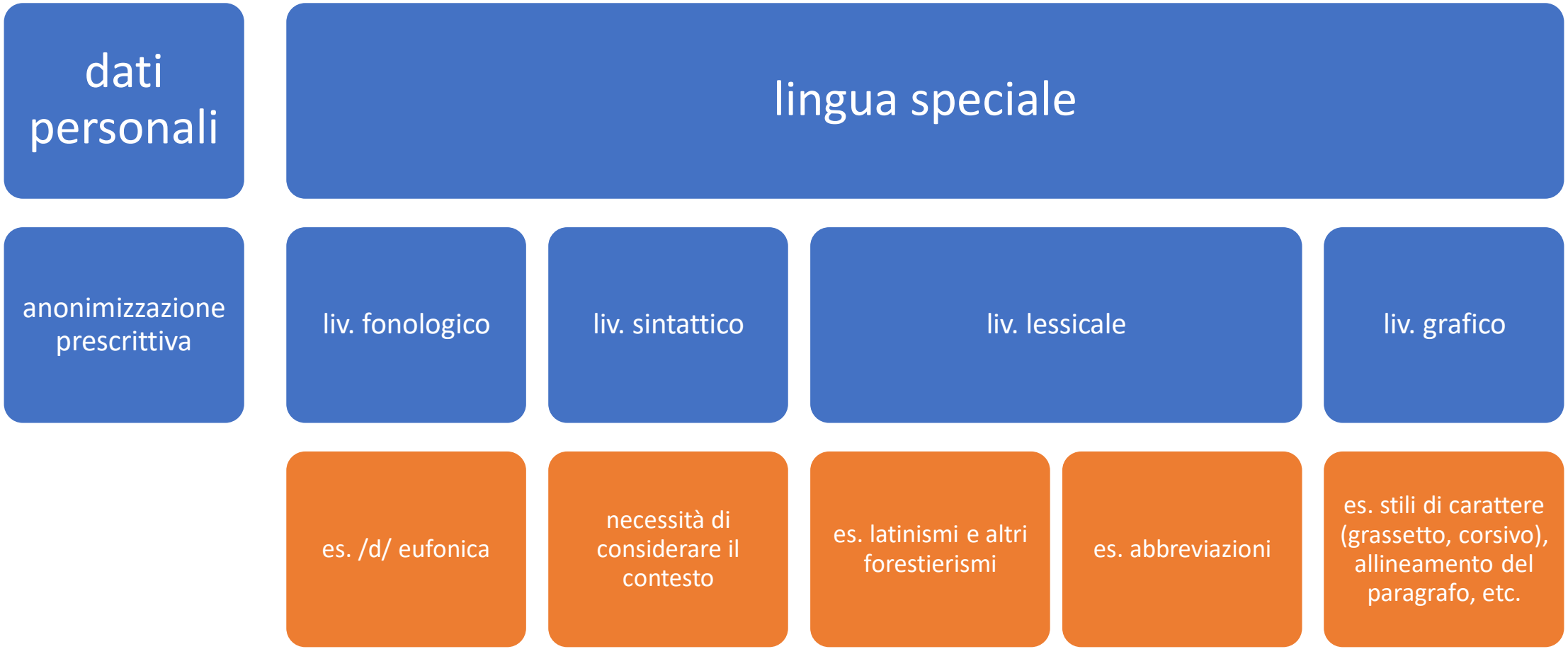
Nuovi strumenti digitali per la creazione e l'uso di corpora specialistici: il progetto "Atti Chiari"

XXXI Convegno Ass.I.Term
10 dicembre 2021

DANIELE FUSI

Peculiarità dei testi in rapporto all'analisi



Anonimizzazione

Approccio granulare

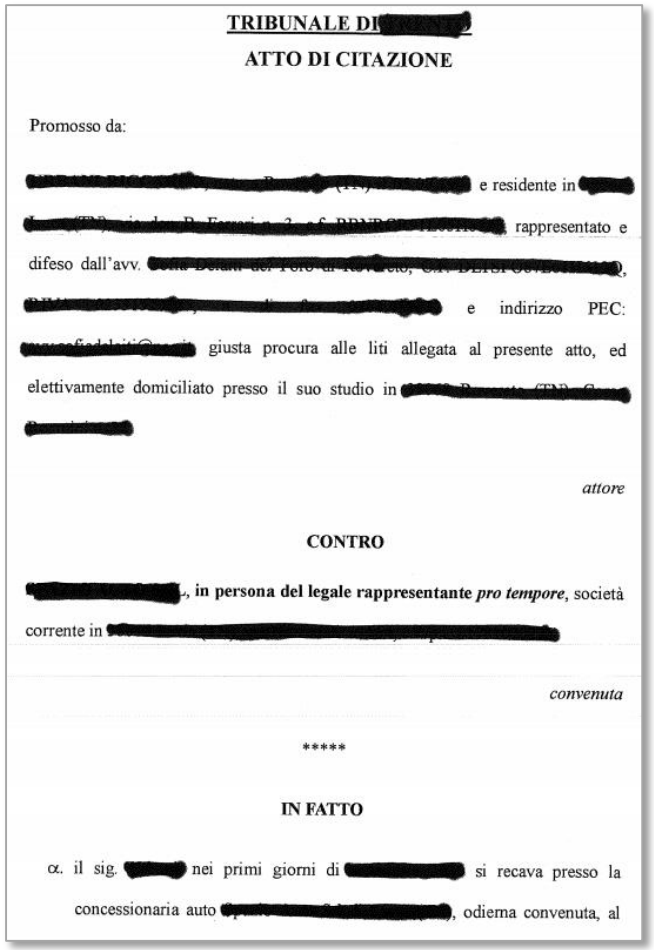
Anonimizzazione: approccio tradizionale

Procedura

1. documento originale
2. cancellazione di ogni dato sensibile
3. documento mutilo

Svantaggi

- testo mutilo:
  - reduce la leggibilità del testo
  - pregiudica l'analisi linguistica
- nessuna granularità del processo



TRIBUNALE DI [redacted]

ATTO DI CITAZIONE

Promosso da:

[redacted] e residente in [redacted], rappresentato e difeso dall'avv. [redacted], e indirizzo PEC: [redacted] giusta procura alle liti allegata al presente atto, ed elettivamente domiciliato presso il suo studio in [redacted]

attore

CONTRO

[redacted], **in persona del legale rappresentante *pro tempore***, società corrente in [redacted]

