	Value 1	Feature 2	Value 2
aizzò	lemma	aizzare	morfologia	VER:ind+past+3+s
Aja	lemma	Aja	morfologia	NPR
Ajax	lemma	Ajax	morfologia	NPR
al	lemma	al	morfologia	ARTPRE-M:s
Al	lemma	Al	morfologia	NPR
ala	lemma	ala	morfologia	NOUN-F:s
ala	lemma	alare	morfologia	VER:impr+pres+2+s
ala	lemma	alare	morfologia	VER:ind+pres+3+s
alacre	lemma	alacre	morfologia	ADJ:pos+f+s
alacre	lemma	alacre	morfologia	ADJ:pos+m+s
alacremente	lemma	alacremente	morfologia	ADV
alacri	lemma	alacre	morfologia	ADJ:pos+f+p
alacri	lemma	alacre	morfologia	ADJ:pos+m+p
alacrissima	lemma	alacre	morfologia	ADJ:sup+f+s
alacrissime	lemma	alacre	morfologia	ADJ:sup+f+p
alacrissimi	lemma	alacre	morfologia	ADJ:sup+m+p
alacrissimo	lemma	alacre	morfologia	ADJ:sup+m+s
alai	lemma	alare	morfologia	VER:ind+past+1+s

… GATE – *Rule based* …

Glossari derivati da Morph-it per PR "MORPH-IT-GAZETTEER"

List name	Major	Minor	Language	Annotation type
DBmorph-it.lst	MajorUndef	MinorUndef	IT	LookupMorphIt
Lemmaimpresa.lst	lemmaimpresa	lemmaimpresa	IT	LookupLemmaImpresa
Lemmamezzocomunicazione.lst	lemmamezzocomunicazione	lemmamezzocomunicazione	IT	LookupLemmaMezzoComunicazione
Lemmamezzomobile.lst	lemmamezzomobile	lemmamezzomobile	IT	LookupLemmaMezzoMobile
Lemmastupefacente.lst	lemmastupefacente	lemmastupefacente	IT	LookupLemmaStupefacente

Value	Feature 1	Value 1	Feature 2	Value 2
officine	lemma	officina	morfologia	NOUN-F:p
onlus	lemma	associazione	morfologia	NOUN-F:s
pescheria	lemma	pescheria	morfologia	NOUN-F:s
pescherie	lemma	pescheria	morfologia	NOUN-F:p

List name	Major	Minor	Language	Annotation type
DBmorph-it.lst	MajorUndef	MinorUndef	IT	LookupMorphIt
Lemmaimpresa.lst	lemmaimpresa	lemmaimpresa	IT	LookupLemmaImpresa
Lemmamezzocomunicazione.lst	lemmamezzocomunicazione	lemmamezzocomunicazione	IT	LookupLemmaMezzoComunicazione
Lemmamezzomobile.lst	lemmamezzomobile	lemmamezzomobile	IT	LookupLemmaMezzoMobile
Lemmastupefacente.lst	lemmastupefacente	lemmastupefacente	IT	LookupLemmaStupefacente

Value	Feature 1	Value 1	Feature 2	Value 2
chiamarci	lemma	chiamare	morfologia	VER:inf+pres+ci
chiamare	lemma	chiamare	morfologia	VER:inf+pres
chiamarla	lemma	chiamare	morfologia	VER:inf+pres+la
chiamarle	lemma	chiamare	morfologia	VER:inf+pres+le
chiamarli	lemma	chiamare	morfologia	VER:inf+pres+li

List name	Major	Minor	Language	Annotation type
DBmorph-it.lst	MajorUndef	MinorUndef	IT	LookupMorphIt
Lemmaimpresa.lst	lemmaimpresa	lemmaimpresa	IT	LookupLemmaImpresa
Lemmamezzocomunicazione.lst	lemmamezzocomunicazione	lemmamezzocomunicazione	IT	LookupLemmaMezzoComunicazione
Lemmamezzomobile.lst	lemmamezzomobile	lemmamezzomobile	IT	LookupLemmaMezzoMobile
Lemmastupefacente.lst	lemmastupefacente	lemmastupefacente	IT	LookupLemmaStupefacente

Value	Feature 1	Value 1	Feature 2	Value 2
fermate	lemma	fermare	morfologia	VER:impr+pres+2+p
fermate	lemma	fermare	morfologia	VER:ind+pres+2+p
fermate	lemma	fermare	morfologia	VER:part+past+p+f
fermatela	lemma	fermare	morfologia	VER:impr+pres+2+p+
fermateli	lemma	fermare	morfologia	VER:impr+pres+2+p+

List name	Major	Minor	Language	Annotation type
DBmorph-it.lst	MajorUndef	MinorUndef	IT	LookupMorphIt
Lemmaimpresa.lst	lemmaimpresa	lemmaimpresa	IT	LookupLemmaImpresa
Lemmamezzocomunicazione.lst	lemmamezzocomunicazione	lemmamezzocomunicazione	IT	LookupLemmaMezzoComunicazione
Lemmamezzomobile.lst	lemmamezzomobile	lemmamezzomobile	IT	LookupLemmaMezzoMobile
Lemmastupefacente.lst	lemmastupefacente	lemmastupefacente	IT	LookupLemmaStupefacente

Value	Feature 1	Value 1	Feature 2	Value 2
aspirasser	lemma	aspirare	morfologia	VER:sub+imp
aspirassero	lemma	aspirare	morfologia	VER:sub+imp
aspirassero	lemma	aspirare	morfologia	VER:sub+imp
aspirassi	lemma	aspirare	morfologia	VER:sub+imp
aspirassi	lemma	aspirare	morfologia	VER:sub+imp

Pratica / Esempi interattivi

PR “MORPH-IT GAZETTEER”

- ✓Esecuzione del PR.
- ✓Navigazione dei risultati sul documento.
- ✓La navigazione/gestione da IDE dei glossari.
- ✓ La configurazione e la forma dei file di glossario.

… GATE – *Rule based* …

Un flusso di processi

!	Name	Type
	Document Reset PR	Document Reset PR
	ANNIE English Tokeniser_00141	ANNIE English Tokeniser
	ANNIE STD SENTENCE SPLITTER	ANNIE Sentence Splitter
	GLOSSARY-GAZETTEER	ANNIE Gazetteer
	MORPH-IT-GAZETTEER	ANNIE Gazetteer
	JAPEPLUS-NE-TRANSDUCER	JAPE-Plus Transducer

Effetto del PR "JAPEPLUS-NE-TRANSDUCER"

Mario Rossi, nato a Milano il 22/07/1953, risiede a Milano dove lavora presso la società VORTEX in qualità di Direttore Commerciale. La VORTEX srl, alla data dell'incontro, non risultava ancora implicata nei fatti.

Type	Set	Start	End	Id	Features
SoggettoFisico		0	11	317	{AIRE=, Codice Fiscale=, Data di nascita=22/07/1953, Indirizzo=, Indirizzo residenza=Milano, Luogo di nascita=Milano, MATER

Mario Rossi, nato a Milano il 22/07/1953, risiede a Milano dove lavora presso la società VORTEX in qualità di Direttore Commerciale. La VORTEX srl, alla data dell'incontro, non risultava ancora implicata nei fatti.

Type	Set	Start	End	Id	Features
SoggettoGiuridico		89	95	1509	{Codice Fiscale=, Data di costituzione=, Luogo di costituzione=, Partita IVA=, R.E.A.=, Sede=, denomin
SoggettoGiuridico		136	142	1510	{Codice Fiscale=, Data di costituzione=, Luogo di costituzione=, Partita IVA=, R.E.A.=, Sede=, denomin

… GATE – *Rule based* …

```
//Rossi Mario
//BIANCHI Davide Maria
//Monache Vincenzo
(
    (
        (
            (
                {Token.kind == word, Token.length > 1, Token.orth != lowercase,
                !LookupGlossary.majorType == titstpersona, Token notWithin {LookupGlossary.majorType == organizzazione},
                 Token notWithin {LookupGlossary.majorType == sfformedubbie}, Token notWithin {LookupGlossary.majorType == prefissosf}}
                (AA)?
            )[1,3]
        ):cognome
        (
            (
                {LookupGlossary.majorType == nome, Token.orth != lowercase, Token notWithin {LookupGlossary.majorType == sfformedubbie}}
            )[1,2]
            |
            (INIZIALI)
        ):nome
    ):sf4
    // il pattern che segue ha il fine di escludere casi di identificazione come SF di nomi puntato seguito immediatamente da  un number
    // spesso frequenti in dizioni del tipo XXXXX N. 1, XXXXX Q. 2 ecc. Es. ALLEGATO N. 1..
    (
        {Token.kind != number}
        |
        {Split}
    )
) :bindNop
-->
:sf4.SoggettoFisico = {rule = "SF4", nome = :nome@string, cognome = :cognome@string, titoli=""}
```

```
//Apostrofo, accento
Macro: AA
(
    ({Token.kind == punctuation, Token.string == "'"})|
    ({Token.kind == punctuation, Token.string == "’"})|
    ({Token.kind == punctuation, Token.string == "\`"})|
    ({Token.kind == punctuation, Token.string == "`"})|
    ({Token.kind == punctuation, Token.string == "´"})|
    ({Token.kind == punctuation, Token.string == "‘"})|
    ({Token.kind == punctuation, Token.string == "’"})|
    ({Token.kind == punctuation, Token.string == "\'"})|
    ({Token.kind == punctuation, Token.string == "’"})|
    ({Token.kind == punctuation, Token.string == "´"})|
    ({Token.kind == punctuation, Token.string == "'"})
)
```

… GATE – *Rule based* …

```
// riconoscimento domicilio o residenza, solo cap (opzionale), località obbligatoria, eventuale frazione
// prefisso residenza obbligatorio, altrimenti si replica annotazione di località
Rule: Ind03
Priority: 80
(
   (
    (PrefRes)
    (
        (
            (CAP)?
            (Separatore)?
            ({Localita}):loc
            (PROVINCIA)?:prov
            (
             (Separatore)?
             (QualificatoreFrazMenoDi)
             (DenominazioneFrazione)
            )?
        )
     ):tagInd
   )
:match)--> RHSofRule
```

```
// RHS per tutte le regole di questo file jape
Macro: RHSofRule
:match
{
    // La rilettura del glossario DenIndEscluse.lst avviene una volta solo all'esecuzione della prima
    // fase first_token.jape nella sezione ControllerStarted: {} e ripreso qui attraverso la feature
    // "DenIndEscluseArray" associata al controller dell'applicazione
    // In questo modo la rilettura avviene per ogni documento e non per ogni regola che applica la macro RHS.

    // determina il controller dell'applicazione
    Controller Cntr = ctx.getController();

    // legge i valori del glossario posti come stringa (valori separati da "|"
    // splitting per "|" e ricostruzione di denIndEscluse come ArrayList<String>
    // Questo metodo è impiegato poiché non si è compreso come passare dal controller in
    // sezione ControllerStarted: {} direttamente l'ArrayList<String>
    String GlossaryArray = null;
    String[] SplitGlossaryArray = null;

    // DenIndEscluse.lst => denIndEscluse
    GlossaryArray = Cntr.getFeatures().get("DenIndEscluseArray").toString();
    SplitGlossaryArray = GlossaryArray.split("\\|");
    ArrayList<String> denIndEscluse = new ArrayList<String>(Arrays.asList(SplitGlossaryArray));

    // DaEscludere.4rh => daEscludere
    GlossaryArray = Cntr.getFeatures().get("DaEscludereArray").toString();
    SplitGlossaryArray = GlossaryArray.split("\\|");
    List<String> daEscludere = new ArrayList<String>(Arrays.asList(SplitGlossaryArray));
```

… GATE – *Rule based* …

MultiPhase: Test-NE-Transducer
Phases:

first_phase_controls
first_token
cf_iva

date_transducing
soggetto_fisico
scambio_cognomi

cognomi
stupefacente

soggetto_straniero

coreferenza_soggetto_fisico

localita
indirizzo
soggetto_giuridico
associazione_criminale
proprieta_soggetto_fisico
proprieta_soggetto_giuridico
ambiguita_01_soggettogiuridico

mezzo_offesa

clear_sf
clear_sg

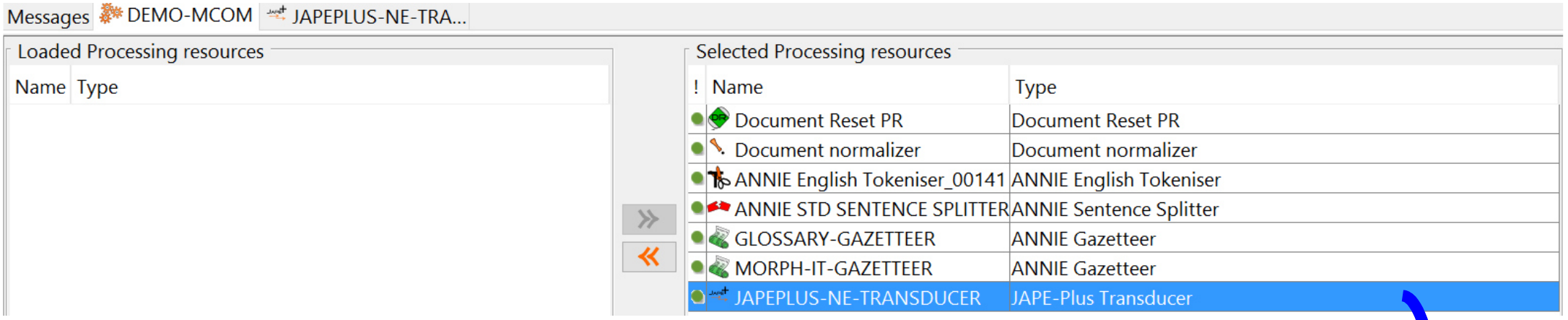
recupero_cognomi
recupero_soggettogiuridico
mezzo_comunicazione_email
mezzo_comunicazione_tel_fax
mezzo_comunicazione_IMEI
mezzo_comunicazione_PIN
mezzo_comunicazione_sequenze

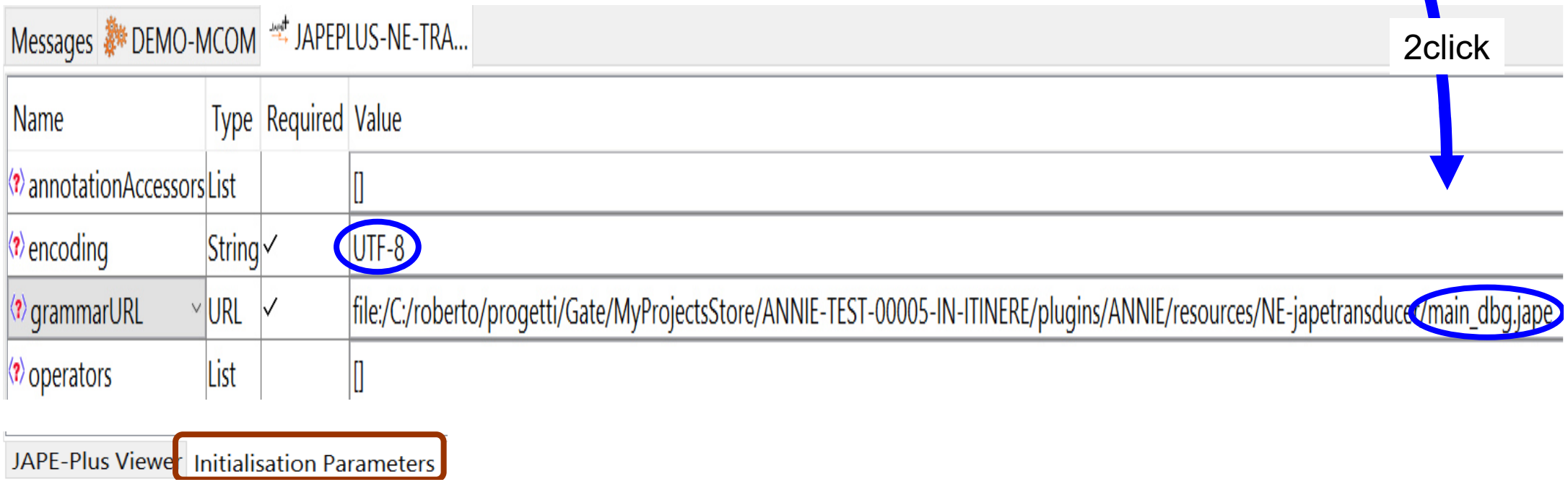
mezzo_mobile
mezzo_mobile_sequenze

clear_mc
clear_mm
proprieta_mezzo_comunicazione
proprieta_mezzo_mobile
clear_sf_post
clear_sg_post
clear_loc
clear_ind
clear

… GATE – *Rule based* …

NE JAPEPLUS TRANSDUCER (elementi notevoli di configurazione)





… GATE – *Rule based* …

Una "grammatica" JAPE (file txt UTF-8 *.jape)

Phase: <nome Fase (grammatica)>

// annotazioni input:
Input: LookupGlossary Token DataCompleta SoggettoFisico CFIVA Localita Indirizzo Split

// controllo di priorità:
Options: control = appelt negationGrouping = false

// sezione delle macro
<macro 1>
...
<macro n>
...

// sezione delle regole
<regola 1>
....
<regola m>

… GATE – *Rule based* …

Una "grammatica" JAPE (file txt UTF-8 *.jape)

```
// ************************************************************************
// M A C R O .....
// ************************************************************************
Macro: <nomeMacro>
(
        ...
        .......
)
.........
Macro: <nomeMacroRegolaRhs>
:<nomeBindingMacroRhs>
{
        ....
        .........
}
```

… GATE – *Rule based* …

Una "grammatica" JAPE (file txt UTF-8 *.jape)

```
// ************************************************************************
// J A P E   R U L E S .....
// ************************************************************************
Rule: <nomeRegola>
Priority: 100
(
        ...
        .......
)
-->
:<nomeBindingMacroRhs>)--> <nomeMacroRegolaRhs>
        oppure
:<annotation binding>.<annotation type> = {rule = "<nome regola>",
                <feature name> = :<binding>@string, <feature name> = :<binding>@string, .....}
```

LHS: Left Hand Side

RHS: Right Hand Side

Sul Left Hand Side di una regola di una "grammatica" JAPE

… GATE – *Rule based* …

```
Phase: prove
Input:  Token LookupGlossary Split
Options: control = appelt // negationGrouping = false

Rule: PersFisEsSempl
// Attenzione che è solo un esempio incompleto di tutti i casi da gestire !!!
Priority: 100

//Rossi Mario
//BIANCHI Davide Maria
//Monache Vincenzo
(
    (
        (
            {Token.kind == word, Token.length > 2, Token.orth != lowercase,
             Token notWithin {LookupGlossary.majorType == sfformedubbie}}
         ) [1,3]
     ):cognome
    (
        (    {LookupGlossary.majorType == nome, Token.orth != lowercase}  )  [1,2]
        |
        (
            {Token.orth == upperInitial, Token.kind == "word", Token.length == 1}
            {Token.kind == "punctuation", Token.string == "."}
         ) [1,2]
     ):nome
):sf4
-->
:sf4.SoggettoFisicoEsempio = {rule = "PersFisEsSempl", nome = :nome@string, cognome = :cognome@string}
```

… GATE – *Rule based* …

Sul Left Hand Side di una regola di una "grammatica" JAPE

☑ SoggettoFisicoEsempio

Marco Rossi => qui c'è prima il nome
Maria Bianchi => qui c'è prima il nome
Verdi Aldo => OK prima cognome e poi nome il glossario
BIANCHI Davide Maria => OK cognome seguito da due nomi
Bianchi aldo => qui è cognome e nome ma nome è minuscolo
bianchi Mario => qui è cognome nome ma cognome è minuscolo

- Confermato Mario in data 22/03/2017 => "Confermato" (cognome) è in "sfformedubbie"
- Confermato Aldo in data 23/03/2017 => "Confermato" (cognome) è in "sfformedubbie"
- Confermati Gianni e Salvatore in data 24/03/2017 => "Confermati" (cognome) NON è in "sfformedubbie"

Rossi C. => OK cognome e iniziale
Verdi P. G. => prende entrambe le iniziali
Bianchi P. g. => qui la seconda iniziale è minuscola
bianchi P. => qui il cognome è minuscolo
Bianchi P. G. R. => max iniziali sono 2 e quindi non comprende la terza

ecc. ecc. ecc.