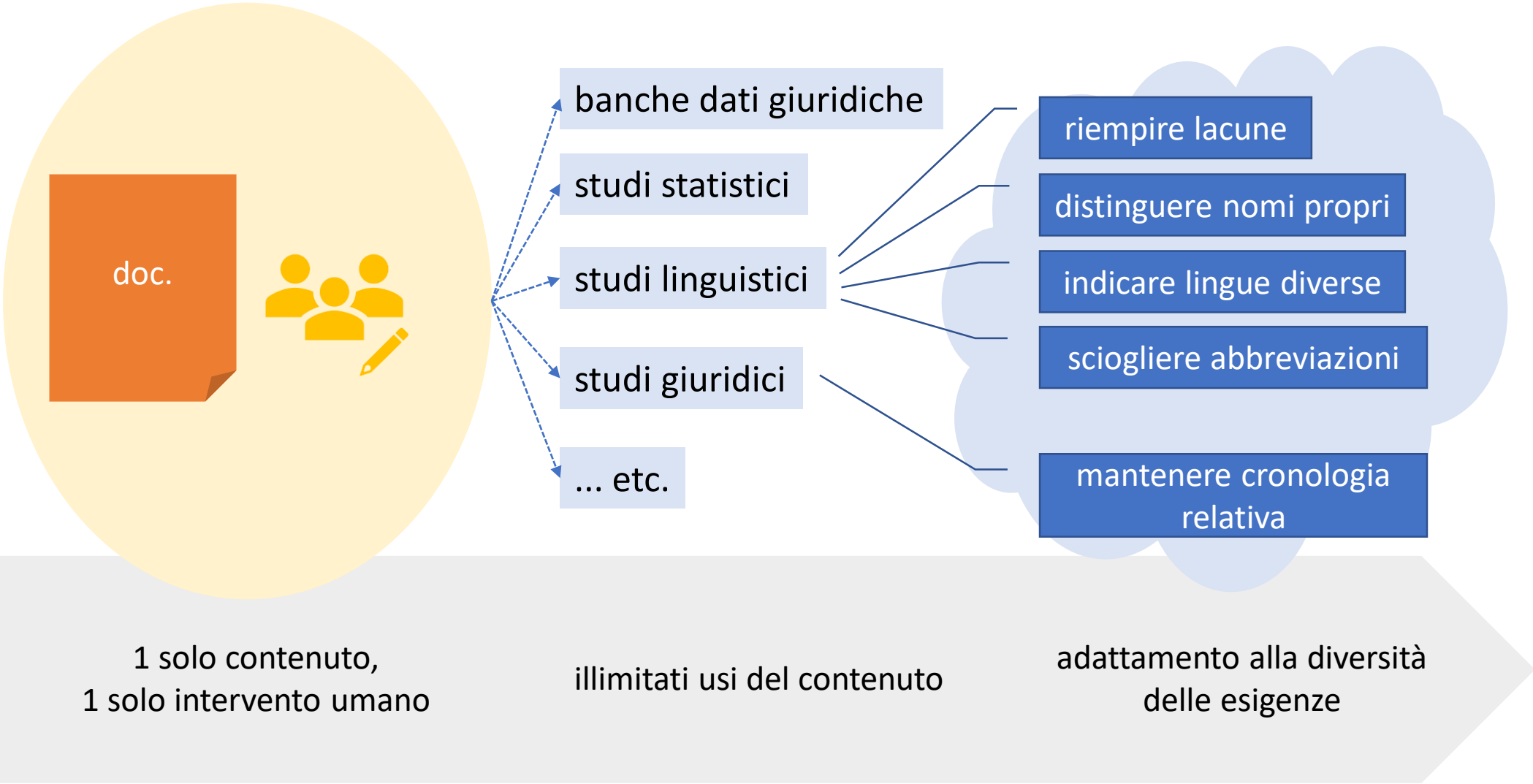
convenuta

*****

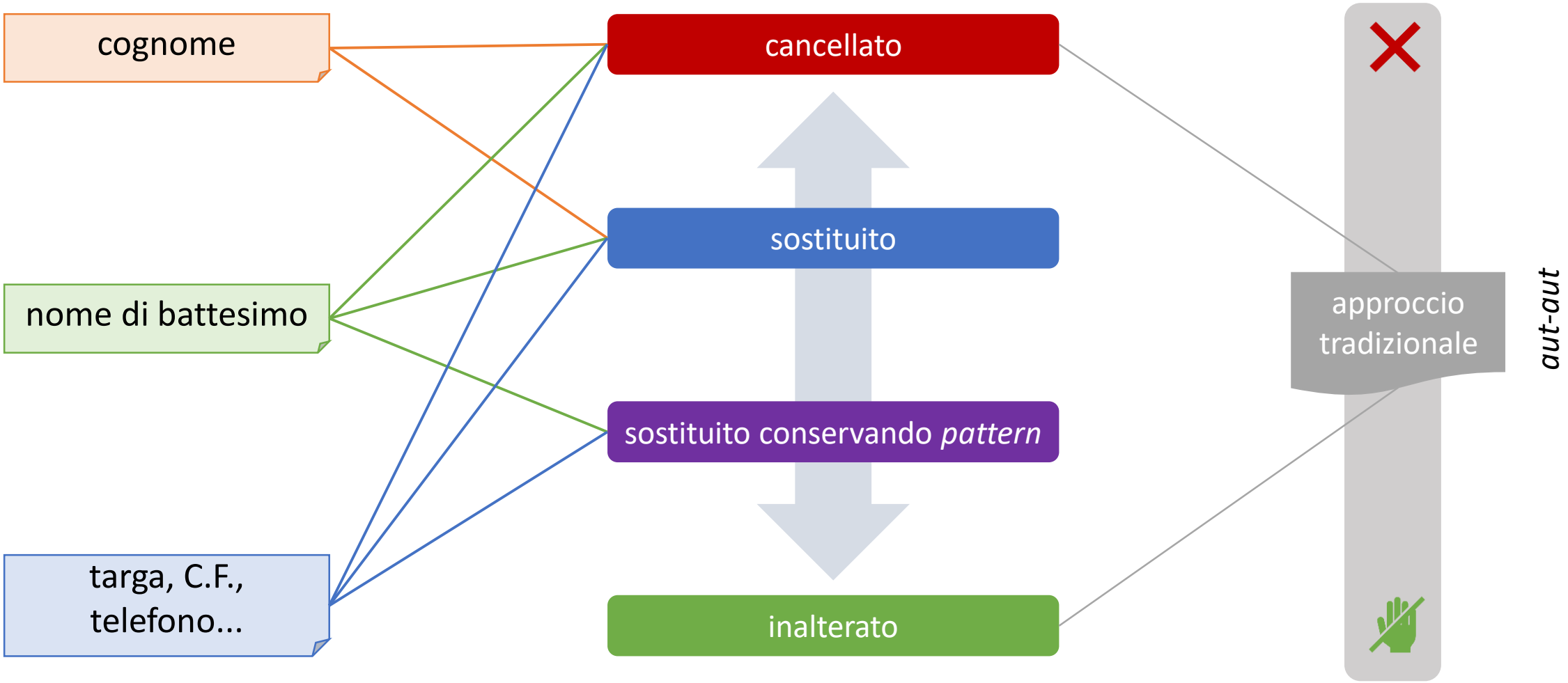
IN FATTO

α. il sig. [redacted] nei primi giorni di [redacted] si recava presso la concessionaria auto [redacted], odierna convenuta, al

Granularità e versatilità



Granularità variabile - esempi



Approccio alternativo: marcatura minimale

Procedura

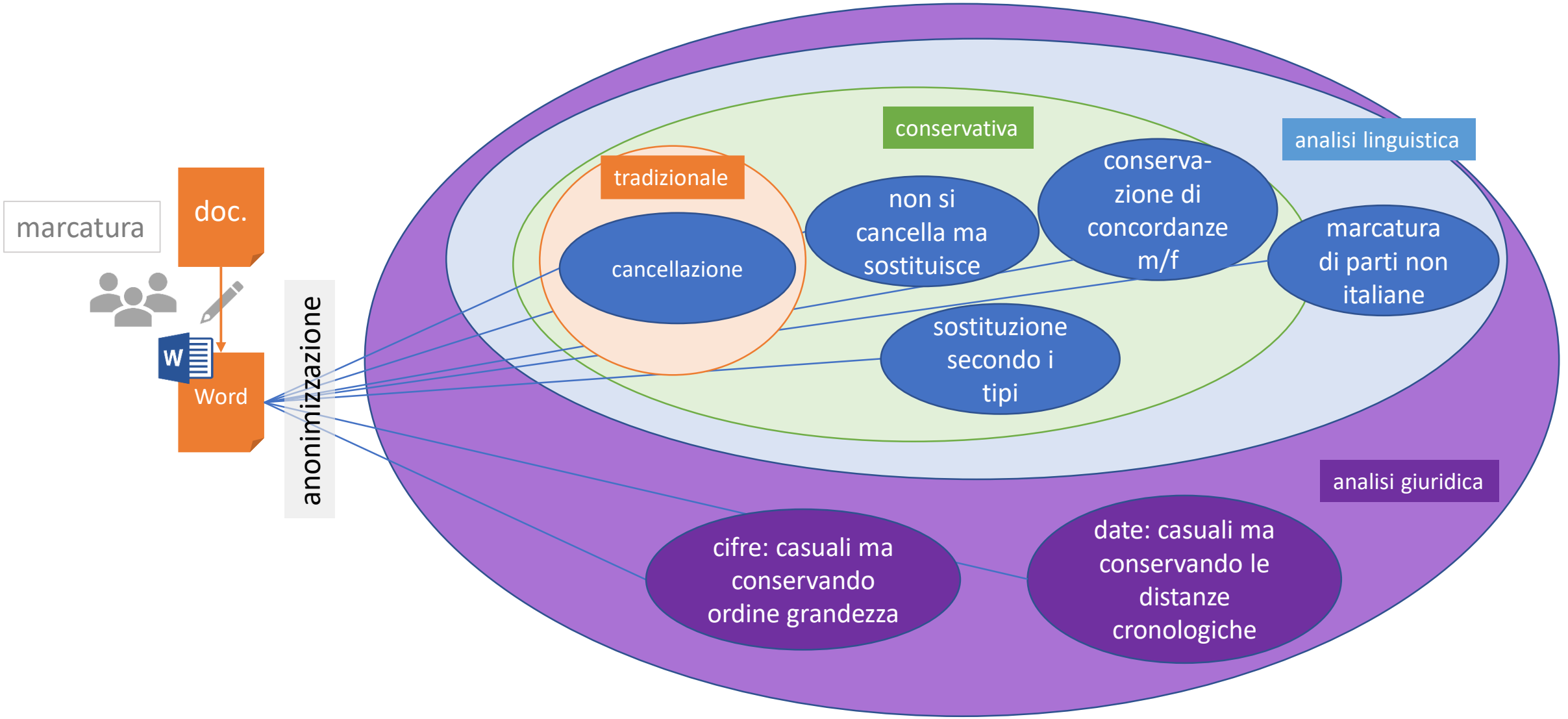
1. documento originale
2. marcatura di ogni dato personale secondo categorie generali
3. anonimizzazione automatica del documento marcato

... il signor {f-m:Mario} {f-l:Verdi}, nato a {t:Roma} il {d:1/2/1970}, c.f. {u:VRDMRA...} ...

questo **aggiunge** anziché togliere informazione; il risultato sarà poi la base per l'anonimizzazione automatica

l'informazione viene eliminata in una seconda fase, consentendo a questa procedura di essere granulare a seconda degli scopi

Adattamento della logica agli usi



Esempio: approccio tradizionale vs. marcatura

TRIBUNALE DI ########

SEZ. FALL.

COMPARSA DI COSTITUZIONE E RISPOSTA

del **Dott.** ######## ########, già Commissario Giudiziale del Concordato Preventivo ######## in liq., elettivamente domiciliato in ########, ########, presso lo studio dell'Avv. ######## ######## (c.f. ########; fax ########; pec: ########) che lo rappresenta ed assiste come da procura in calce alla presente comparsa

NEL PROCEDIMENTO RELATIVO

AL RECLAMO EX ART. 26 L.F.

(n. 1/12 C.P. – n. 11/12 R.R.N.C.)

PROPOSTO DA

######## **in liquidazione** (di seguito ########), in persona del liquidatore ######## Dott. ######## ########

- ricorrente -

PER L'ANNULLAMENTO

dell'ordinanza pronunciata il ########, su istanza del Commissario Giudiziale Dott. ######## del ########, dalla Dott.ssa ######## ########, Giudice Delegato del Tribunale di ######## nella procedura di concordato preventivo ######## (procedura oggetto di una prima proposta di concordato presentata da ########, con ricorso depositato il ########, successivamente modificata da ######## all'udienza del ########, e infine definitivamente rinunciata da ########, con conseguente declaratoria del Tribunale di ########, nella data del ########, di "non doversi provvedere sulla proposta di concordato per rinuncia del proponente").