Type	Set	Start	End	Id	Features
SoggettoFisicoEsempio		81	91	11325	{cognome=Verdi, nome=Aldo, rule=PersFisEsSempl}
SoggettoFisicoEsempio		137	157	11326	{cognome=BIANCHI, nome=Davide Maria, rule=PersFisEsSempl}
SoggettoFisicoEsempio		511	528	11327	{cognome=Confermati, nome=Gianni, rule=PersFisEsSempl}
SoggettoFisicoEsempio		615	623	11328	{cognome=Rossi, nome=C., rule=PersFisEsSempl}

… GATE – *Rule based* …

Sul Left Hand Side di una regola di una "grammatica" JAPE

Sugli operatori impiegabili in LHS (Left Hand Side) di una regola

"*Equality operators*": **==** **!=**

Es.: Token.kind == word … Token.orth != lowercase

"*Comparison operators*": **>** **<** **>=** **<=**

Es.: Token.length > 2

"*Regular exprexions*": =~ (match contenuto) ==~ (match completo) !~ !=~

Es.: {Token.string =~ "[Gg]atti"}

"*Contextual operators*":

{X contains Y} {X notContains Y} {X within Y} {X notWithin Y}

dove **Y** può a sua volta essere un *constraints* es.: {X contains **{Y.foo==bar}**}

Es.: Token notWithin {LookupGlossary.majorType == sfformedubbie}

… GATE – *Rule based* …

Sulla *NEGATION*

{ Token.kind == word, !LookupGlossary.majorType == nome }

È verificata per un'annotation type ***Token*** che nella proprietà ***kind*** abbia il valore ***word*** AND allo ***StartOffset*** non ci sia un'annotation type ***LookupGlossary*** con proprietà ***majorType*** al valore ***nome*** **MA ANCHE quando non ci sia nessuna annotation type LookupGlossary** allo StartOffset.

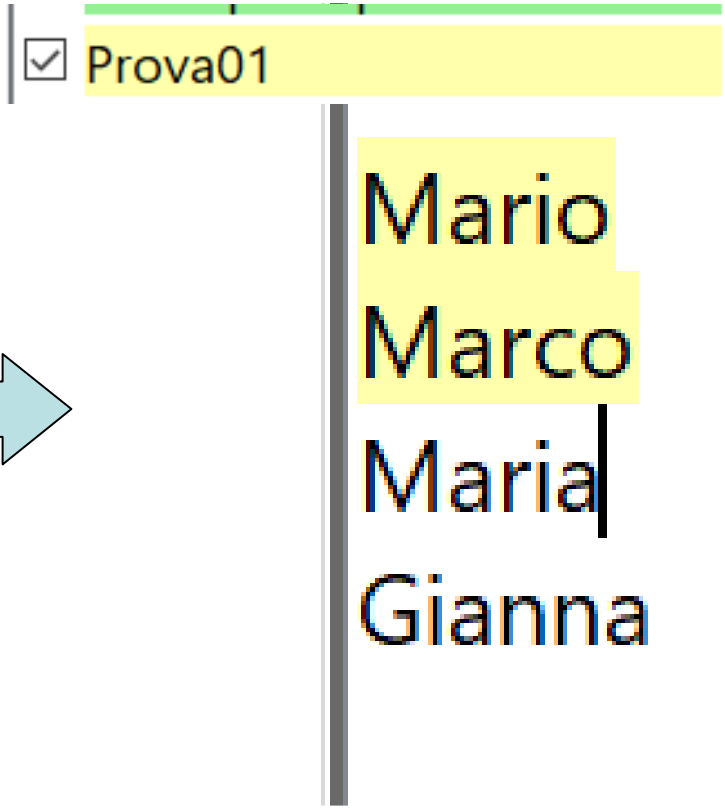
{ Token.kind == word, LookupGlossary.majorType != nome }

È verificata per un'annotation type ***Token*** che nella prorprietà ***kind*** abbia il valore ***word*** AND allo ***StartOffset*** ci sia un'annotation type ***LookupGlossary*** con proprietà ***majorType*** al valore ***nome. NOTA il non esserci allo StartOffset un' annotation Type LookupGlossary impedisce la verifica del pattern.***

… GATE – *Rule based* …

Sulla *NEGATION*

```
Rule: Prova01
Priority: 100
(
      {Token.kind == word,
            !LookupGlossary.majorType == "nome", !LookupGlossary.minorType == "femmina"}
):pr01
-->
:pr01.Prova01 = {rule = "Prova01"}
```



Perché ?

… GATE – *Rule based* …

Sulla *NEGATION*

{Token.kind == word,
!LookupGlossary.majorType == "nome", !LookupGlossary.minorType == "femmina"}

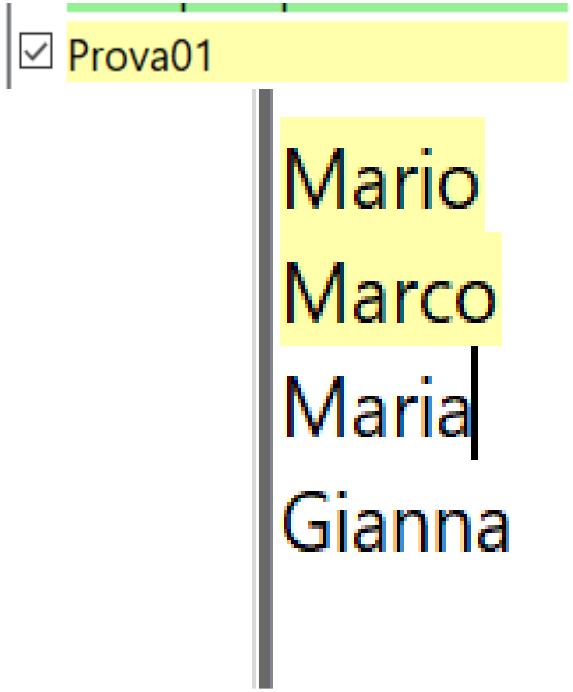
"Mario", "Marco" sono un "nome" ma "maschio" e sono annotati perché, in un {constraint}, di default le negazioni sulla stessa annotation type sono **dipendenti** (equivale a un implicito *Options: … negationGrouping = true …*) che equivale a un **OR** dei constraint di negazione sulla **medesima annotation type**

Es. **Mario**:

[A] Token.kind == word → true
[B] !LookupGlossary.majorType == "nome" → false
[C] !LookupGlossary.minorType == "femmina"} → true

[A] AND ([B] **OR** [C])

true AND (false **OR** true) → true → matching → annotazione !



… GATE – *Rule based* …

Sulla *NEGATION*

{Token.kind == word,
!LookupGlossary.majorType == "nome", !LookupGlossary.minorType == "femmina"}

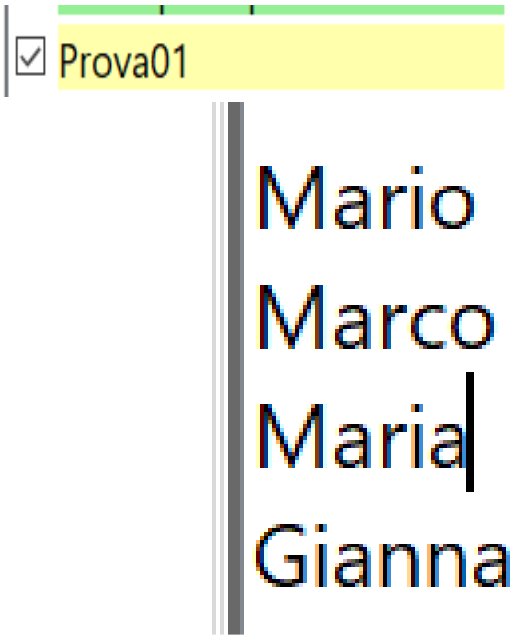
Ma **fissando in testata della grammatica**
Options: … negationGrouping = false
Si impone per ogni regola della grammatica di effettuare un **AND** dei constraint di negazione sulla **medesima annotation type** → le negazioni sulla stessa annotation type sono **indipendenti**

Es. **Mario**:

[A] Token.kind == word → true
[B] !LookupGlossary.majorType == "nome“ → false
[C] !LookupGlossary.minorType == "femmina"} → true

[A] AND ([B] **AND** [C])

True AND (false **AND** true) → false → no matching → no annotazione !



… GATE – *Rule based* …

Sulla *NEGATION*

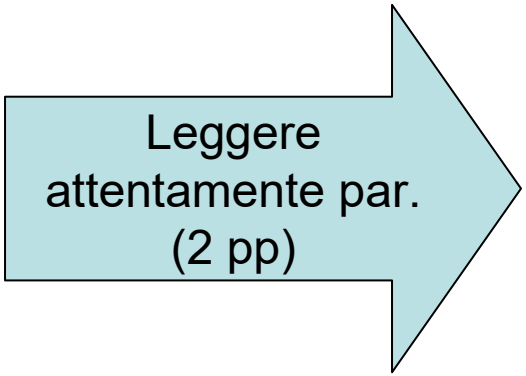
Comunque constraint negati su **annotation type** differenti vengono sempre trattati come **indipendenti.**

Es.:

{SoggettoFisico, !SoggettoGiuridico, !Localita}

Il "matching" avviene su un annotation type ***SoggettoFisico*** ma **solo se** a partire dallo StartOffset dell'annotazione non c'è allo stesso posto un annotation type ***SoggettoGiuridico*** **e (AND)** un annotation type ***Localita.***

Developing Language Processing Components with GATE Version 8 (a User Guide)



8.1.11 Negation

… GATE – *Rule based* …

Sulla *NEGATION*

Non so voi ma, ad oggi, non ho ancora trovato un collega che non applicasse istintivamente la logica del negationGrouping = false (che però non è il default)

Attenzione che

questa opzione di controllo è definita a livello di grammatica (file .JAPE) e vale per tutte le regole della grammatica jape !!!

Phase: prove
Input: Token LookupGlossary ….
Options: control = appelt negationGrouping = false

… GATE – *Rule based* …

Sulla *NEGATION*

Attenzione quindi all'inizio, ma non solo 😉 ***, all'uso delle negazioni:***

✓Oppure si può ricorrere a forme assertive (senza la negazione a sinistra dell'annotation type)

✓Oppure si può fare ampio uso degli operatori di contesto

✓Quando le si usano, e sono certo utili/necessarie in diverse situazioni, occorre aver ben chiara la loro logica

✓Oppure si possono creare regole di esclusione dell'annotazione

```
Rule: EsclusioneAnnotazionePersona
Priority: 5
// we want to match 'Jones, F.W.'
(
< LHS Jape constraints >
)
:persona -->
```

oppure

```
:persona
-->
{ }
```

… GATE – *Rule based* …

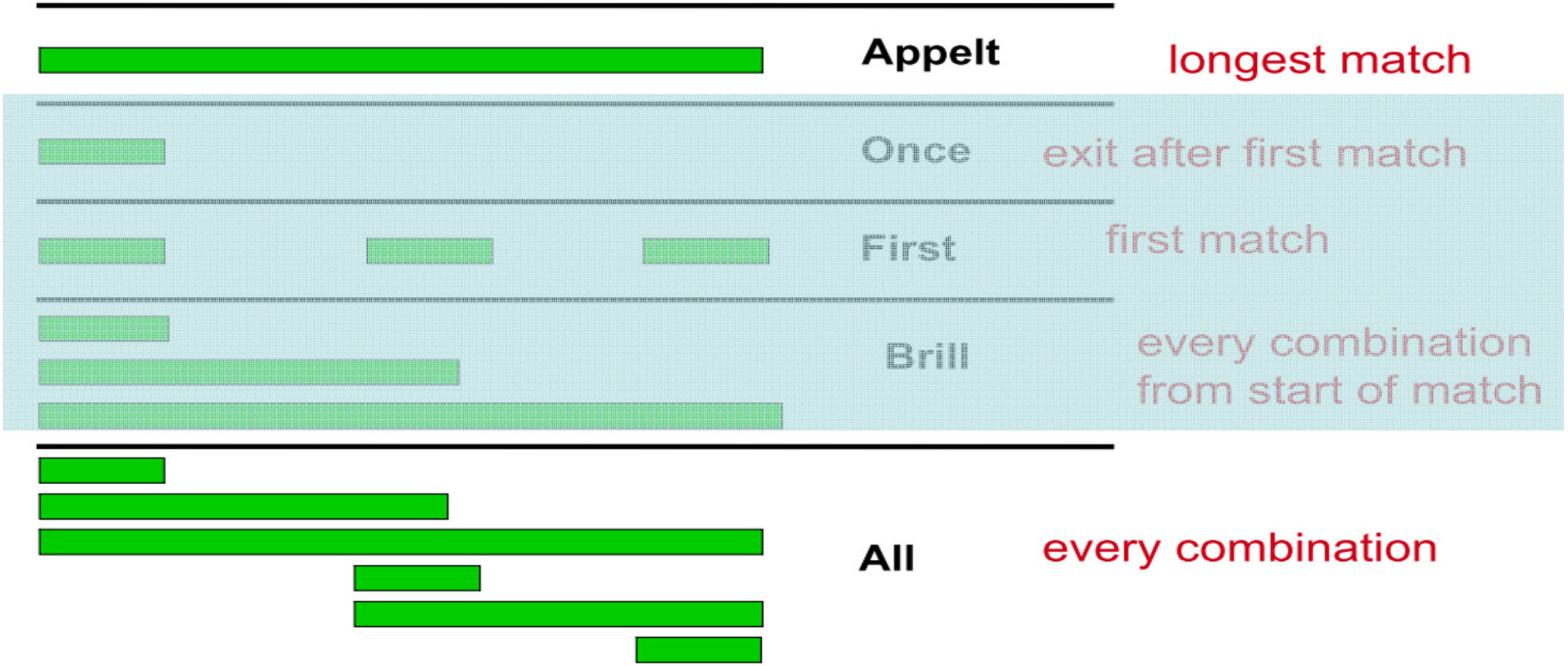
Phase: prove
Input: Token LookupGlossary ….

Options: control = ***appelt*** *negationGrouping = false*

…. Definizioni macro
……. Definizioni regole

{LookupGlossary.majorType == nome}+

Mario **Gino** **Franco**



… GATE – *Rule based* …

Una regola di una grammatica JAPE

```
Rule: SFT3
Priority: 90

//Ing. Davide Maria Bianchi
//Sost. Proc. Francesco maria Borletti Buitoni
//Avvocato simone Raccagni
//Avv sig.ra Maria camilla Verdi
(
    (SEQUENZA_TITOLI):titoli
    (
        (
            ({LookupGlossary.majorType == nome, Token notWithin {LookupGlossary.majorType == sfformedubbie}})[1,2]
            |
            (INIZIALI)
        ):nome
        (
            (
                ({Token.kind == word, Token.length > 1, Token.orth != lowercase, !LookupGlossary.majorType == titstpersona,
                  Token notWithin {LookupGlossary.majorType == sfformedubbie}, Token notWithin {LookupGlossary.majorType == organizzazione}})
                (AA)?
            )[1,3]
        ):cognome
    ):sft3
)
-->
:sft3.SoggettoFisico = {rule = "SFT3", nome = :nome@string, cognome = :cognome@string, titoli = :titoli@string}
```

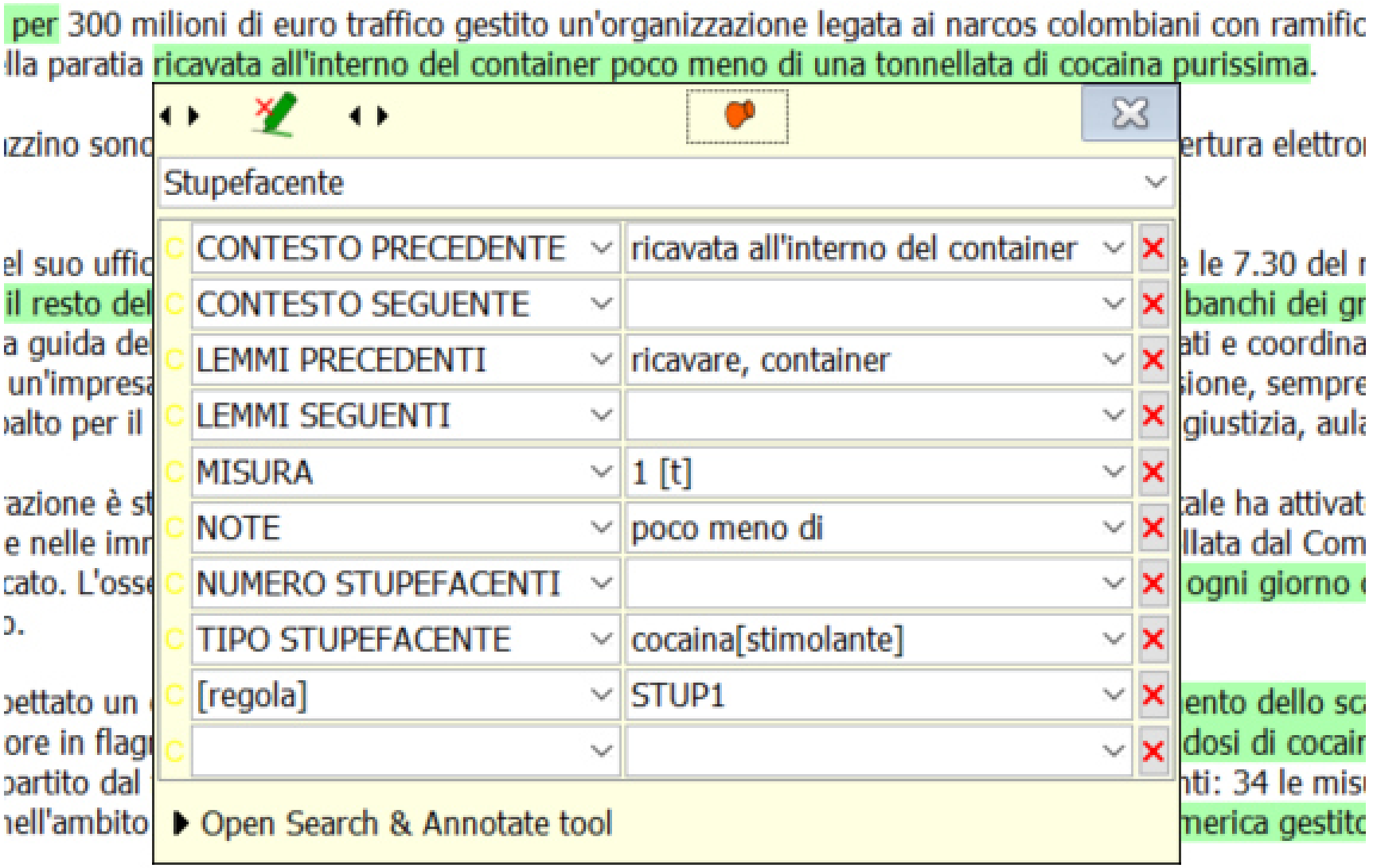
… GATE – *Rule based* …

Una regola di una grammatica JAPE

```
Rule: SGPROP3
Priority: 80

//Omnia spa, costituita il 01-01-2015

(
    (
     ({SoggettoGiuridico}):sg
     (TIPORAGSOC)?
     (PUNTEGGIATURA)?
     (PREFCOST)?
     (ART)?
     ({DataCompleta}):dataCostituzione
     (SEDELEGALE_o_SEDEOPERATIVA_PRIMO)     //opzionale //
     (PUNTEGGIATURA)?
     (SEDELEGALE_o_SEDEOPERATIVA_SECONDO)   //opzionale //
     (REA)?
    )
:matchLegOpe)--> RHSofRuleLegOpe
```

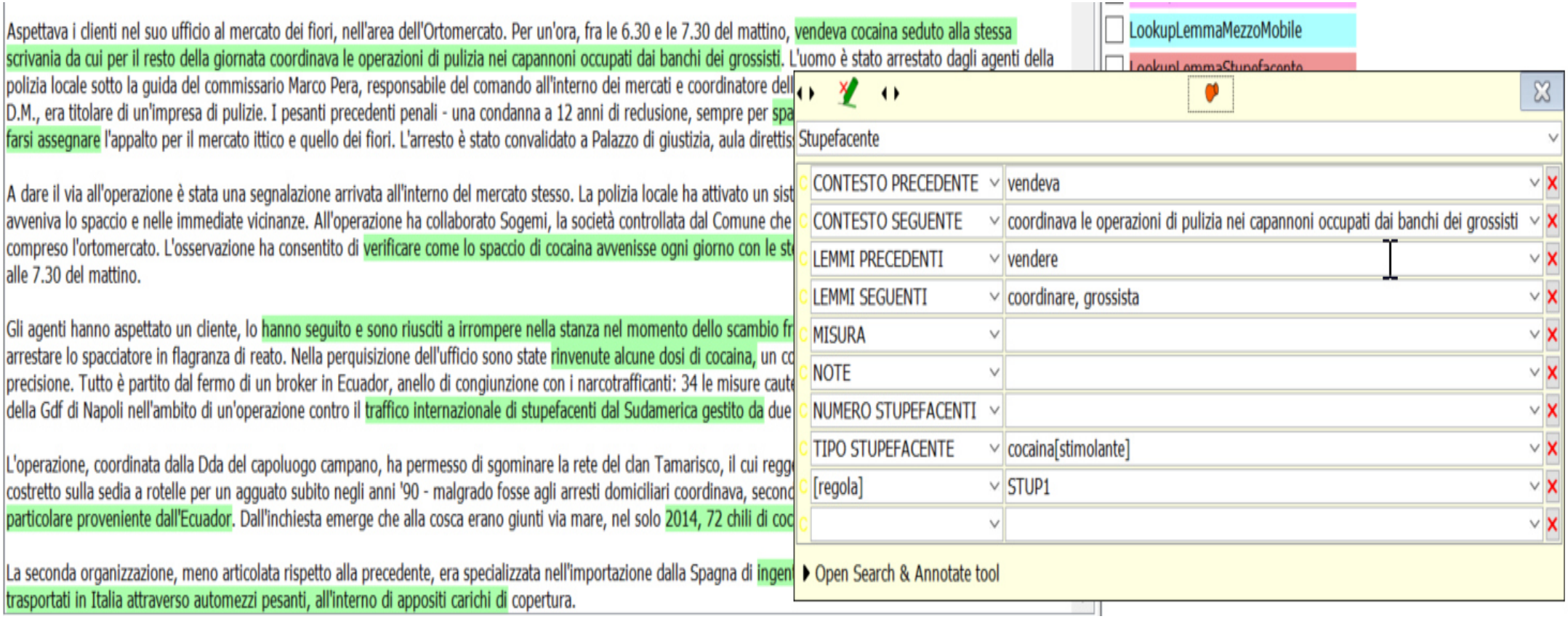
… GATE – *Rule based* …

Altri esempi: stupefacenti



… GATE – *Rule based* …

Altri esempi: stupefacenti



… GATE – *Rule based* …

Altri esempi: stupefacenti

)0; trovate nel mezzo della piazza circa più di 50 pasticche di ecstasy.

ito un'organizzazione legata ai nar
ntainer poco meno di una tonnella

dieci auto gia' approntate con dopp

ll'area dell'Ortomercato. Per un'ora
operazioni di pulizia nei capannon
, responsabile del comando all'inte
edenti penali - una condanna a 12
i fiori. L'arresto è stato convalidat

vata all'interno del mercato stesso.
razione ha collaborato Sogemi, la
verificare come lo spaccio di coca

e sono riusciti a irrompere nella st
uisizione dell'ufficio sono state rinvenute alcune dosi di cocaina, un coltello serramanico e un bilanci

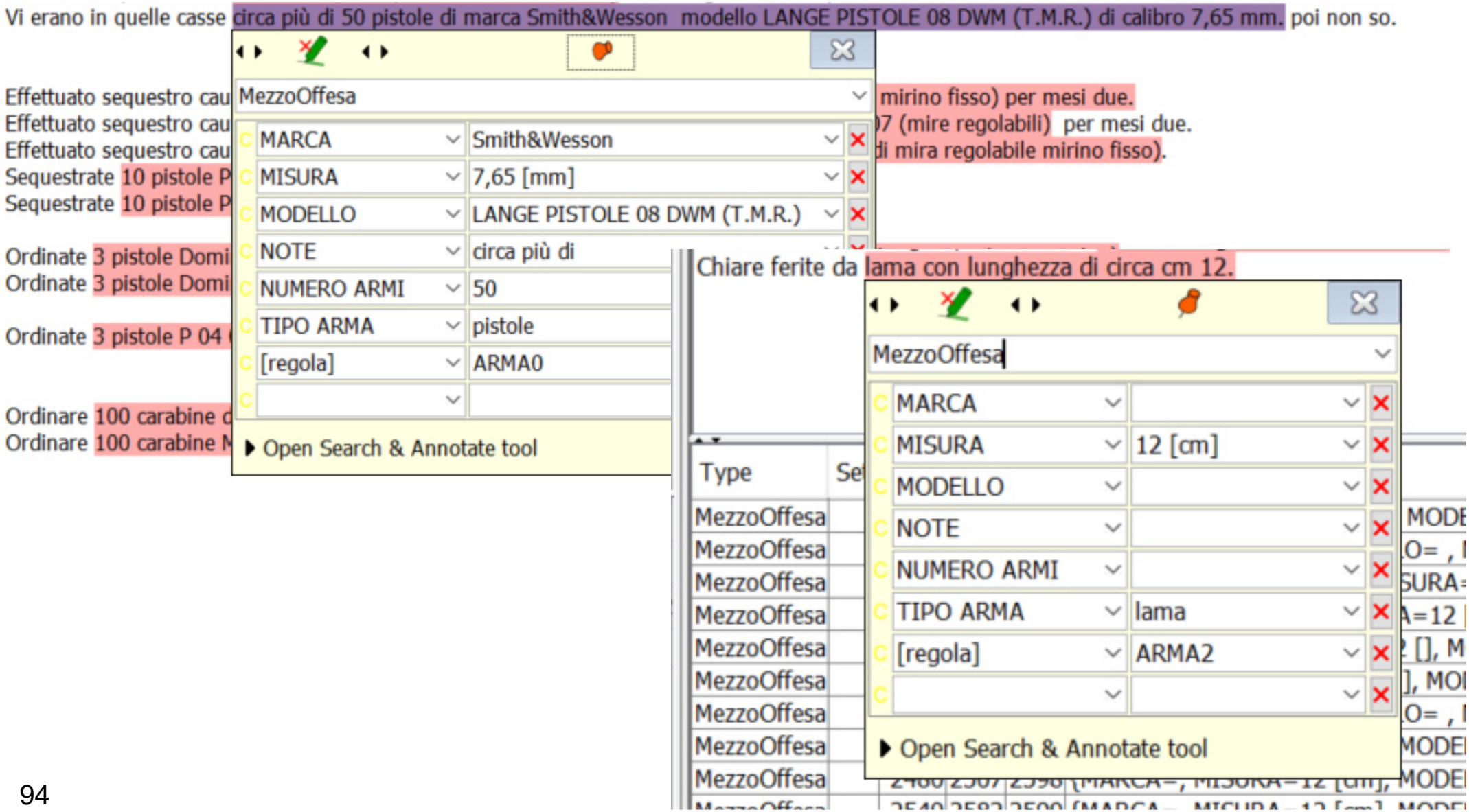
Stupefacente

CONTESTO PRECEDENTE	
CONTESTO SEGUENTE	
LEMMI PRECEDENTI	
LEMMI SEGUENTI	
MISURA	50 [pasticca]
NOTE	circa più di
NUMERO STUPEFACENTI	
TIPO STUPEFACENTE	ecstasy[allucinogeno]
[regola]	STUP1