TRIBUNALE DI INVORIO

sez. FALL.

COMPARSA DI COSTITUZIONE E RISPOSTA

del dott. Immacolato PAOLETTI, già Commissario Giudiziale del Concordato Preventivo Desolina in liq., elettivamente domiciliato in Invorio, Giodoco Gottifredi, 68, presso lo studio dell'avv. Ado Liguoro (c.f. HMVMQY03G67Y621X; fax 389-3174293; pec: hp1462@outlook.it) che lo rappresenta ed assiste come da procura in calce alla presente comparsa

NEL PROCEDIMENTO RELATIVO

AL RECLAMO EX art. 26 L.F.

(n. 1/12 c.p.n. 11/12 R.R.N.C.)

PROPOSTO DA

GUIA in liquidazione (di seguito Calliroe), in persona del liquidatore pro tempore dott. Giberto Taqi

- ricorrente -

PER L'ANNULLAMENTO

dell'ordinanza pronunciata il 9/11/2009, su istanza del Commissario Giudiziale dott. Leccio del 8/4/2007, dalla dott.ssa Emiliana Pittini, Giudice Delegato del Tribunale di Vetralla nella procedura di concordato preventivo Calliroe (procedura oggetto di una prima proposta di concordato presentata da Calliroe, con ricorso depositato il 12/10/2004, successivamente modificata da Calliroe all'udienza del 20/3/2001, e infine definitivamente rinunciata da Calliroe, con conseguente declaratoria del Tribunale di Vetralla, nella data del 17/12/2016, di "non doversi provvedere sulla proposta di concordato per rinuncia del proponente").

Conversione TEI e usi digitali

Dalla marcatura tipografica
alla marcatura semantica

Ricaduta del processo: conversione XML-TEI

marcatura nel documento

... il signor **Mario Verdi** , nato a **Roma** il **1/2/1970** , impiegato presso la ditta **ACME** Spa, con autovettura targata **FO392FI** , recatosi *de relato* in ritardo al lavoro ...

formato rich text

anonimizzatore

TXT

HTML

TEI

```
<body>
  <p rend="c"><choice><abbr>avv.</abbr><expan xml:lang="ita">
  avvocato</expan></choice> <persName type="mn">ASPROMONTE</
  persName> <persName type="s">CALLISTO</persName></p>
  <p rend="c">
    <address>
      <addrLine>ADELISO COLLIRI, 56</addrLine>
    </address>- <num>01845</num> <placeName>SOLIGNANO</placeName>
    </p>
  <p rend="c">PEC: <email>PA8405@HOTMAIL.COM</email></p>
  <p rend="c"><choice><abbr>avv.</abbr><expan xml:lang="ita">
  avvocato</expan></choice> <persName type="mn">FIERO</persName> <
  persName type="s">SGARGI</persName></p>
  <p rend="c">CASSAZIONE E GIURISDIZIONI SUPERIORI</p>
  <p rend="c">
    <address>
      <addrLine>JULIANO MINTO, 49</addrLine>
    </address> - <num>52026</num> <placeName>SOLIGNANO</placeName>
    </p>
  <p rend="c">PEC: <email>HJ4621@TISCALI.IT</email></p>
  <p rend="c"><choice><abbr>avv.</abbr><expan xml:lang="ita">
  avvocato</expan></choice> <persName type="fn">ARMENIA</persName>
   <persName type="s">BERTOLANI</persName></p>
```

Anonimizzazione e pretrattamento

- anonimizzazione = marcatura
- lo scopo della marcatura può andare oltre le necessità pratiche abbracciando quelle dell'analisi

estensione dei marcatori

- forestierismi (con indicazione della lingua)
- abbreviazioni: scioglimento automatico per:
  - maggior leggibilità
  - evitare che i punti siano fraintesi come terminatori di frase

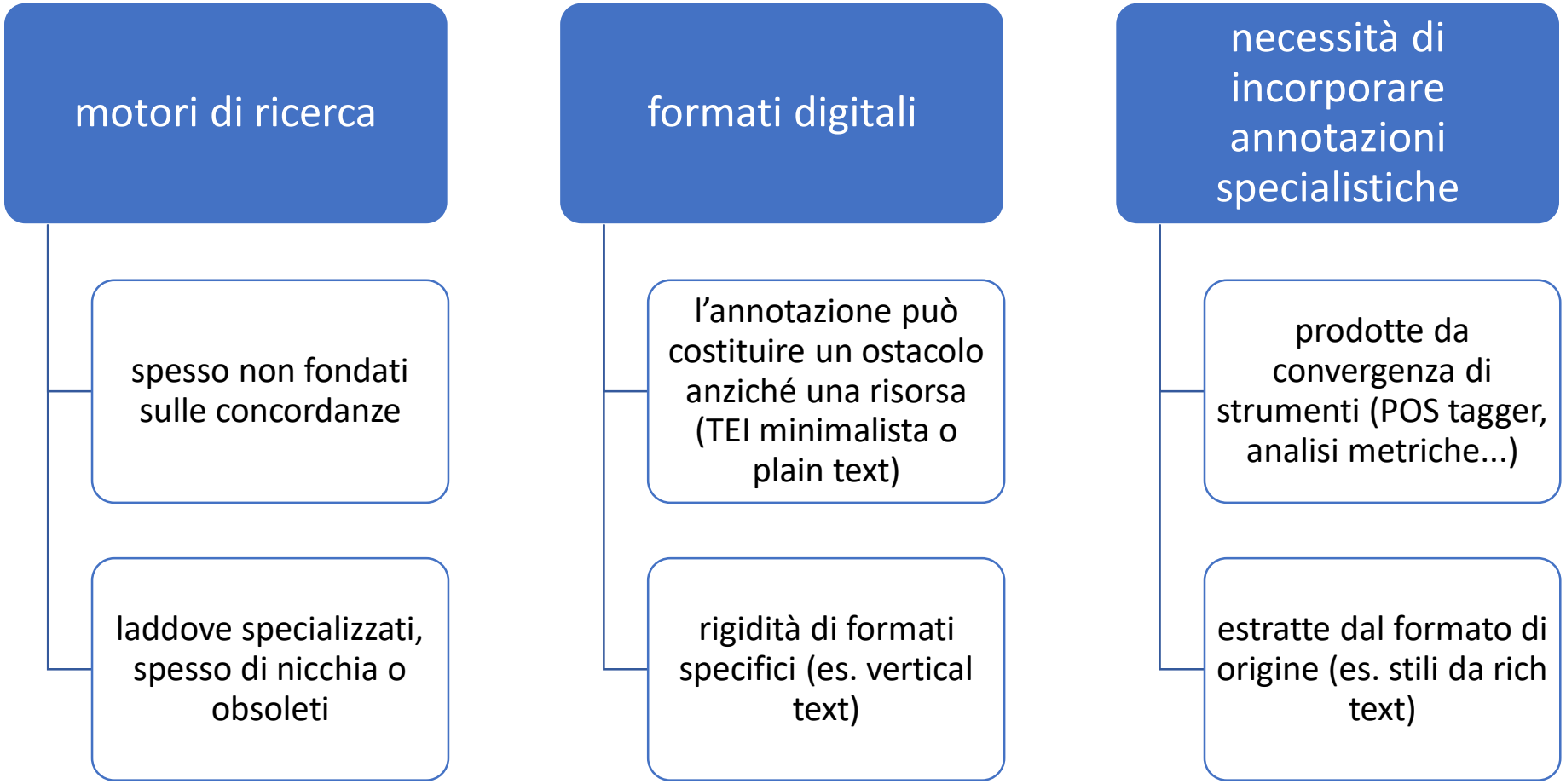
Anonimizzazione e pretrattamento

- anonimizzazione = marcatura
- lo scopo della marcatura può andare oltre le necessità pratiche abbracciando quelle dell'analisi

logica di anonimizzazione

- /d/ eufonica: garantire che ogni nome proprio casuale sia scelto con la stessa iniziale
- date: per analisi giuridica, casuali ma con conservazione della distanza relativa