Open Search & Annotate tool

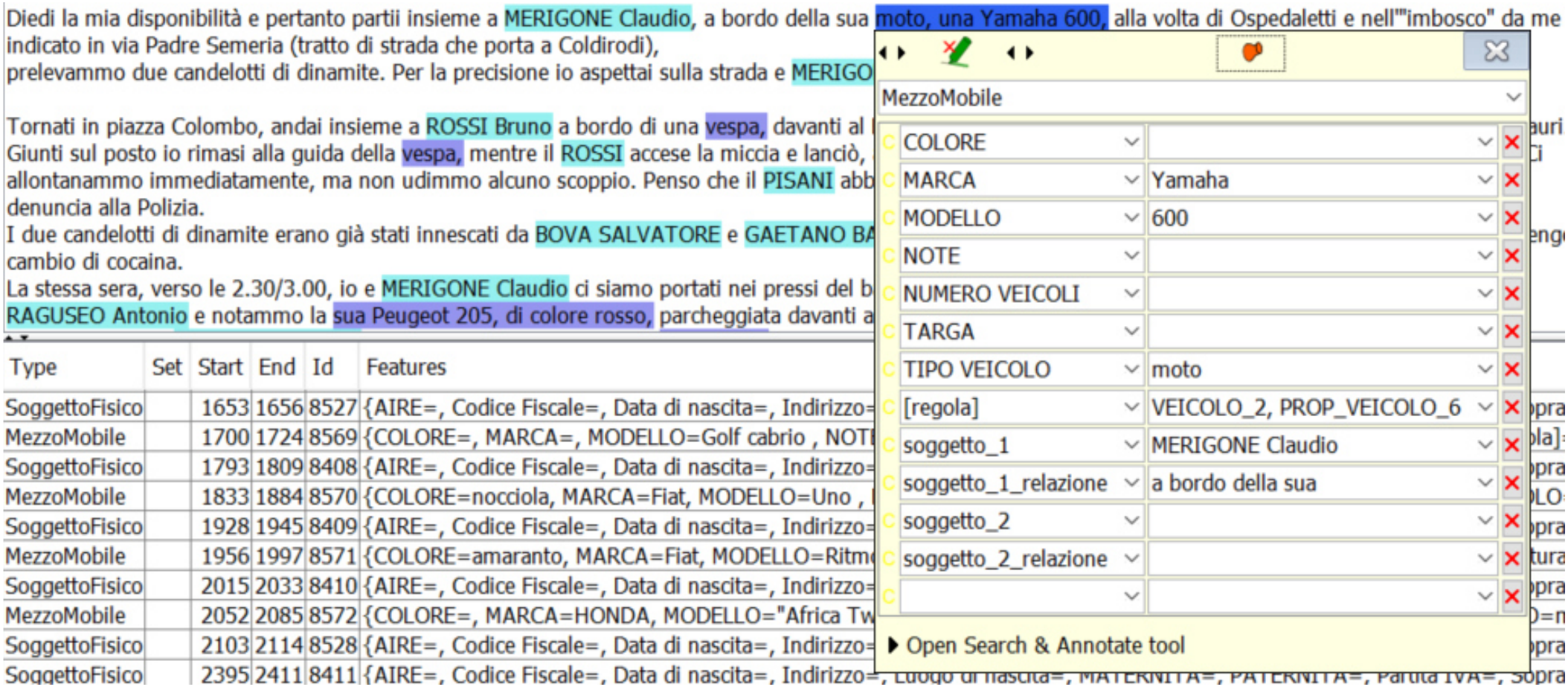
… GATE – *Rule based* …

Altri esempi: armi

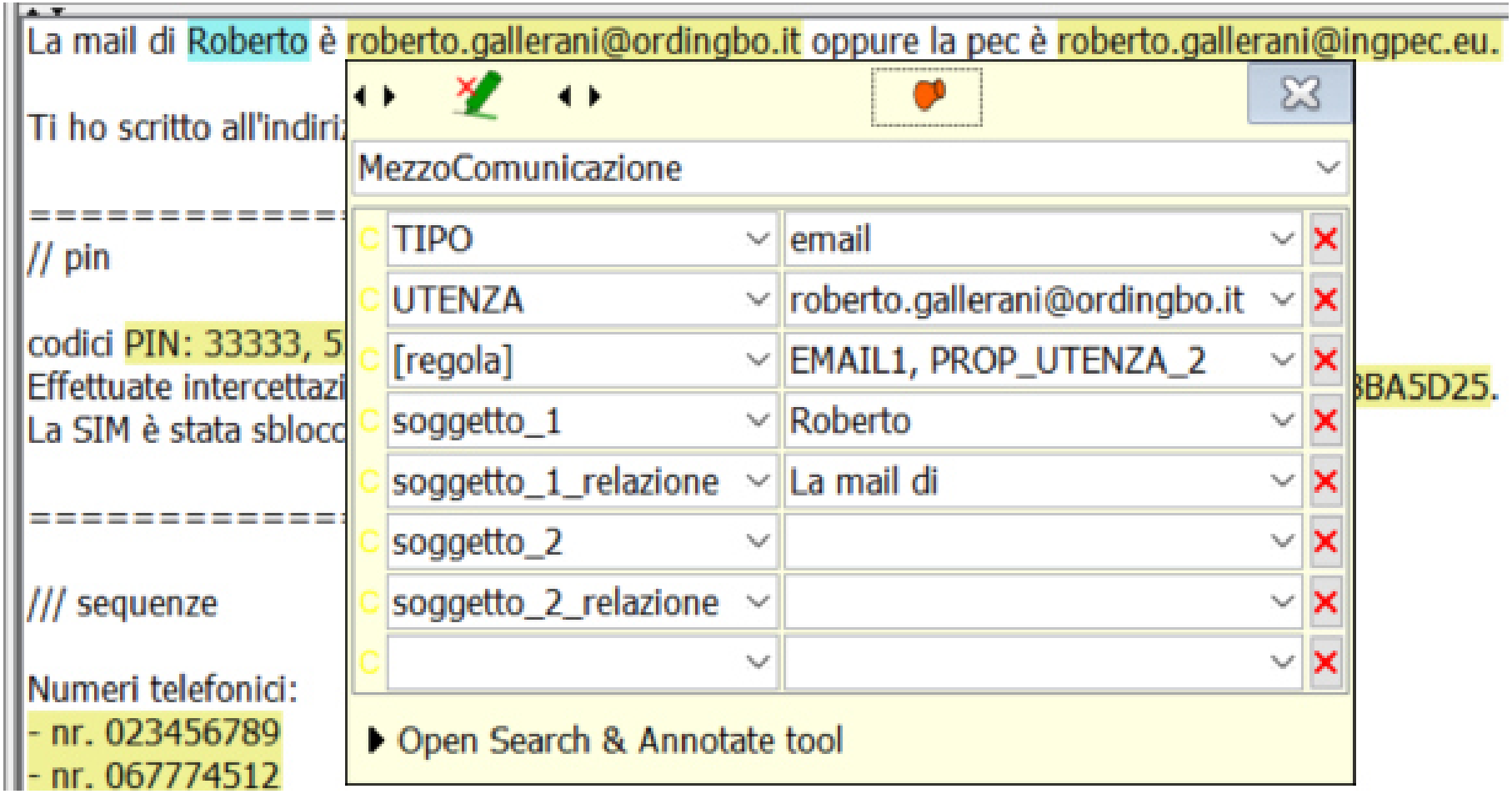


… GATE – *Rule based* …

Altri esempi: mezzi mobili

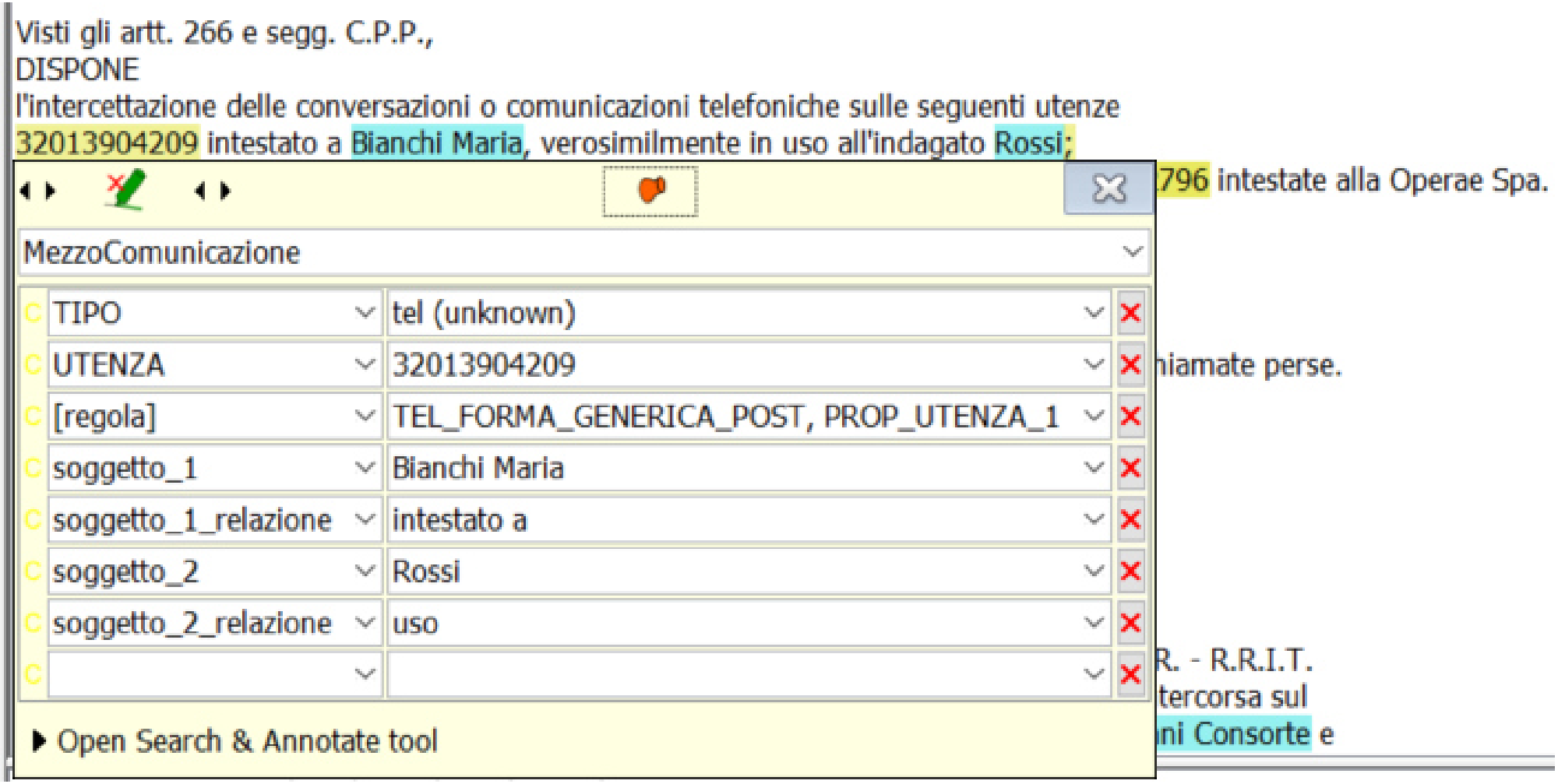


Altri esempi: mezzi comunicazione



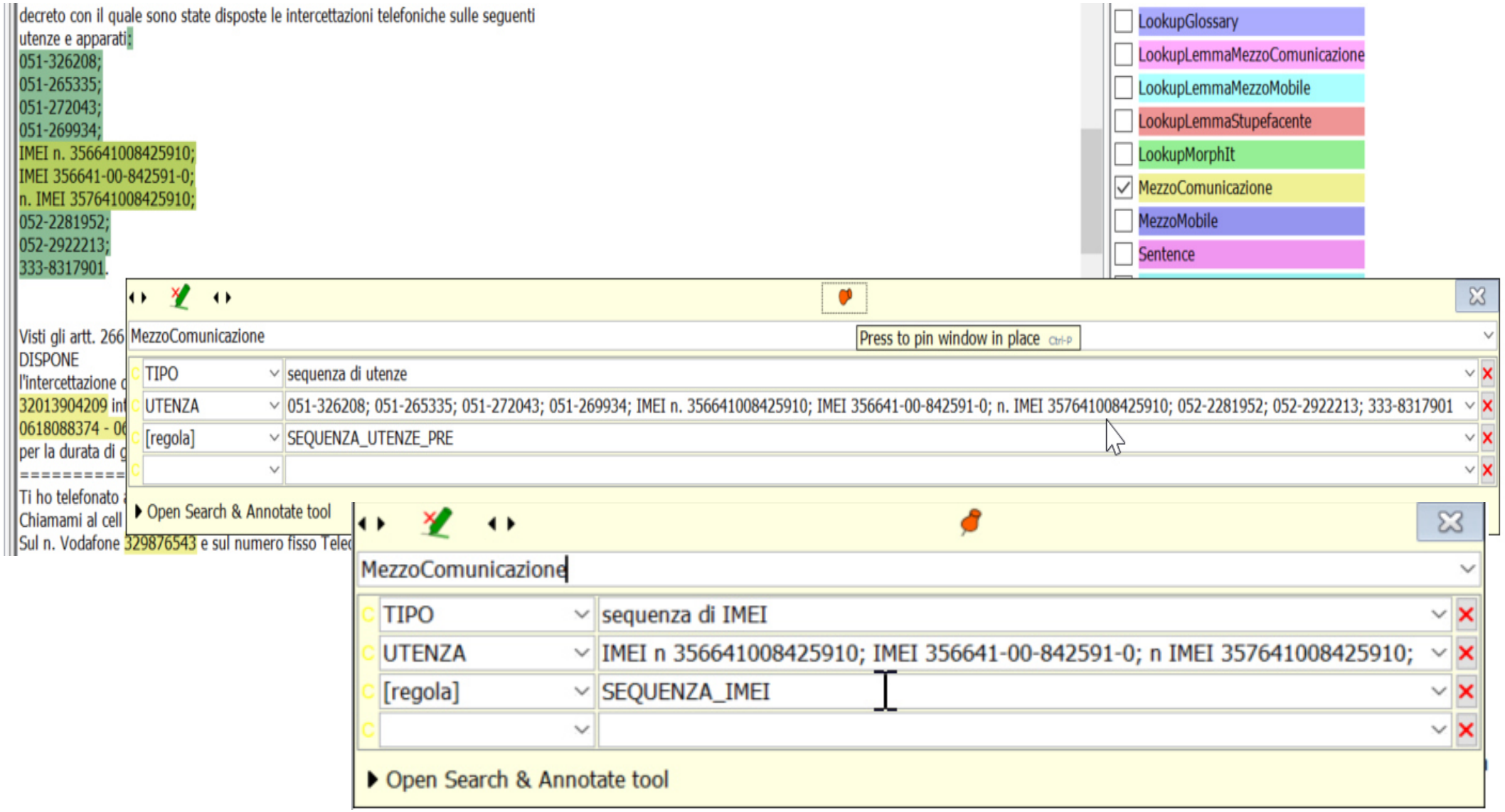
… GATE – *Rule based* …

Altri esempi: mezzi comunicazione



… GATE – *Rule based* …

Altri esempi: mezzi comunicazione



GATE – Rule based: alcuni altri aspetti …

La progettazione e la realizzazione di un sistema di grammatiche JAPE deve prevedere un'organizzazione modulare e condizionale per l'impiego di glossari, grammatiche e pipeline che sappia gestire:

✓ differenti livelli di profondità di IE;

✓ riuso di grammatiche elementari;

✓ utilizzo in diverse modalità dei glossari esistenti (in funzione del livello di profondità o della tipologia dei documenti).

Una pipeline Pi è una sequenza di PR, {PR1, PR2, …, PRn} applicata a un set (corpus) di documenti, al limite anche uno solo.

Un Pri generico di una pipeline può essere sempre eseguito in modo incondizionato oppure in modo condizionato al verificarsi di una specifica condizione.

… GATE – Rule based: alcuni altri aspetti

Un **progetto applicativo** → presenza di diverse pipeline per diverse esigenze di IE. Il progetto deve garantire che esse possano **condividere**, **attraverso** i relativi **PR**, **parti in comune**: glossari, grammatiche jape, script groovy ecc. secondo logiche modulari.

Inoltre considerare:

✓ la "condizionalità" di un PR di una pipeline è solo "binaria": ESEGUI/NON ESEGUI;

✓ le regole di esecuzione (annotazione) delle regole di una grammatica jape (Options: control =)

Per una generica pipeline Pi i PR che la compongono vengono pertanto definiti, laddove necessario, attraverso la forma condizionale.

Per un generico tipo di PR la "condizionalità" consente di impostare una pipeline basata su sequenze multiple di PR, dello stesso tipo e non, eseguibili nella corretta sequenza attraverso l'impostazione di parametri di esecuzione, valorizzabili manualmente in GATE Developer (durante la fase di realizzazione e test) oppure da un'applicazione JAVA attraverso le API del framework GATE Embedded.

Pratica / Esempi interattivi

PR “NE JAPE TRANSDUCER”

- ✓Esecuzione del PR.
- ✓Navigazione dei risultati sul documento.
- ✓La navigazione/gestione delle grammatiche.
- ✓ La configurazione e la forma dei file di grammatiche.

Come misurare i risultati

Occorrenze: numero di occorrenze appartenenti all'entità presenti nel documento

Occorrenze distinte: numero di occorrenze distinte appartenenti all'entità presenti nel documento

Riconoscimenti corretti (*C*): occorrenze individuate dal software perfettamente

Riconoscimenti parzialmente corretti (*P*): occorrenze individuate dal software con alcune imprecisioni, come ad esempio una annotazione che eccede di alcuni caratteri rispetto a quella corretta

Falsi Positivi (*FP*): riconoscimenti fatti dal software che non corrispondono a reali occorrenze dell'entità

Falsi Negativi (*FN*): occorrenze presenti nel documento ma non individuate dal software

Come misurare i risultati

- **Precision** (*Pr*): indicatore che fa trasparire quanto ciò che il software identifica sia effettivamente corretto. Risulta tanto più elevato quanto meno sono presenti falsi positivi.

$$Precision = \frac{C + \alpha \cdot P}{C + FP + P}$$

dove $\alpha \epsilon [0,1]$ indica quanto peso si vuole attribuire ai riconoscimenti parzialmente corretti. Nel test in questione è stato fissato $\alpha = 0.5$.

- **Recall** (*Rec*): indicatore che fa trasparire se il software trova effettivamente tutto ciò che nel documento va evidenziato. In minor misura sono presenti falsi negativi, più il valore dell'indicatore è elevato.

$$Recall = \frac{C + \alpha \cdot P}{C + FN + P}$$

dove $\alpha \epsilon [0,1]$ indica quanto peso si vuole attribuire ai riconoscimenti parzialmente corretti. Nel test in questione è stato fissato $\alpha = 0.5$.

Come misurare i risultati

- **F-measure**: indicatore di sintesi calcolato con media pesata tra i due precedenti. In un sistema di Natural Language Processing è piuttosto semplice forzare uno dei due precedenti indicatori per raggiungere percentuali prossime al 100%, ma è molto difficile farlo per entrambi contemporaneamente, cioè ottenere anche un F-measure prossimo al 100%.

$$F - measure = \frac{(\beta^2 + 1) \cdot Pr \cdot Rec}{(\beta^2 \cdot Pr) + Rec}$$

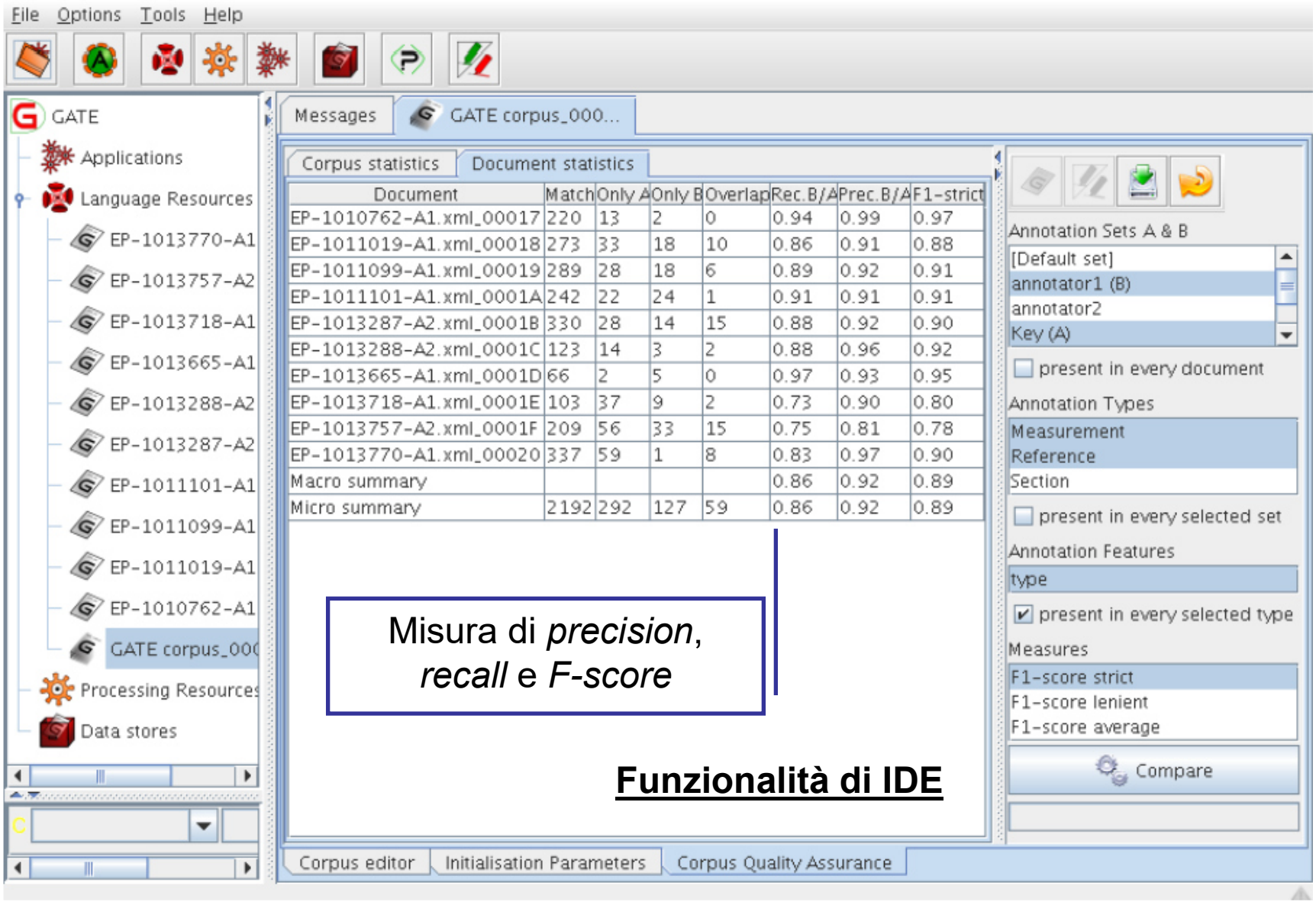
β = *Recall / Precision*	→ indica quanto pesare tra loro *Recall* e *Precision*.
Se β vale 1	→ hanno uguale importanza.
Se β vale 0.5	→ *Precision* pesa il doppio di *Recall.*
Se β vale 2	→ *Recall* pesa il doppio di *Precision.*

Come misurare i risultati

Misure		Documento XXXXXXXXXXXX							
GG/MM/AAAA		Presenti	Riconosc. corretti	Riconosc. Parz. corretti	Falsi Positivi	Falsi Negativi	Precision	Recall	F-measure
Soggetto Fisico	Occorrenze	70	61	1	6	8	90,44%	87,86%	**89,13%**
Località	Occorrenze	24	21	0	1	3	95,45%	87,50%	**91,30%**

… GATE – *Rule based* …

Come misurare i risultati (tra due annotation set su stesso docum.)



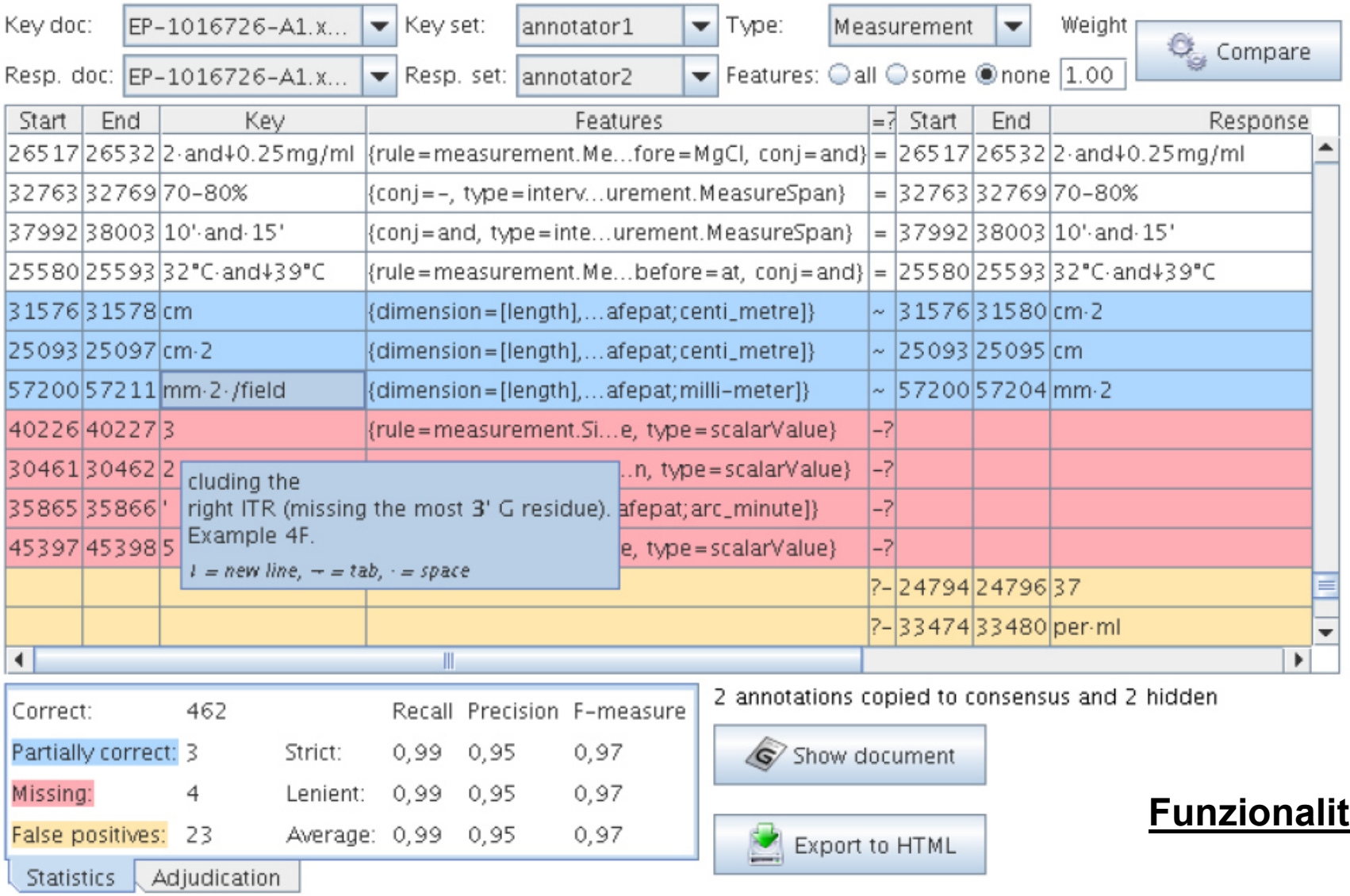
uno etichettato come **A**, (***gold standard***)

e l'altro

B di annotazioni da confrontare *(generate da una pipeline)*

… GATE – *Rule based* …

Come misurare i risultati (diff di due documenti)



Funzionalità di IDE

… GATE – *Rule based* …

Come misurare i risultati (Corpus Benchmark Tool)

```
main directory (can have any name)
|
|_"clean" (directory containing unannotated documents
|
|_"marked" (directory containing annotated documents in XML form)
|
|_"processed" (directory containing the datastore which is generated
              when you 'store corpus for future evaluation')
```

```
threshold=0.7
annotSetName=Key
outputSetName=ANNIE
annotTypes=Person;Organization;Location;Date;Address;Money
annotFeatures=type;gender
```

Soglia *precision/recall*

Annotation set delle annotazioni utente

Annotation set delle annotazioni generate

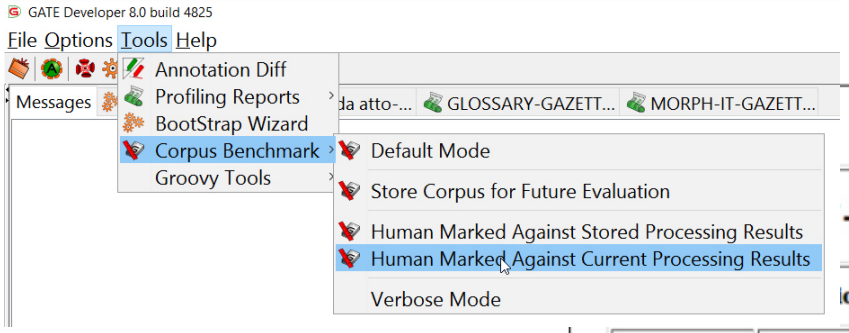
Annotation type interessati

Feature considerate

Funzionalità di IDE

… GATE – *Rule based* …

Come misurare i risultati (Corpus Benchmark Tool)



[2202,2210] **England**: *[12,19]* **Commerce**: *[2618,2626]* **England**: *[2260,2267]* **Bank**: *[344,348]*

-jul-2001.xml

		ion	Recall	Annotations
Mention	0.851063829787234		0.6896551724137931	MISSING ANNOTATIONS in the automatic texts: **1997**: *[652,656]* **the Airline Group**: *[1669,1686]* **London Underground**: *[388,406]* **The Airline Group**: *[1266,1283]* **The Airline Group**: *[2412,2429]* **Swanwick**: *[2634,2642]* **2000**: *[2018,2022]* **The Airline Group**: *[938,955]* **The Airline Group**: *[1751,1768]* **Swanwick**: *[2376,2384]* **Monarch Airlines**: *[1052,1068]* **Virgin Atlantic**: *[45,60]* **Britannia Airways**: *[1030,1047]* **Labour**: *[634,640]* **London Area and Terminal Control Centre**: *[2029,2068]* SPURIOUS ANNOTATIONS in the automatic texts: **peak days**: *[2194,2203]* **PPP**: *[376,379]* **The Airline Group**: *[938,955]* **London**: *[2029,2035]* PARTIALLY CORRECT ANNOTATIONS in the automatic texts: **last year Nats handled more than 2m air traffic movements with volumes growing by 5 per cent in 2000**: *[1922,2022]* **Abbey**: *[1367,1372]* **March**: *[1743,1748]* **March**: *[1486,1491]* **January**: *[2402,2409]* **next 10**: *[2520,2527]*

ft-bmi-09-may-2001.xml

Annotation Type	Precision	Recall	Annotations
Mention	0.9791666666666666	0.94	

Statistics

Annotation Type	Correct	Partially Correct	Missing	Spurious	Precision	Recall	F-Measure
Mention	434	16	34	8	0.9650655021834061	0.9132231404958677	0.9384288747346073

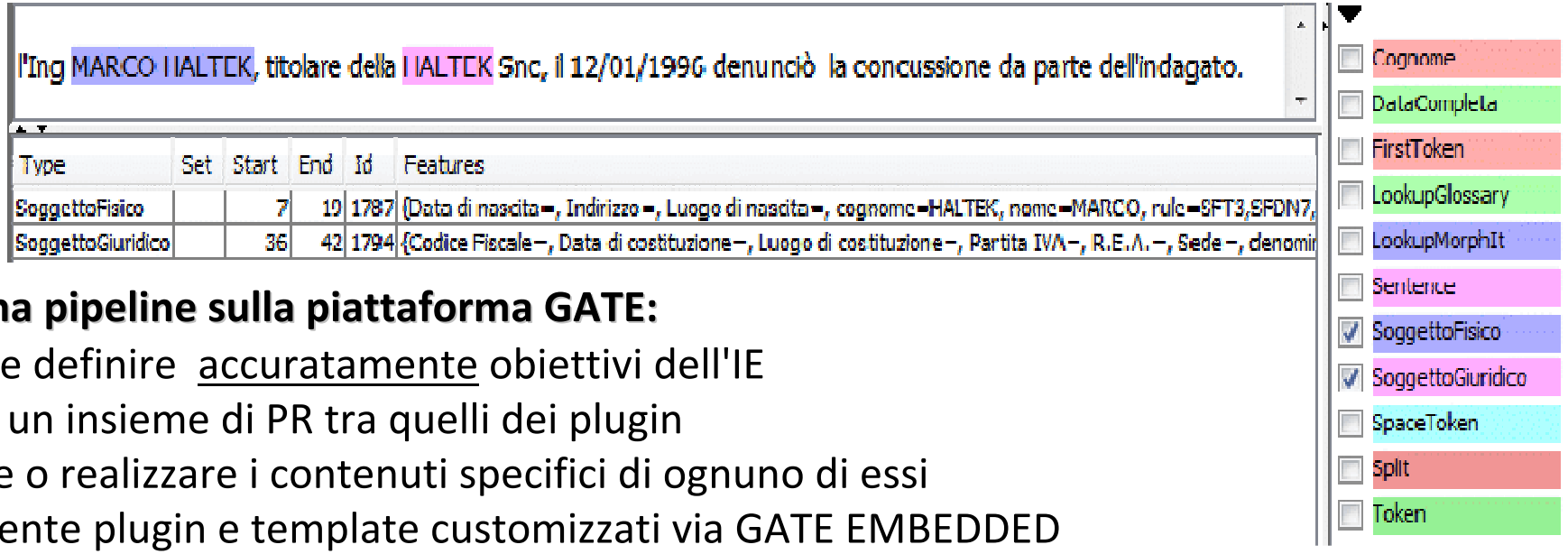
Overall average precision: 0.959158791298373
Overall average recall: 0.9108515466531439
Overall average fMeasure : 0.9332844120837758
Finished!

… GATE – *Rule based*

Considerazioni generali nell'approccio "ruled based"

✓ Organizzazione di glossari e grammatiche di regole è correlata alla lingua, ai formati di codifica, al "dominio".

✓ Uso di glossari e grammatiche di regole non deve far pensare a GATE come a un puro "motore di pattern matching testuale". Il valore del risultato è determinato dall'azione delle regole.



Sviluppare una pipeline sulla piattaforma GATE:

✓ Analizzare e definire accuratamente obiettivi dell'IE

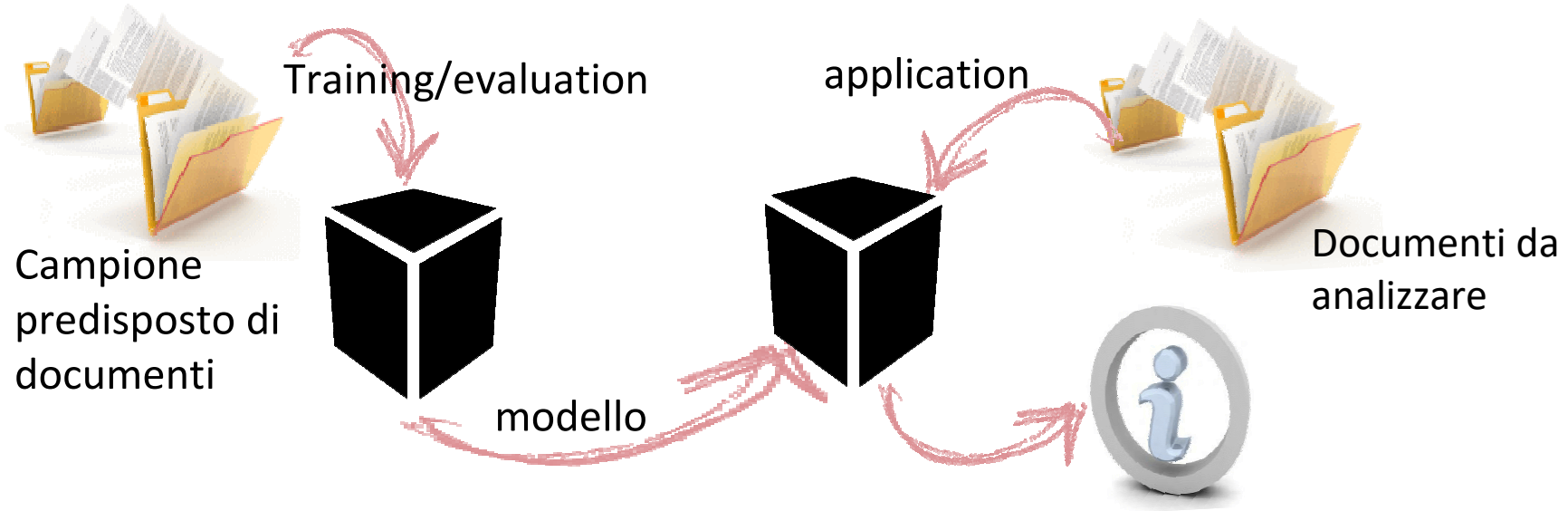
✓ Comporre un insieme di PR tra quelli dei plugin

✓ Revisionare o realizzare i contenuti specifici di ognuno di essi

✓ Eventualmente plugin e template customizzati via GATE EMBEDDED

✓ Tutte le annotazioni generate possono essere registrate su DB o file esterni per essere poi rielaborate, mediante la creazione di PR Groovy o di classi java basate sul framework GATE EMBEDDED

GATE – *Machine learning (SVM) …*

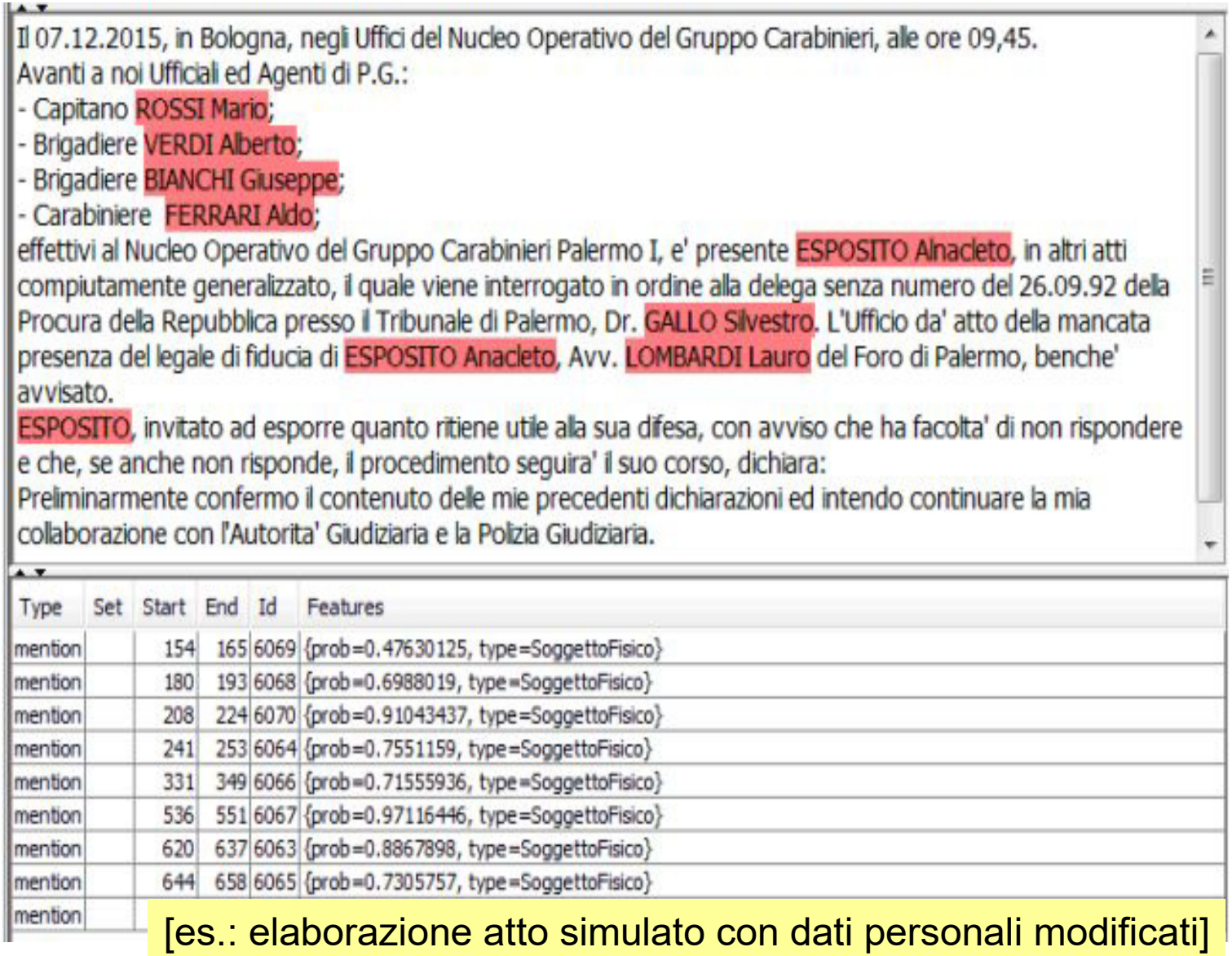


Passi del "machine learning" :

- ✓ annotazione manuale di un campione documentale
- ✓ configurazione dei criteri di analisi del campione (caratteristiche e proprietà degli elementi di interesse e di quelli in prossimità)
- ✓ "training" del motore sul campione ed "evaluation" dei risultati
- ✓ Il motore produrrà una serie di artefatti (matrici di rappresentazione statistica delle condizioni di annotazione)
- ✓ Il modello verrà applicato dal motore a documenti diversi dal campione ("application")

… GATE – *Machine learning (SVM)* …

“Application”
Il modello statistico viene applicato a un insieme di documenti diverso da quello campione, determinando come risultato un’annotazione per ogni sequenza di token rispondente alla distribuzione con l’indicazione della probabilità di buon riconoscimento.



Il 07.12.2015, in Bologna, negli Uffici del Nucleo Operativo del Gruppo Carabinieri, alle ore 09,45.
Avanti a noi Ufficiali ed Agenti di P.G.:
- Capitano ROSSI Mario;
- Brigadiere VERDI Alberto;
- Brigadiere BIANCHI Giuseppe;
- Carabiniere FERRARI Aldo;

effettivi al Nucleo Operativo del Gruppo Carabinieri Palermo I, e' presente ESPOSITO Alnacleto, in altri atti compiutamente generalizzato, il quale viene interrogato in ordine alla delega senza numero del 26.09.92 della Procura della Repubblica presso il Tribunale di Palermo, Dr. GALLO Silvestro. L'Ufficio da' atto della mancata presenza del legale di fiducia di ESPOSITO Anacleto, Avv. LOMBARDI Lauro del Foro di Palermo, benche' avvisato.
ESPOSITO, invitato ad esporre quanto ritiene utile alla sua difesa, con avviso che ha facolta' di non rispondere e che, se anche non risponde, il procedimento seguira' il suo corso, dichiara:
Preliminarmente confermo il contenuto delle mie precedenti dichiarazioni ed intendo continuare la mia collaborazione con l'Autorita' Giudiziaria e la Polizia Giudiziaria.

Type	Set	Start	End	Id	Features
mention		154	165	6069	{prob=0.47630125, type=SoggettoFisico}
mention		180	193	6068	{prob=0.6988019, type=SoggettoFisico}
mention		208	224	6070	{prob=0.91043437, type=SoggettoFisico}
mention		241	253	6064	{prob=0.7551159, type=SoggettoFisico}
mention		331	349	6066	{prob=0.71555936, type=SoggettoFisico}
mention		536	551	6067	{prob=0.97116446, type=SoggettoFisico}
mention		620	637	6063	{prob=0.8867898, type=SoggettoFisico}
mention		644	658	6065	{prob=0.7305757, type=SoggettoFisico}
mention					

[es.: elaborazione atto simulato con dati personali modificati]

… GATE – *Machine learning (SVM)* …

Annotazione manuale di un campione documentale

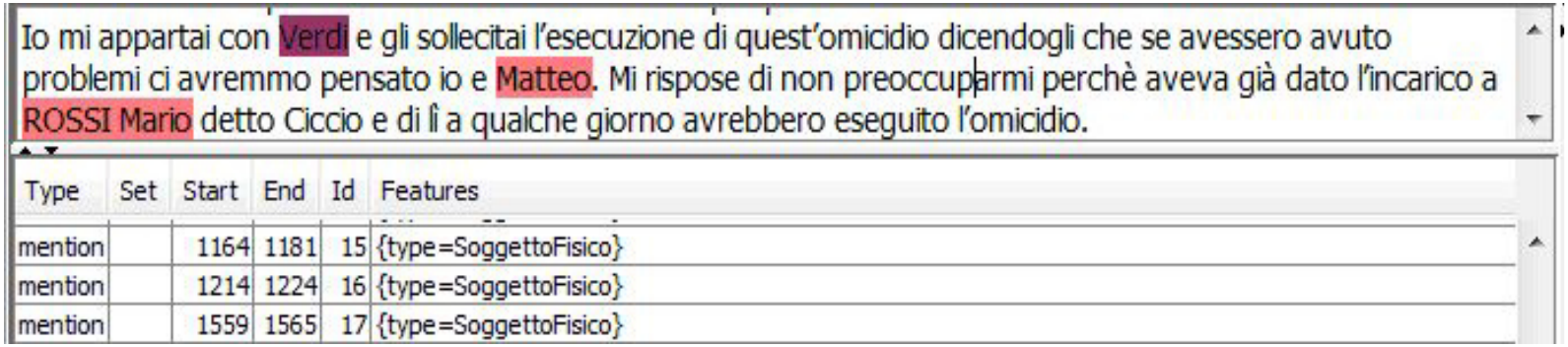
Ciascuno dei documenti del campione dovrà essere opportunamente annotato indicando le istanze di interesse, nel caso in esempio indicate attraverso il nome di entità “SoggettoFisico”.

Configurazione dei criteri di analisi del campione

✓ algoritmo impiegato (nel nostro caso SVM)

✓ metodo e opzioni tecniche di esecuzione dell’algoritmo

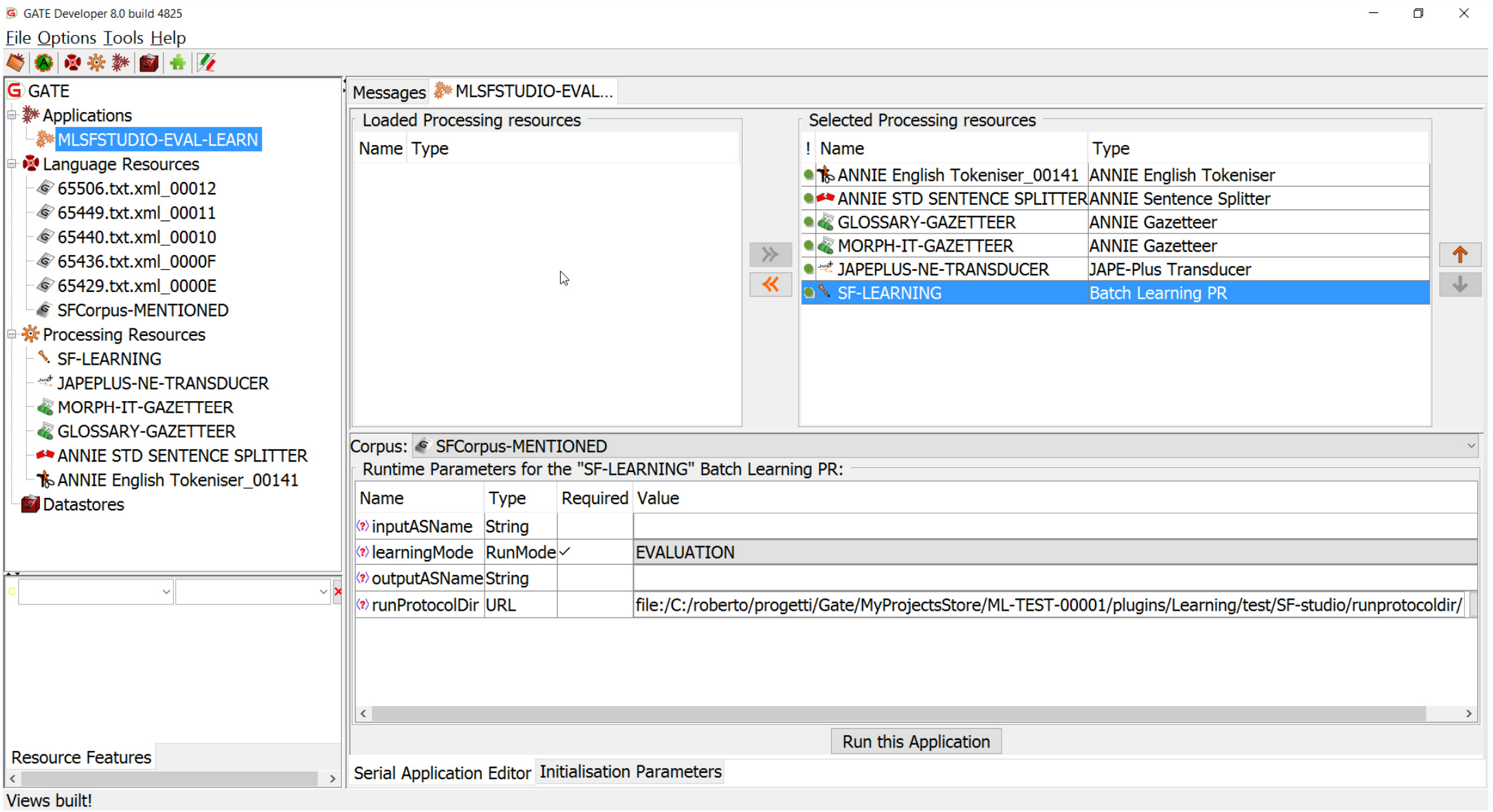
ad es. l’indicazione degli attributi che entrano in gioco nel processo di elaborazione statistica