Adattabilità degli strumenti ai corpora

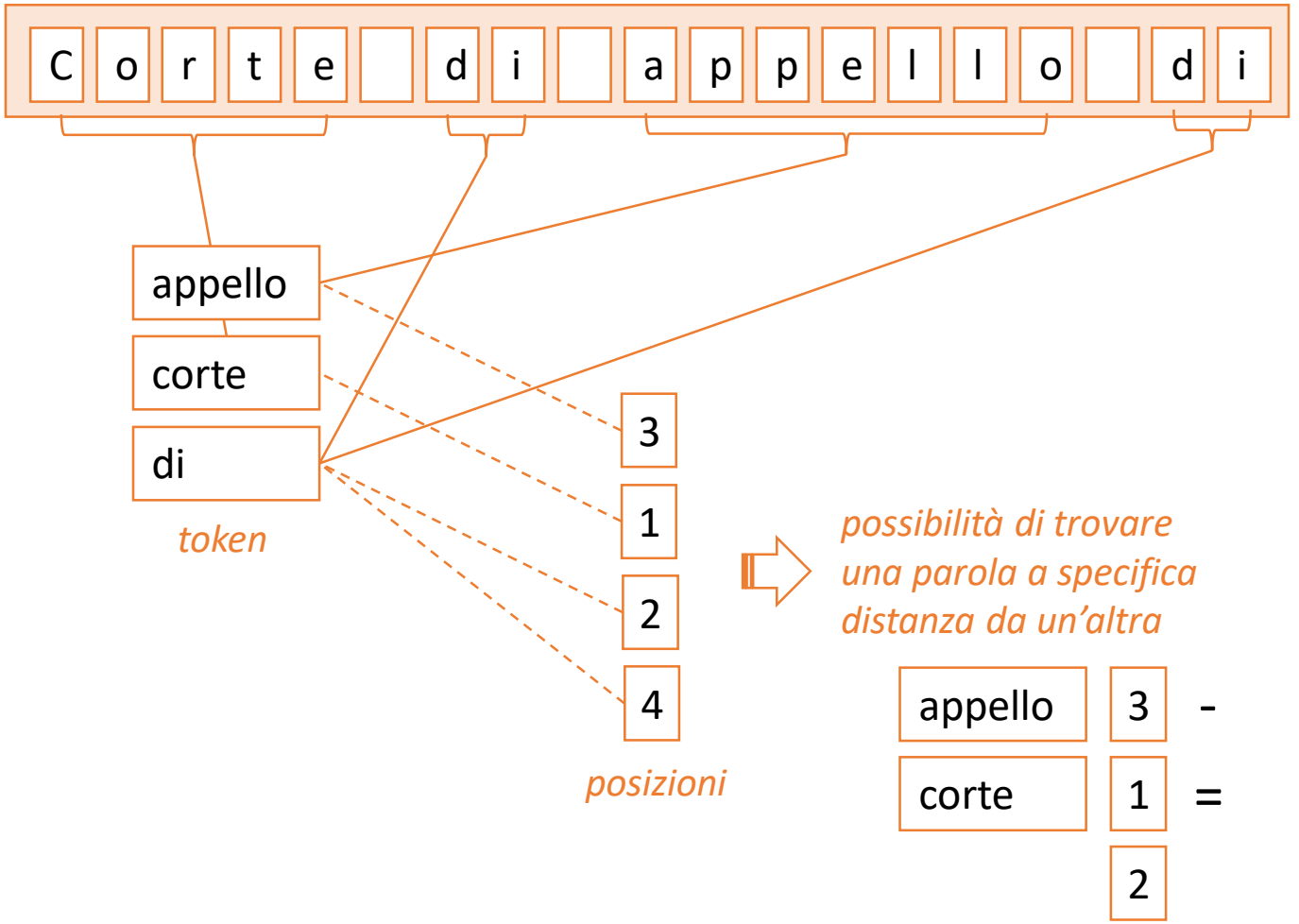


Motore di ricerca

https://github.com/vedph/pythia

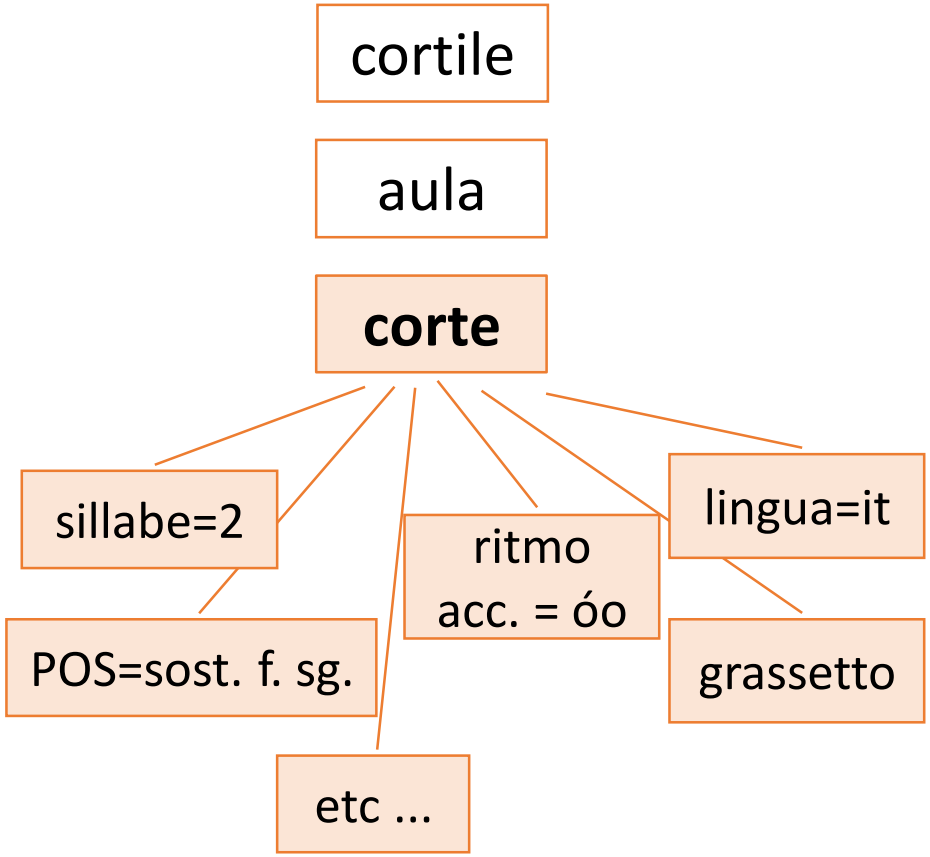
Approccio tradizionale

- documento come sequenza di caratteri
- estrazione di sotto-sequenze corrispondenti alle "parole" (*token*)
- eventuale aggiunta della posizione



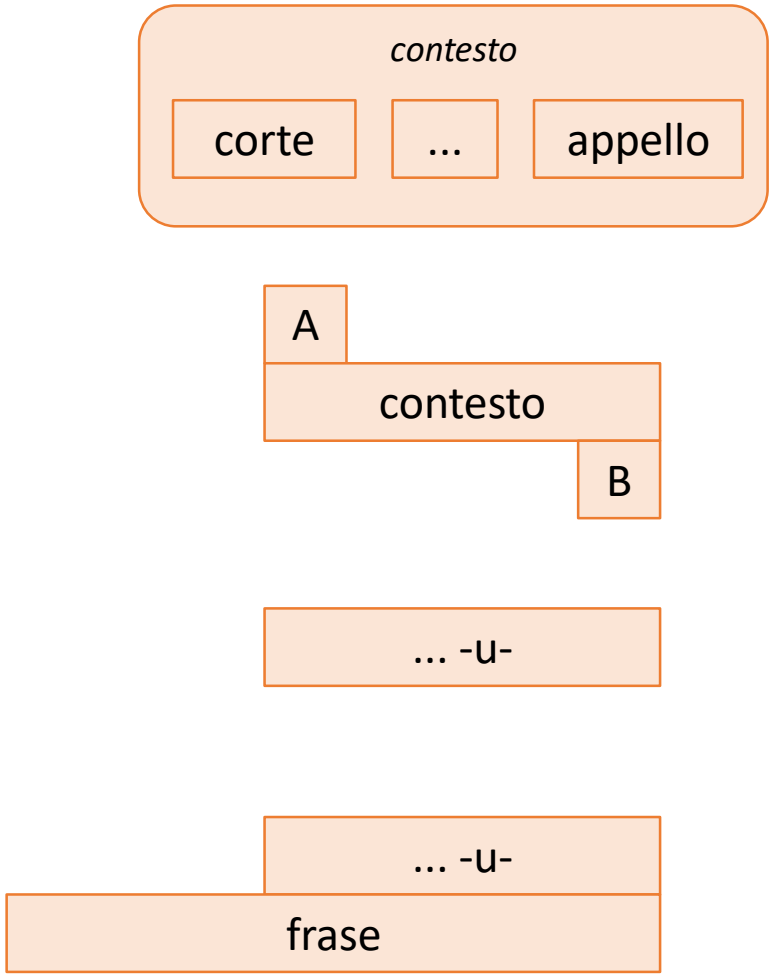
Limiti: parola

- possibile aggiungere più sequenze a uno stesso *token* (es. sinonimi)
- essenzialmente limitato a una sequenza di caratteri, cui qualsiasi altro metadato, per quanto eterogeneo, deve essere ridotto: es. numero di sillabe, POS, lingua, data, grassetto...
- specie in presenza di un gran numero di annotazioni specialistiche la quantità dei metadati è assai superiore ai caratteri del testo