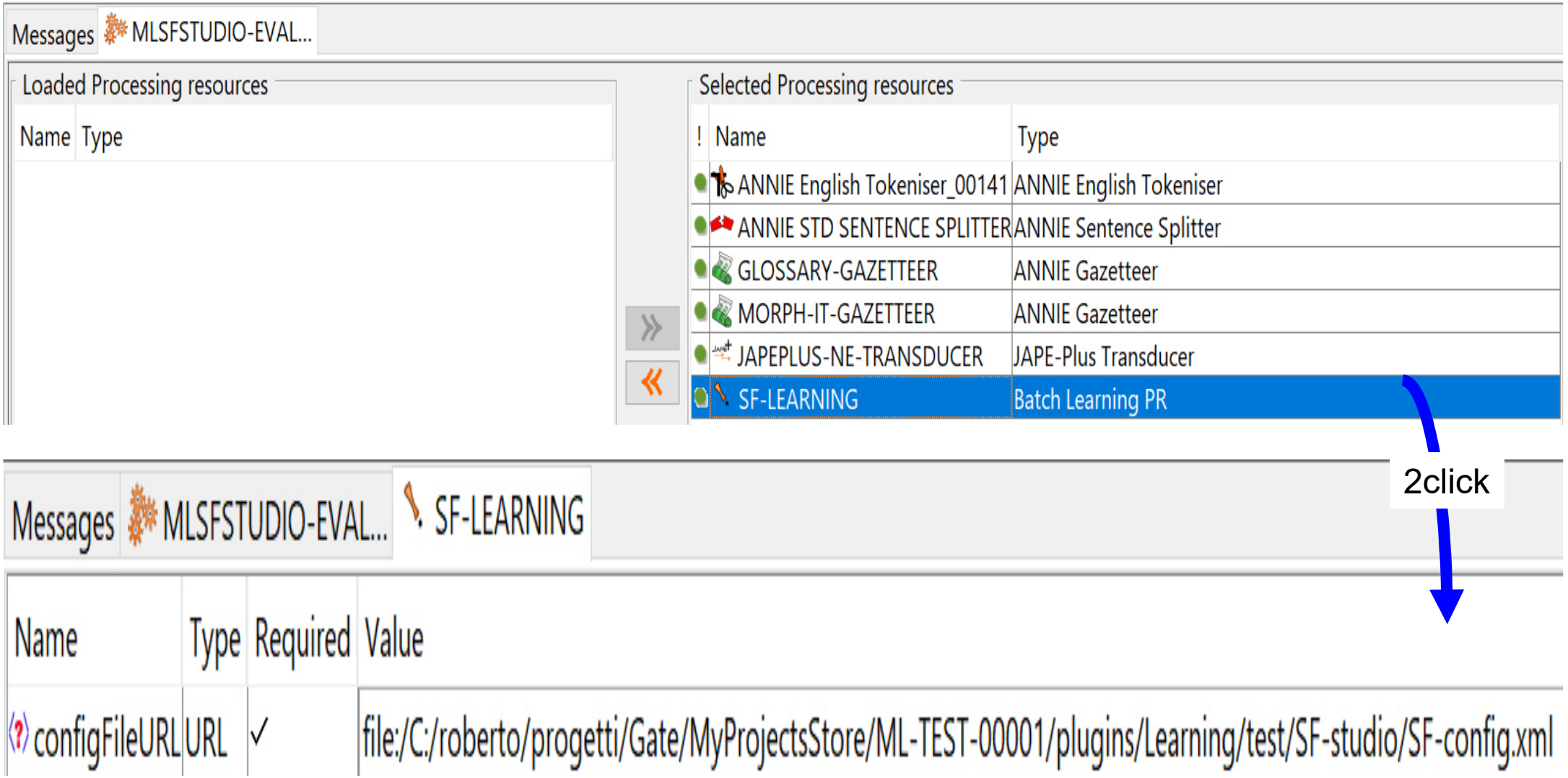
Tipo annotazione	Caratteristiche espresso dal valore della feature	Intervallo di prossimità
Token	Orth: l’essere maiuscolo l’iniziale, interamente maiuscolo, tutto minuscolo	[-1,1]
Token	Kind: l’essere un numero, una parola o un simbolo di punteggiatura	[-2,2]
Token	String: il valore della stringa del token	[-3,3]
LookupGlossary	majorType: tipologia del glossario	[-3,3]
LookupGlossary	minorType: sottotipologia del glossario	[-3,3]
LookupMorphIt	Lemma: lemma della forma relativa alla stringa del token	[-3,3]
LookupMorphIt	Morfologia: morfologia della forma relative alla stringa del token	[-3,3]

… GATE – *Machine learning (SVM)* …

"Evaluation"



… GATE – *Machine learning (SVM)* …



… GATE – *Machine learning (SVM)* …

Elementi di configurazione che pilotano il PR di ML (SVM)

```
<?xml version="1.0"?>
<ML-CONFIG>
  <!-- Verbosity: -1 none; 0 minimum; 1 normal; 2 debug -->
  <VERBOSITY level="2"/>
  <!-- Surround: true consigliato per NER learning in quanto si utilizza -->
  <!--           nell'apprendimento una composizione di "pezzi" ciascuno con il suo start/end -->
  <SURROUND value="true"/>

  <!-- Filtering:  ratio=0.0 e dis=far sono i valori default che di fatto non eseguono il "filtering" -->
  <FILTERING ratio="0.0" dis="far"/>

  <!-- le soglie sottostanti sono al momento poste ai valori di default -->
  <PARAMETER name="thresholdProbabilityEntity" value="0.2"/>
  <PARAMETER name="thresholdProbabilityBoundary" value="0.4"/>

  <!-- ================================================================= -->

  <multiClassification2Binary method="one-vs-others"/>

  <EVALUATION method="kfold"
        runs="4"/>

  <ENGINE nickname="SVM" implementationName="SVMLibSvmJava"
     options=" -c 0.7 -t 0 -m 2400 -tau 0.4  "/>
```

……

……

… GATE – *Machine learning (SVM)*

Elementi di configurazione che pilotano il PR di ML (SVM)

<DATASET>

```
<INSTANCE-TYPE>Token</INSTANCE-TYPE>

<ATTRIBUTELIST>
  <NAME>TokenOrth</NAME>
  <SEMTYPE>NOMINAL</SEMTYPE>
  <TYPE>Token</TYPE>
  <FEATURE>orth</FEATURE>
  <RANGE from="-1" to="1"/>
</ATTRIBUTELIST>

<ATTRIBUTELIST>
  <NAME>TokenKind</NAME>
  <SEMTYPE>NOMINAL</SEMTYPE>
  <TYPE>Token</TYPE>
  <FEATURE>kind</FEATURE>
  <RANGE from="-2" to="2"/>
</ATTRIBUTELIST>

<ATTRIBUTELIST>
  <NAME>TokenString</NAME>
  <SEMTYPE>NOMINAL</SEMTYPE>
  <TYPE>Token</TYPE>
  <FEATURE>string</FEATURE>
  <RANGE from="-3" to="3"/>
</ATTRIBUTELIST>
```

</DATASET>

```
<ATTRIBUTELIST>
  <NAME>LookupGlossaryMajorType</NAME>
  <SEMTYPE>NOMINAL</SEMTYPE>
  <TYPE>LookupGlossary</TYPE>
  <FEATURE>majorType</FEATURE>
  <RANGE from="-3" to="3"/>
</ATTRIBUTELIST>

<ATTRIBUTELIST>
  <NAME>LookupGlossaryMinorType</NAME>
  <SEMTYPE>NOMINAL</SEMTYPE>
  <TYPE>LookupGlossary</TYPE>
  <FEATURE>minorType</FEATURE>
  <RANGE from="-3" to="3"/>
</ATTRIBUTELIST>
```

```
<ATTRIBUTELIST>
  <NAME>LookupMorphItLemma</NAME>
  <SEMTYPE>NOMINAL</SEMTYPE>
  <TYPE>LookupMorphIt</TYPE>
  <FEATURE>lemma</FEATURE>
  <RANGE from="-3" to="3"/>
</ATTRIBUTELIST>

<ATTRIBUTELIST>
  <NAME>LookupMorphItMorfologia</NAME>
  <SEMTYPE>NOMINAL</SEMTYPE>
  <TYPE>LookupMorphIt</TYPE>
  <FEATURE>morfologia</FEATURE>
  <RANGE from="-3" to="3"/>
</ATTRIBUTELIST>
```

… GATE – *Machine learning (SVM)* …

“Evaluation”

k-fold cross-validation
PR segmenta corpus in k partizioni di uguale dimensione e usa ogni partizione com “test set” e i rimanenti documenti come “training set”. Parametro di configurazione *run* determina il numero di partizioni.

Hold-out validation
PR seleziona casualmente un insieme di documenti per il “test set” e usa I restanti come “training set”. Parametro di configurazione ***ratio*** indica la percentuale sul totale dei documenti del corpus da usare come “training set”

Es.:

<EVALUATION method=“kfold" runs=4/> (è il caso di esempio della slide precedente)

<EVALUATION method=“holdout " ratio=0.66/>

… GATE – *Machine learning (SVM)* …

"Evaluation"

Messages | MLSFSTUDIO-EVAL...

```
Pre-processing the 5 documents...
Learning starts.
For the information about this learning see the log file
C:\roberto\progetti\Gate\MyProjectsStore\ML-TEST-00001\plugins\Learning\test\SF-studio\savedFiles\logFileForNLPLearning.save
The number of threads used is 1
** Evaluation mode started:
Kfold k=4, numDoc=5, len=1.

*** Fold 1
Number of docs for training: 4
1 65436.txt.xml_0000F
2 65440.txt.xml_00010
3 65449.txt.xml_00011
4 65506.txt.xml_00012
Number of docs for application: 1
1 65429.txt.xml_0000E
```

XVAL Fold 0: (correct, partialCorrect, spurious, missing)= (41.0, 1.0, 20.0, 1.0); (**precision, recall, F1**)= (**0.66129035, 0.95348835, 0.78095233**);)

Overall results as:
(correct, partialCorrect, spurious, missing)= (42.25, 0.5, 9.75, 3.25);

(**precision, recall, F1**)= (**0.81793886, 0.9163628, 0.860167**)

GATE Machine Learning (SVM)

- ✓Esecuzione di un progetto di esempio (soggettoFisico).
- ✓Corpus training/evaluation e test.
- ✓PR di Batch Learning (EVALUATION,TRAINING,APPLICATION).
- ✓ La configurazione e il tuning.

GATE – Rule based/*Machine learning*

Considerazioni generali sui due approcci:

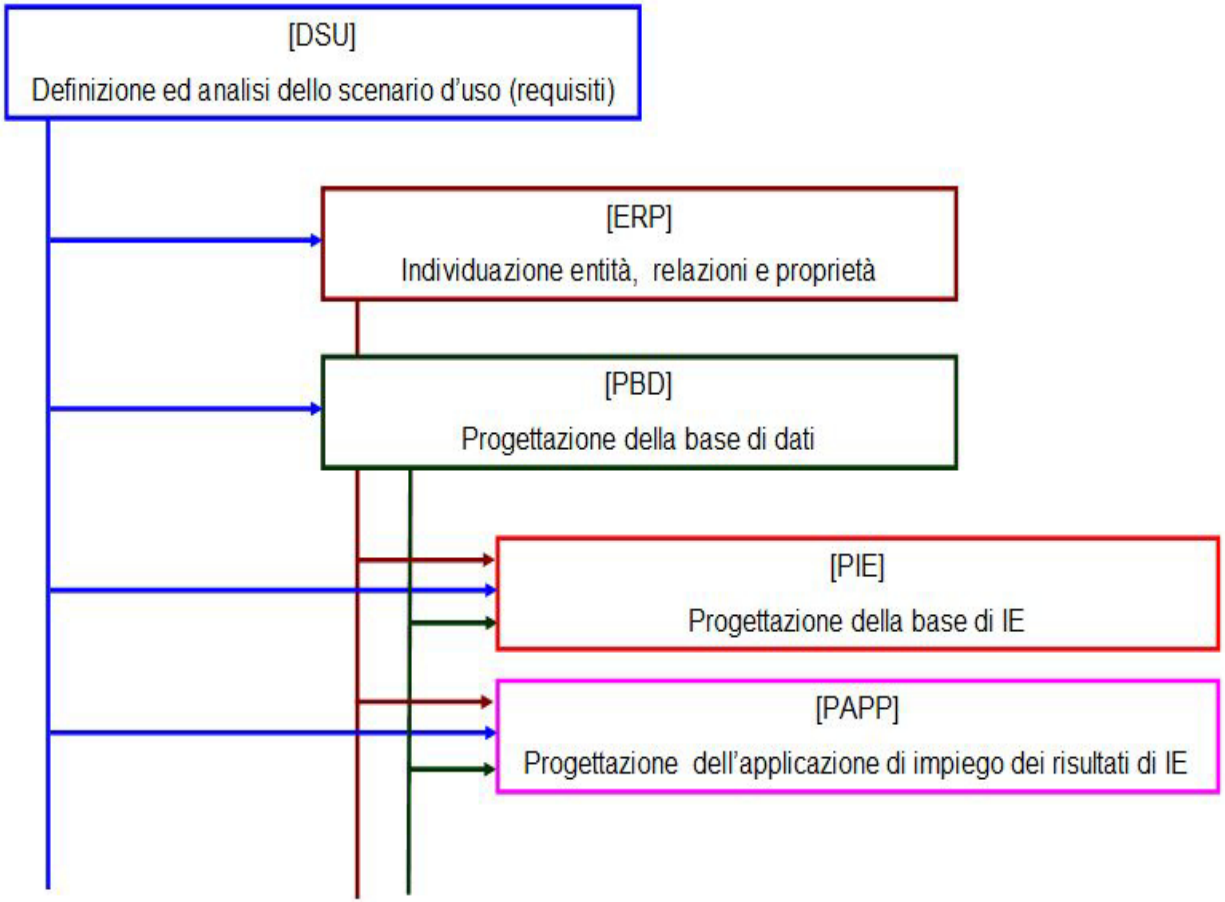
✓ in “machine learning” prevale l’ impegno di addestramento e verifica

✓ in “rule based” prevale l’impegno di disegnare un insieme efficace di glossari e grammatiche di regole

✓ In entrambi i casi è richiesto un team con un equilibrato mix di competenze ed esperienze in campo linguistico, conoscenza del dominio e padronanza degli strumenti tecnologici

✓ Gli approcci “rule based” e “machine learning” non sono mutuamente esclusivi ma possono essere combinati

Il metodo



L'approccio e il metodo …

[DSU]

Definizione ed analisi dello scenario d'uso (requisiti)

Destinatari

Esigenze

Aspettative (obiettivi)

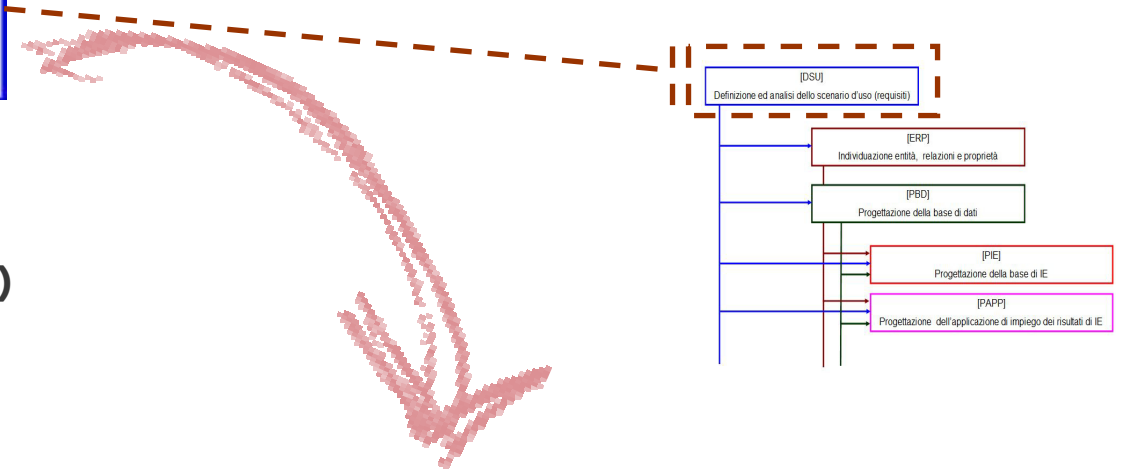
Risultati (misura dei risultati à metriche di misura)

Funzionalità

Casi d'uso

Criteri e livelli di accessibilità e impiego in relazione alle diverse esigenze dello scenario

Sorgenti documentali/informative non strutturate da trattare (caratteristiche, formati, ubicazione, reperibilità, ecc)



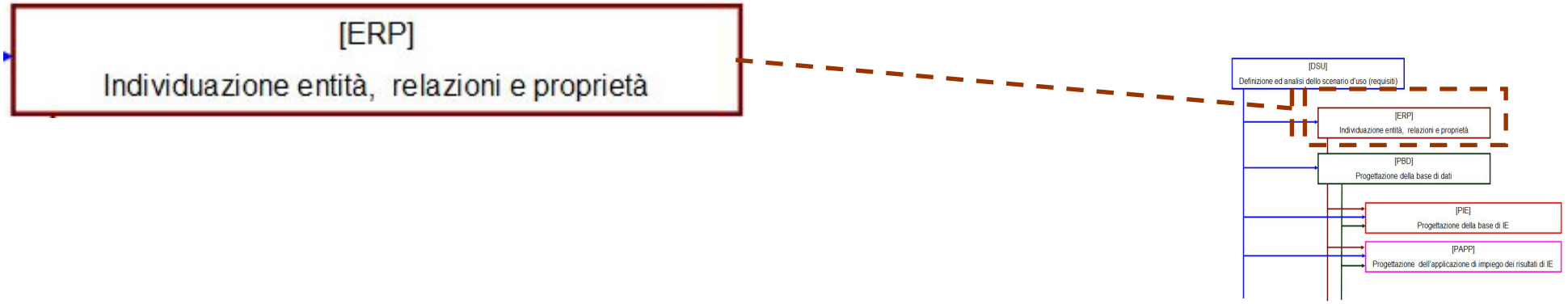
Delineato, analizzato e documentato lo scenario d'uso di interesse si perverrà alla definizione:

✓ **del contesto linguistico (italiano, inglese, ecc.)**

✓ **del dominio di applicazione recante terminologie, forme espressive e aspetti semantici del contesto, da "dominare" pienamente attraverso le competenze degli "esperti di dominio".**

Importante è predefinire obiettivi delle metriche di misura dei risultati.

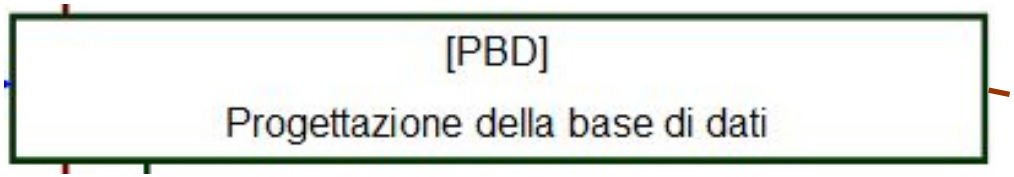
… L'approccio e il metodo …

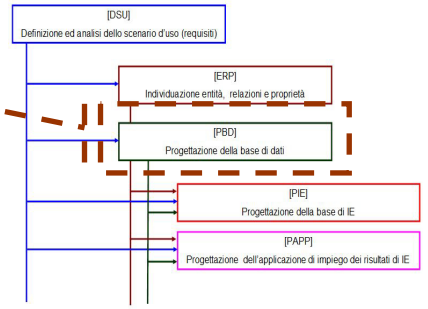


✓ Analisi, costruzione e documentazione dell'**ontologia** completa del dominio di interesse per lo scenario, articolata in **classi**, **sottoclassi**, **proprietà** e **relazioni** e, infine, alla valorizzazione delle **istanze principali** in relazione al contesto.

✓ Verifica di una corretta identificazione e denominazione delle entità informative individuate in relazione al dominio e alle pratiche di impiego e di rappresentazione dei destinatari d'uso.

… L'approccio e il metodo …





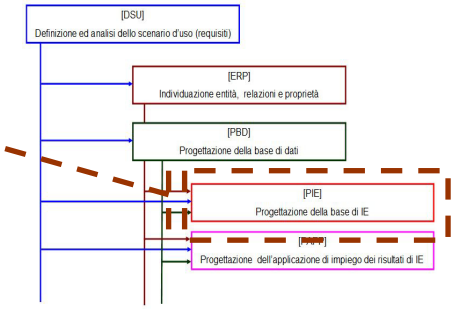
✓ Il disegno della base dati, prima concettuale e poi logico-fisico, determinerà, in funzione delle esigenze funzionali e degli scopi della soluzione, le strutture dati necessarie per il miglior utilizzo delle informazioni estraibili attraverso i processi di IE.

✓ In funzione dello scenario la base dati potrà essere di tipo relazionale, NoSQL o mista, e ciò comporterà approcci e forme di rappresentazione differenti ma sicuramente tutte ruotanti intorno al dominio di interesse espresso dalle ontologie e in funzione delle esigenze tecnologiche IT impiegate per il soddisfacimento dei requisiti dell'applicazione.

… L'approccio e il metodo …

[PIE]

Progettazione della base di IE



Punto di partenza:

✓Sorgenti non strutturate individuate in [DSU]
✓Ontologie definite in [ERP]
✓Base di dati disegnata in [PBD]
✓Esigenze funzionali definite in [DSU]
✓Gradualità dell'implementazione funzionale dell'applicazione sulla base di quanto rilevato in [DSU] ed economicamente sostenibile

I principali compiti di questa fase:

- Analisi delle sorgenti non strutturate
  - Formati di interesse, tipologie di documenti
  - Espressioni linguistiche di riferimento (operando in approccio "rule-based" ciò risulta necessario per la determinazione e l'impiego di glossari, elementi morfologico-lessicali e sintattici)
  - Effetto dell'applicazione di eventuali diversi livelli di analisi e interpretazione delle sorgenti in funzione della loro tipologia e/o del grado di dettaglio richiesto/ottenibile

- Individuazione dell'impiego di eventuali forme di "machine learning" propedeutiche o di supporto alle grammatiche "rule-based".

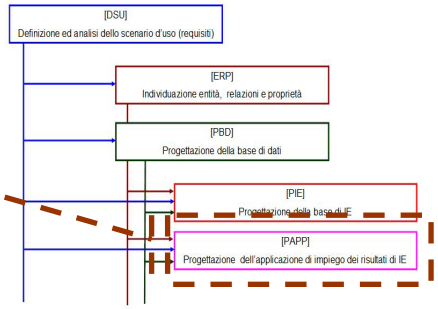
- Progettazione delle modalità e delle sequenze di training per le eventuali parti di "machine learning".

- Progettazione dei glossari e delle pipeline (sequenze di processi) di IE per la generazione delle annotazioni strutturate in relazione alle esigenze funzionali dell'applicazione

… L'approccio e il metodo

[PAPP]

Progettazione dell'applicazione di impiego dei risultati di IE



Punto di partenza:

✓Ontologie definite in [ERP]
✓Risultati di [PIE]
✓Base di dati disegnata in [PBD]
✓Esigenze funzionali definite in [DSU]
✓Gradualità dell'implementazione funzionale dell'applicazione sulla base di quanto rilevato in [DSU] ed economicamente sostenibile

Pincipali compiti di questa fase:

✓progettazione del sistema di "orchestrazione" (schedulazione e gestione dell'elaborazione) delle pipeline di IE
✓progettazione dell'integrazione dei risultati di IE alla base dati (relazionale e NoSql)
✓progettazione di funzioni/servizi in relazione all'utilizzo dei dati estratti da IE e disponibili sulla base dati
✓progettazione della gestione (alimentazione, manutenzione, monitoraggio, controllo) dei glossari/dizionari di supporto.

Alcuni esempi di altri campi di applicazione …

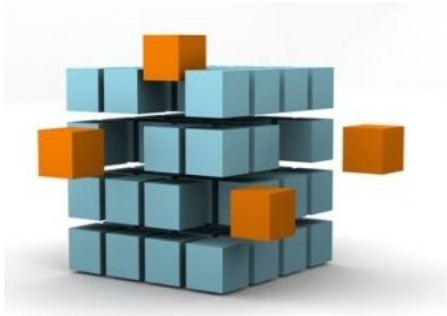
Relazioni di accompagnamento al bilancio

I documenti, pur ricorrendo a rappresentazioni sintetiche, come tabelle, grafici, ecc, riportano in genere molte delle principali informazioni attraverso l'uso del linguaggio naturale proprio del "dominio" amministrativo/finanziario….

… in essi si "accumulano" un qrande quantità di informazioni e "concetti" sviluppati, di volta in volta, dagli "esperti di dominio" (amministratori, revisori, ecc) e che rischiano di andare perduti o, quantomeno, non completamente utilizzati.

Un "sistema" IT che analizzi questi documenti sulla base del loro contenuto in linguaggio naturale alla ricerca di "concetti", informazioni e dati da rendere fruibili in forma strutturata e catalogata permette di estendere la conoscenza di una organizzazione, facilitando inoltre il compito degli "esperti di dominio", lasciati liberi da compiti di rilettura e rianalisi completa dei precedenti documenti, in ogni operazione di auditing e valutazione dell'organizzazione.

… Alcuni esempi di altri campi di applicazione …



Consolidamento di anagrafiche di prodotto

In ogni impresa la proliferazione di nuove anagrafiche e di duplicati è causata spesso dai limiti di gestione e dalla complessità dei processi dei sistemi informativi utilizzati, nei quali dati e informazioni, oltre che su “database” strutturati, sono contenuti anche in fonti discorsive di varia natura (pagine web, cataloghi, relazioni e rapporti tecnici, documenti marketing e commerciali, ecc).

Inoltre nei sempre più frequenti processi di fusione, acquisizione, creazione di reti di imprese risulta determinante la capacità di mettere a fattor comune, in un complesso omogeneo e coerente, questa mole di informazioni in genere organizzate, catalogate, codificate e descritte secondo criteri e modalità differenti…

…per analizzare e ricostruire automaticamente il patrimonio di anagrafiche dell’organizzazione a partire da tutte le sorgenti disponibili, strutturate e non, supportando e facilitando processi di normalizzazione, razionalizzazione, ottimizzazione e riorganizzazione.

… Alcuni esempi di altri campi di applicazione …

Gestione magazzino

•avvio dell'inventario con selezione dei prodotti da inventariare.
•conta fisica dei prodotti selezionati per l'inventario.
•inserimento dei dati di esistenza dei prodotti inventariati nel programma di gestione.
•chiusura della procedura d'inventario per la valorizzazione finale.

… ma vi sono spesso documenti aggiuntivi del "ciclo del magazzino".Si pensi ad esempio a quelli relativi all'analisi e al controllo dei rischi, alla gestione della qualità del prodotto o a quelli che descrivono modalità e aspetti legati al deperimento di merci e prodotti aventi effetto diretto sui criteri di valorizzazione.

Anche in questi casi il numero, la qualità e il valore delle informazioni contenute va molto al di la di quanto presente nel sistema informativo tradizionale.

… Alcuni esempi di altri campi di applicazione …

Verbali di inventario in ambito civilistico

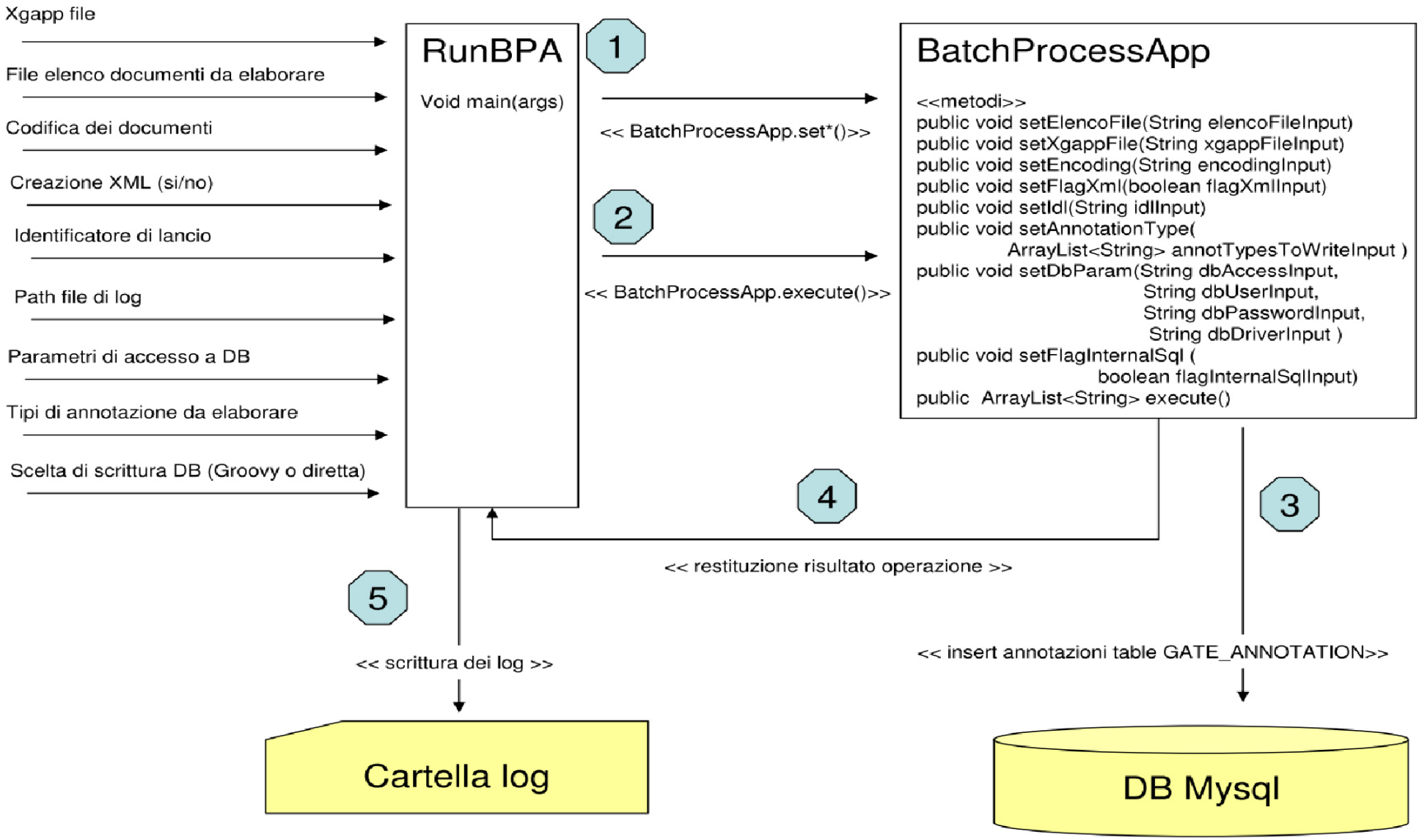
•la descrizione degli immobili, mediante l'indicazione della loro natura, della loro situazione, dei loro confini e dei numeri del catasto e delle mappe censuarie;
•la descrizione e la stima dei mobili, con la specificazione del peso o del marchio per gli oggetti d'oro e d'argento;
•l'indicazione della quantità e specie delle monete per il danaro contante;
•l'indicazione delle altre attività e passività;
•la descrizione delle carte, scritture e in fine dall'ufficiale procedente.

Documenti che, pur nascendo in forma non strutturata e secondo forme e linguaggio proprio del "dominio" giuridico, riportano nel loro contenuto una vasta e ricca quantità di informazioni.

Alcuni esempi di altri campi di applicazione

… e ultimo ma non ultimo …

Esecuzione di un'app GATE …



… Esecuzione di un'app GATE …

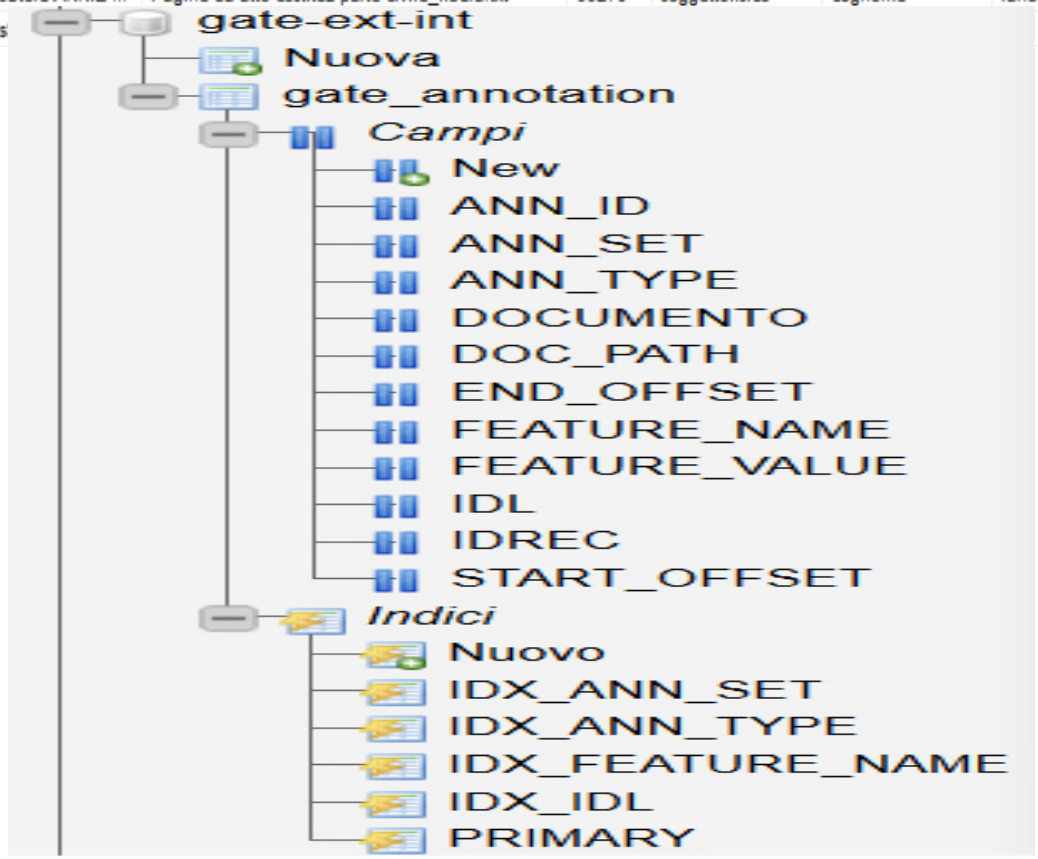
Start Tool

Annotazioni Navigazione

Id. di lancio … 20170310155725353 | Doc. path … | Documento …. | Annotazione … soggettofisico | Proprietà … | Valore proprietà … | Cerca | X

X

IDL	Path	Documento	Id Ann.	Tipo Ann.	Proprietà	Valore	Start Offset	End Offset
20170310155725353	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65270	soggettofisico		/	126	140
20170310155725353	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65270	soggettofisico	codice fiscale	/	126	140
20170310155725353	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65270	soggettofisico	cognome	rando	126	140
20170310155725353	C:\roberto\progetti\Gate\MyProjects						126	140



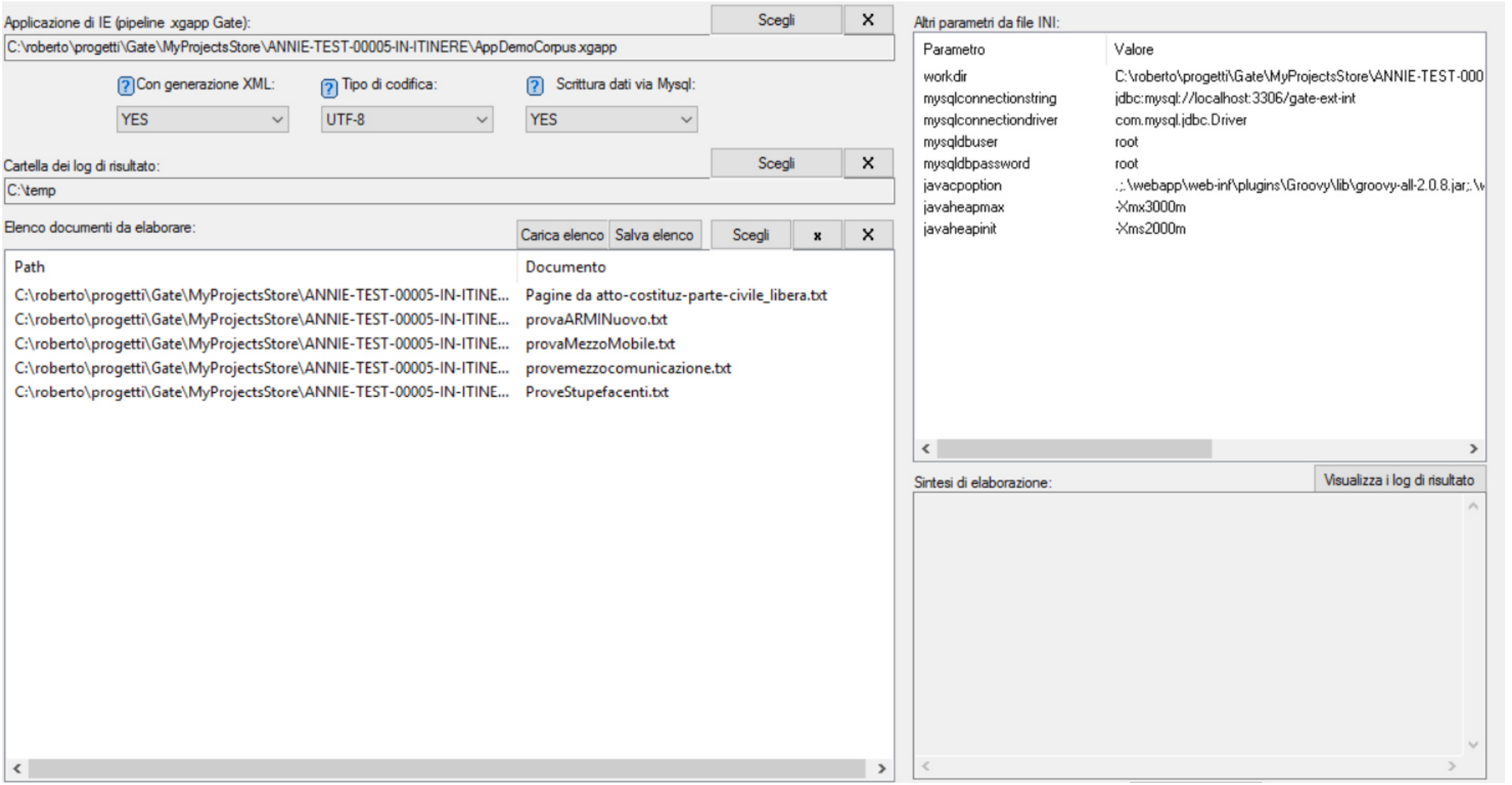
… Esecuzione di un'app GATE …

RunBPA-runjava.cmd - Blocco note

File Modifica Formato Visualizza ?

```
@echo off
cls

echo Esecuzione RunBPA EXTAPP di test integrazione GATE ...
echo .
echo Start - %TIME%
echo .

java                                                                    ^
  -Xmx2048m -Xms2048m                                                   ^
  -cp ".;.\webapp\web-inf\plugins\Groovy\lib\groovy-all-2.0.8.jar;.\webapp\web-inf;.\webapp\web-inf\lib\*"      ^
  webapp.classes.runbpa.RunBPA                                          ^
  -g "..\..\appextintconditional.xgapp"                                 ^
  -e UTF-8                                                              ^
  -l ".\ElencoDocumenti-3.txt"                                          ^
  -x YES                                                                ^
  -p "C:\temp"                                                          ^
  -s "jdbc:mysql://localhost:3306/gate-ext-int"                         ^
  -u root                                                               ^
  -k root                                                               ^
  -d "com.mysql.jdbc.Driver"                                            ^
  -w YES                                                                ^
  -a soggettofisico                                                     ^
  -a soggettogiuridico                                                  ^
  -a associazionecriminale

echo .
echo End   - %TIME%
echo .

pause
```

… Esecuzione di un'app GATE



… Esecuzione di un'app GATE

```
// fissiamo gate.home nelle convenzioni del pacchetto di test
File gateHome = new File(gateHomeDir);
Gate.setGateHome(gateHome);

// fissiamo gate.user.config su gate.home
// gate.user.session in embedded java non viene usata e quindi non è necessario definirla
Gate.setUserConfigFile(new File(gateHome, gateUserConfig));


// inizializzaione GATE prima dell'esecuzione di qualsiasi API Gate
Gate.init();


  gappFile = new File(xgappFile);

  // caricamento dell'applicazione Gate
  CorpusController application =
      (CorpusController)PersistenceManager.loadObjectFromFile(gappFile);

  // Creazione di un Corpus. Si riutilizza sempre lo stesso per ogni iterazione.
  // Il nome interno del Corpus non ha particolare significato.
  Corpus corpus = Factory.newCorpus("BatchProcessApp Corpus");
  application.setCorpus(corpus);


// lettura della lista dei file documento da elaborare
BufferedReader br = new BufferedReader(new FileReader(elencoFileDocumenti));
String docFileName = br.readLine();

 while (docFileName != null) {

        // caricamento del documento corrente con la codifica specificata (nelle ipotesi UTF-8)
        File docFile = new File(docFileName);
        if (docFile.exists() ) { // se il file esiste ....
                    doc = Factory.newDocument(docFile.toURI().toURL(), encoding);

                    // collocazione del documento nel corpus
                    corpus.add(doc);
        }
```