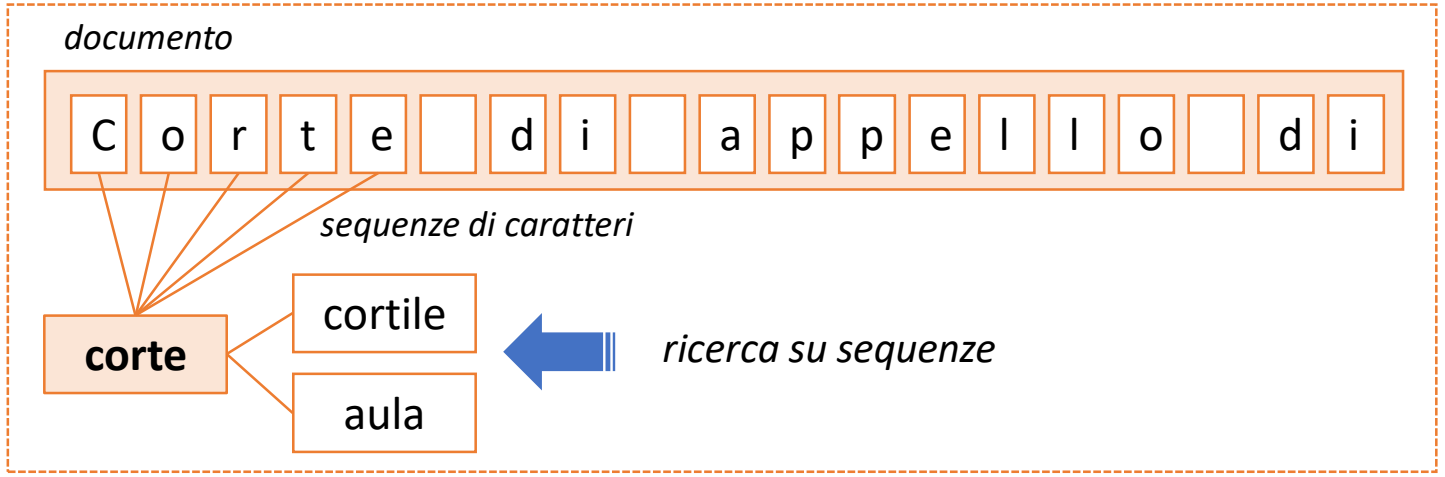
Limiti: contesto

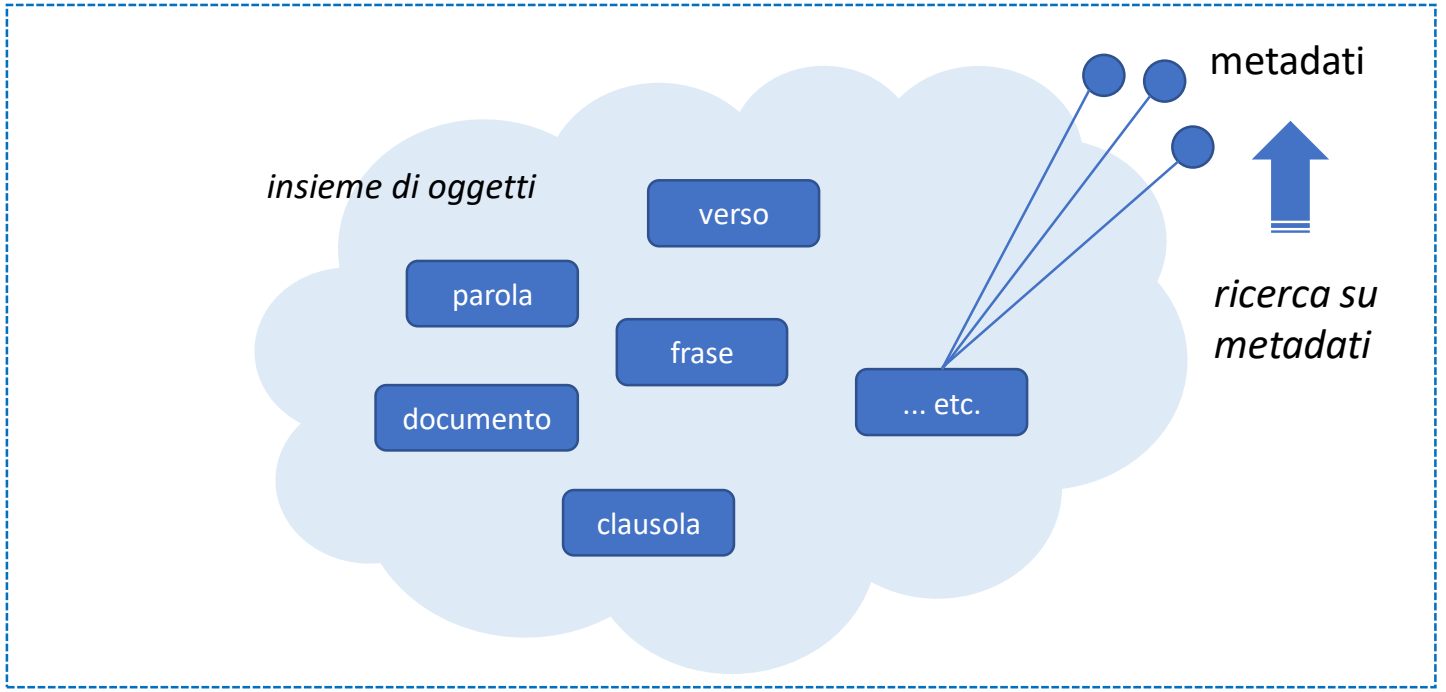
- trovare una parola prima/dopo un'altra
- trovare due parole nello stesso contesto, dove il contesto è una qualsiasi struttura testuale: frase, verso, strofe, paragrafo, clausola ritmica...
- trovare una parola in una certa posizione in un contesto (es. inizio o fine verso)
- trovare un contesto di un certo tipo senza specificare il contenuto: es. tutte le clausole ritmiche x
- trovare contesti in vario rapporto spaziale fra loro: es. clausola a fine di frase