… Esecuzione di un'app GATE

```
// esecuzione della pipeline
application.execute();

mysqlBeginTransaction();

int retCode = annotationReadAndWriteDB(docFile);

mysqlCommitTransaction();

// rimozione del documento dal corpus
corpus.clear();

// l'intero documento in formato XML
docXMLString = doc.toXml();

// creazione del file XML con l'aggiunta dell'estensione ".out.xml"
String outputFileName = docFile.getName() + ".out.xml";
File outputFile = new File(docFile.getParentFile(), outputFileName);

// Scrittura del file XML utilizzando la codifica in input
FileOutputStream fos = new FileOutputStream(outputFile);

BufferedOutputStream bos = new BufferedOutputStream(fos);
OutputStreamWriter out;
if(encoding == null) {
    out = new OutputStreamWriter(bos);
}
else {
    out = new OutputStreamWriter(bos, encoding);
}

out.write(docXMLString);
out.close();


// rilascio del documento corrente non più necessario
Factory.deleteResource(doc);


// prossimo documento (file) dalla lista in input
docFileName = br.readLine();
}
```

… Esecuzione di un'app GATE

```
private int annotationReadAndWriteDB(File docFile) {
// lettura delle annotazioni del documento corrente. Per ogni annotazione rientrante nell'elenco dei tipi di annotazione da elaborare viene effettuata la registrazione su DB MYSQL dei dati
    FeatureMap          mapFeatureNameValue=null;
    int                 sqlFlagError=0; // OK

    AnnotationSet inputAS = doc.getAnnotations();

    for (Annotation annotation : inputAS) {

        String currentAnnType = (String) annotation.getType();
        currentAnnType         = currentAnnType.toLowerCase();
        if ( annotTypesToWrite != null && annotTypesToWrite.contains(currentAnnType) ) {
            String StartOffset  =  String.valueOf(annotation.getStartNode().getOffset());
            String EndOffset    =  String.valueOf(annotation.getEndNode().getOffset());
            String annId        =  String.valueOf(annotation.getId());

            //mapFeatureNameValue.clear();
            mapFeatureNameValue = (FeatureMap) annotation.getFeatures();
            Set  annotationFeatureNames  = mapFeatureNameValue.keySet();
            Iterator annotationFeatureNamesIterator = annotationFeatureNames.iterator();

            while ( annotationFeatureNamesIterator.hasNext() ) {
                Object Key          =   annotationFeatureNamesIterator.next();
                String featureName  =   (String) Key;
                featureName         =   featureName.toLowerCase();

                Object Value        =   mapFeatureNameValue.get(Key);
                String featureValue = "";
                if ( Value != null ) {
                    featureValue =  (String) Value;
                }
                featureValue = featureValue.toLowerCase().replace("\'", "\\'");

                String sqlCmd = ("insert into gate_annotation values(null,'" + idl + "','" + docFile.getParent().replace("\\", "\\\\") + "','" + docFile.getName() + "','default'," + annId + ",'" +
                currentAnnType + "',"+ StartOffset + "," + EndOffset + ",'" + featureName + "','" + featureValue + "')");

                //System.out.println(sqlCmd);
                try {
                    mysqlExec(sqlCmd);
                }
                catch (Exception Ex) {
                    System.out.println("Errore in insert SQL interna su statement: \r\n" + sqlCmd);
                    sqlFlagError = -1;
                }
            }
        }
    }
    return sqlFlagError;
}
```

… Esecuzione di un'app GATE …

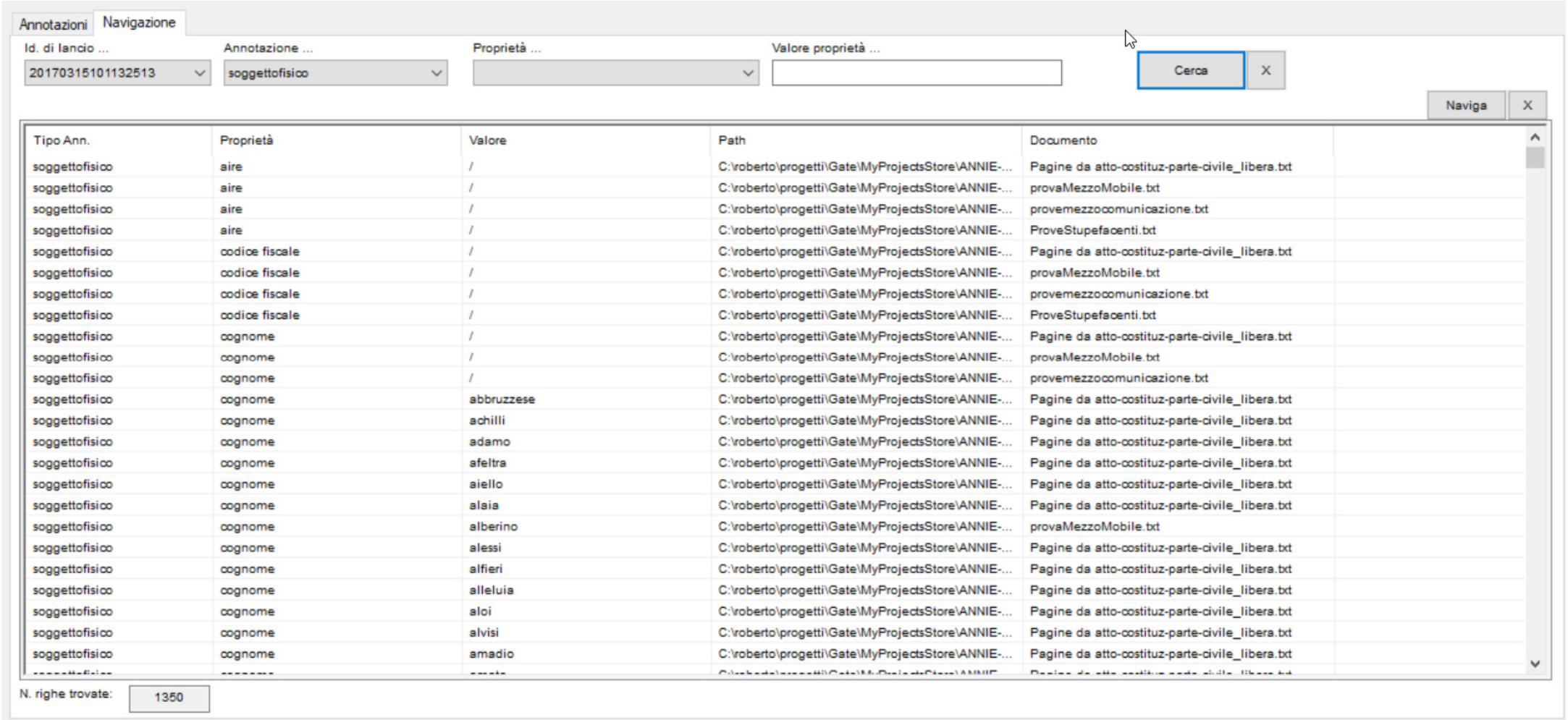
Annotazioni | Navigazione

Id. di lancio … 20170315101132513 | Doc. path … | Documento …. | Annotazione … soggettofisico | Proprietà … | Valore proprietà … | Cerca | X

IDL	Path	Documento	Id Ann.	Tipo Ann.	Proprietà	Valore	Start Offset	End Offset
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65295	soggettofisico	aire	/	126	140
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65295	soggettofisico	codice fiscale	/	126	140
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65295	soggettofisico	cognome	rando	126	140
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65295	soggettofisico	data di nascita	/	126	140
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65295	soggettofisico	indirizzo	/	126	140
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65295	soggettofisico	indirizzo studio	via falloppia piano 53, mo...	126	140
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65295	soggettofisico	luogo di nascita	/	126	140
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65295	soggettofisico	maternita	/	126	140
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65295	soggettofisico	nome	vincenza	126	140
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65295	soggettofisico	partita iva	/	126	140
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65295	soggettofisico	paternita	/	126	140
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65295	soggettofisico	rule	sft3,sfdn6	126	140
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65295	soggettofisico	soprannome	/	126	140
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65295	soggettofisico	titoli	avv.	126	140
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65296	soggettofisico	aire	/	416	432
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65296	soggettofisico	codice fiscale	/	416	432
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65296	soggettofisico	cognome	ciotti	416	432
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65296	soggettofisico	data di nascita	/	416	432
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65296	soggettofisico	indirizzo	/	416	432
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65296	soggettofisico	luogo di nascita	/	416	432
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65296	soggettofisico	maternita	/	416	432
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65296	soggettofisico	nome	pio luigi	416	432
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65296	soggettofisico	partita iva	/	416	432
20170315101132513	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt	65296	soggettofisico	paternita	/	416	432

N. righe trovate: 9579

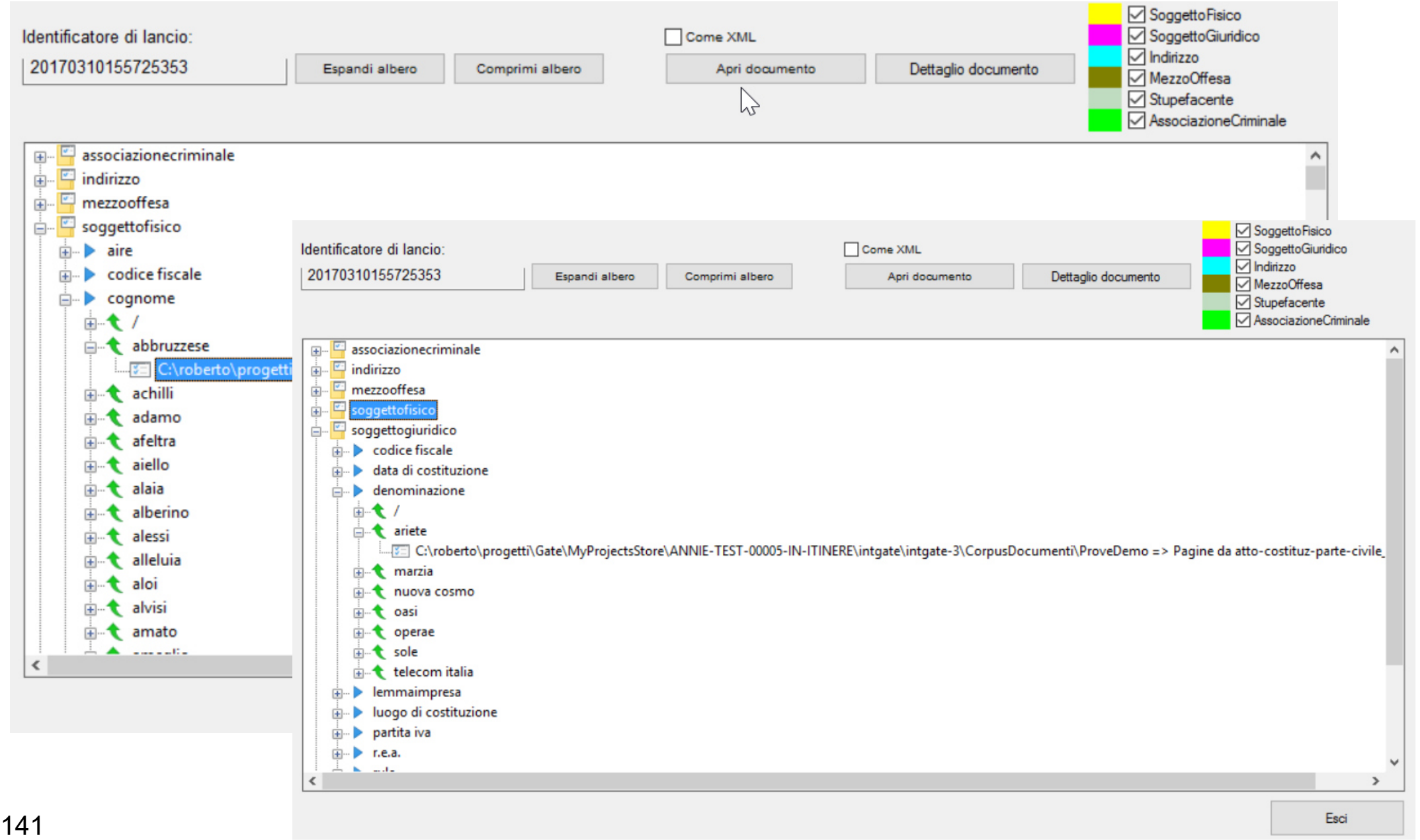
… Esecuzione di un'app GATE …



Tipo Ann.	Proprietà	Valore	Path	Documento
soggettofisico	aire	/	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt
soggettofisico	aire	/	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	provaMezzoMobile.txt
soggettofisico	aire	/	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	provemezzocomunicazione.txt
soggettofisico	aire	/	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	ProveStupefacenti.txt
soggettofisico	codice fiscale	/	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt
soggettofisico	codice fiscale	/	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	provaMezzoMobile.txt
soggettofisico	codice fiscale	/	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	provemezzocomunicazione.txt
soggettofisico	codice fiscale	/	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	ProveStupefacenti.txt
soggettofisico	cognome	/	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt
soggettofisico	cognome	/	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	provaMezzoMobile.txt
soggettofisico	cognome	/	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	provemezzocomunicazione.txt
soggettofisico	cognome	abbruzzese	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt
soggettofisico	cognome	achilli	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt
soggettofisico	cognome	adamo	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt
soggettofisico	cognome	afeltra	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt
soggettofisico	cognome	aiello	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt
soggettofisico	cognome	alaia	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt
soggettofisico	cognome	alberino	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	provaMezzoMobile.txt
soggettofisico	cognome	alessi	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt
soggettofisico	cognome	alfieri	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt
soggettofisico	cognome	alleluia	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt
soggettofisico	cognome	aloi	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt
soggettofisico	cognome	alvisi	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt
soggettofisico	cognome	amadio	C:\roberto\progetti\Gate\MyProjectsStore\ANNIE-...	Pagine da atto-costituz-parte-civile_libera.txt

N. righe trovate: 1350

Esempi d'uso di un'app GATE …



… Esempi d'uso di un'app GATE

Documento: \IE-TEST-00005-IN-ITINERE\CorpusDocumenti\ProveDemo => Pagine da atto-costituz-parte-civile_libera.txt

Entità: dirizzo', 'mezzooffesa', 'stupefacente', 'associazionecriminale'

Proprietà:

Valore proprietà:

SoggettoFisico SoggettoGiuridico Indirizzo MezzoOffesa Stupefacente AssociazioneCriminale

cosmo Cerca Pulisci

70. MORRA Giovanni, nato a Cutro (KR) il 20.09.1965, residente a Sorbolo (PR), in via Mimmi Fochi nr. 24, con domicilio eletto c/o la propria residenza, assistito e difeso di fiducia dall'Avv. Gino Scalzi del Foro di Parma

71. SIRRI Carmelo, nato a Cutro (KR) l'11.05.1967, residente a Reggio Emilia, via Chiesi nr. 16, con domicilio eletto c/o la propria residenza, assistito e difeso di fiducia dall'Avv. Piero Lolli del Foro di Reggio Emilia

72. GOZZO Carlo Alberto, nato a Follo (SP) il 03.01.1945, residente a Parma in Via Fratelli Cervi nr. 2, con domicilio eletto c/o la sede della "Nuova Cosmo s.r.l.", assistito e difeso di fiducia dall'Avv. Miriam Valori del foro di Parma

73. SANNA Diego, nato a Palmi (RC) il 15.8.1957, residente a Milano in via Frosinone nr. 72, attualmente detenuto c/o Casa Circondariale di Voghera, con domicilio
eletto c/o la propria residenza, assistito e difeso di fiducia dall'Avv. Laura Valenti del Foro di Pescara e dall'Avv. Giorgio Alfieri del Foro di Milano

74. MIRRI BERNARDINI Orlando, nato a Pietrasanta (LU) il 12.01.1976, residente a Vareggio (LU), in via Machiavelli nr. 149, assistito e difeso di fiducia dall'Avv. Salvatore Mariani del Foro di Lucca, con domicilio eletto c/o lo studio del difensore

75. FLORIS Alfredo, nato a Locri (RC) il 27.09.1989, residente a Portigliola (RC), c.da Pirettina nr. 9, di fatto domiciliato a Montecchio Emilia (RE), via Pampari nr. 5, difeso di ufficio dall'Avv. Piero Calemi del Foro di Bologna

76. ZANNA VITO Alcide, nato a Crotone il 07.01.1980, residente a Reggio Emilia via Montessori nr. 2/2, attualmente detenuto c/o Casa Circondariale di Genova Marassi, assistito e difeso di fiducia dall'Avv. Alessandro Scassi del Foro di Reggio Emilia e dall'Avv. Mario Bianchi del Foro di Modena, con domicilio eletto c/o lo studio dell'Avv. Scassi

77. VERDI VITO Antonio, detto Giuseppe, nato a Crotone il 20.03.1978, residente a Cadelbosco di Sopra (RE), via Dante Alighieri nr. 53/2, di fatto domiciliato a Scandiano (RE), frazione Arceto, via 2 Agosto 1980 Vittime di Bologna nr. 4, attualmente detenuto c/o Casa Circondariale di Bologna, assistito e difeso di fiducia dall'Avv. Piero Orlandi e dall'Avv. Stella Orsini entrambi del Foro di Bologna

78. SALEMI Giuseppina, nata a Crotone il 15.06.1974, residente a Cutro, via Firenze nr. 1, con domicilio eletto c/o la propria residenza, assistita e difesa di fiducia dall'Avv. Gianni Sarzani del Foro di Reggio Emilia

79. CALANDRA Savino, detto Salvatore, nato a Crotone il 03.03.1976, residente a Reggio Emilia, via Ferri nr. 21/6, di fatto domiciliato in Viale Regina Elena nr. 16, assistito e difeso di fiducia dall'Avv. Mario Amadio del Foro di Reggio Emilia e dall'Avv. Luca Anagni del Foro di Modena, con domicilio eletto c/o lo studio del difensore Avv. Amadio

80. FOGGIA Filippo, nato in Germania il 22.05.1975, residente a Crotone, in via I Trav. di via Crisone nr. 15, con domicilio eletto c/o la propria residenza, assistito e difeso di fiducia dall'Avv. Gabriele Marconi del Foro di Crotone

81. FAZZI Franco, nato a Reggio Emilia il 15.05.1981, ivi residente in via Unione Sovietica nr. 37, assistito e difeso di fiducia dall'Avv. Alberto Rossi del Foro di Reggio Emilia, con domicilio eletto c/o lo studio del difensore

<<

Proprietà	Valore
tiporagionesociale	/
tiposoggetto	/
Annotazione 67549	
Entità	soggettogiuridico
codice fiscale	/
data di costituzione	/
denominazione	nuova cosmo
lemmaimpresa	/
luogo di costituzione	/
partita iva	/
r.e.a.	/
rule	sg003_2,sgprop2
sede	/
tiporagionesociale	srl
tiposoggetto	impresacommerciale
Annotazione 67551	
Entità	associazionecriminale
contesto	ndrangheta
rule	ac1
Annotazione 67552	
Entità	mezzooffesa
[regola]	arma2
marca	/
misura	/
modello	
note	/
numero armi	/
tipo arma	armi

[es.: elaborazione atto pubblico-dati personali modificati]

Esecuzione di un app GATE

✓ Esempio di un' app di cross reference.

✓ Esempio di un'app di esecuzione di una pipeline su documenti variabili, con registrazione su DB MYSQL delle annotazioni.

✓ Esempio di un'app che impiega le annotazione su DB MYSQL.

Riflessioni e idee …

roberto.gallerani@ordingbo.it
https://www.gallerani.it

Grazie per l'attenzione

"We can not solve our problems with the same level of thinking that created them."
Albert Einstein