Nuovo approccio

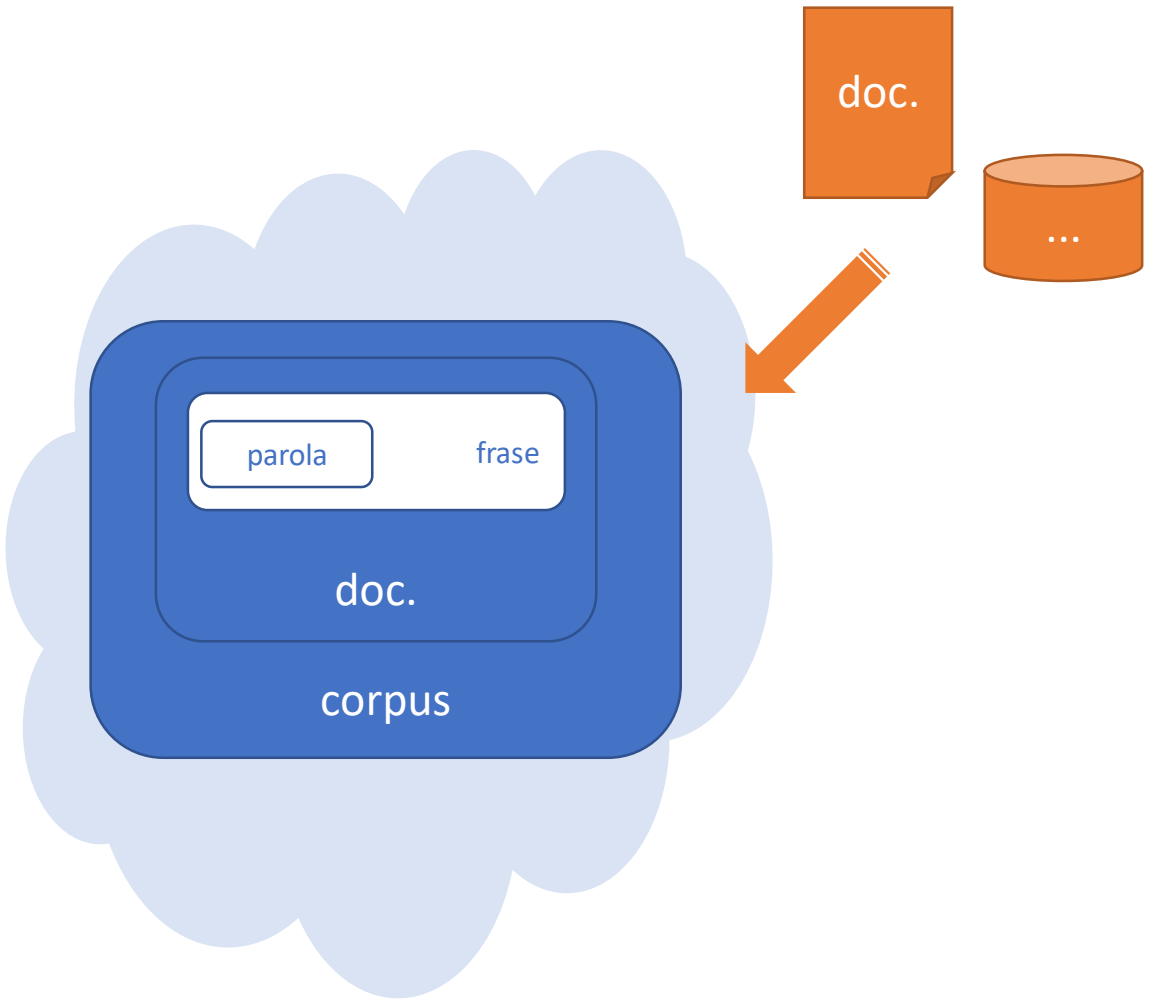
- approccio tradizionale:
  1. estrazione di sequenze di caratteri variamente filtrate e dotate di alternative
  2. ricerca di sequenze indicizzate
- nuovo approccio:
  1. estrazione di oggetti con metadati
  2. ricerca su metadati





Oggetti e metadati

- un oggetto può essere fuori (es. corpus) o dentro (es. parola, frase) un documento
- se dentro, fra i suoi metadati ha anche la posizione, espressa come un punto / segmento
- il repertorio di metadati per ogni oggetto è illimitato e aperto
- i documenti possono essere affiancati da altre fonti

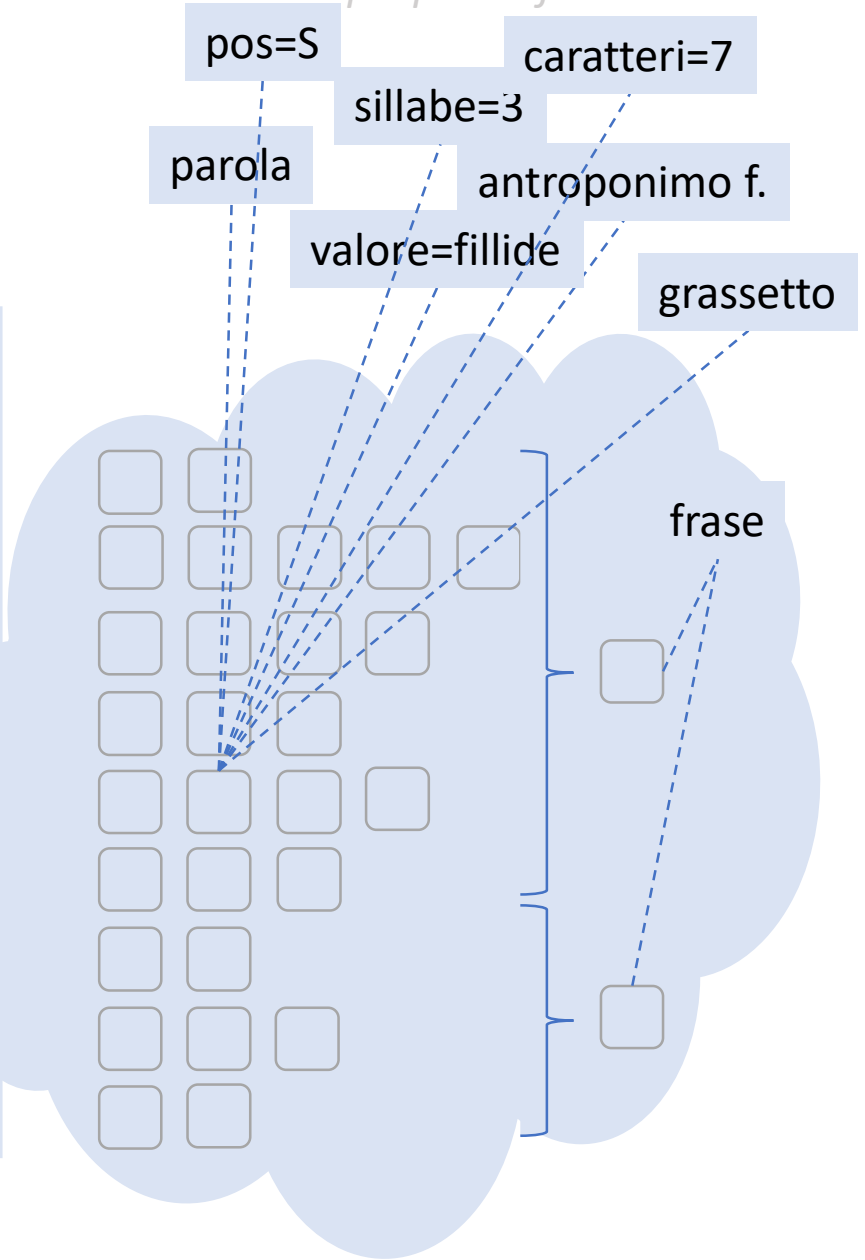


Smaterializzazione del testo

*testo: sequenza di **caratteri** con **formattazione** tipografica*

*insieme di **oggetti**, siano essi parole o strutture testuali*

Assorbente la specifica di cui sopra, in ogni caso il Papà potrà vedere **Antonio** e **Fillide** a fine settimana alterni. Il termine è prorogabile sino al lunedì mattina.
...



Oggetti e metadati

token ("parola")

struttura (es. frase, verso...)

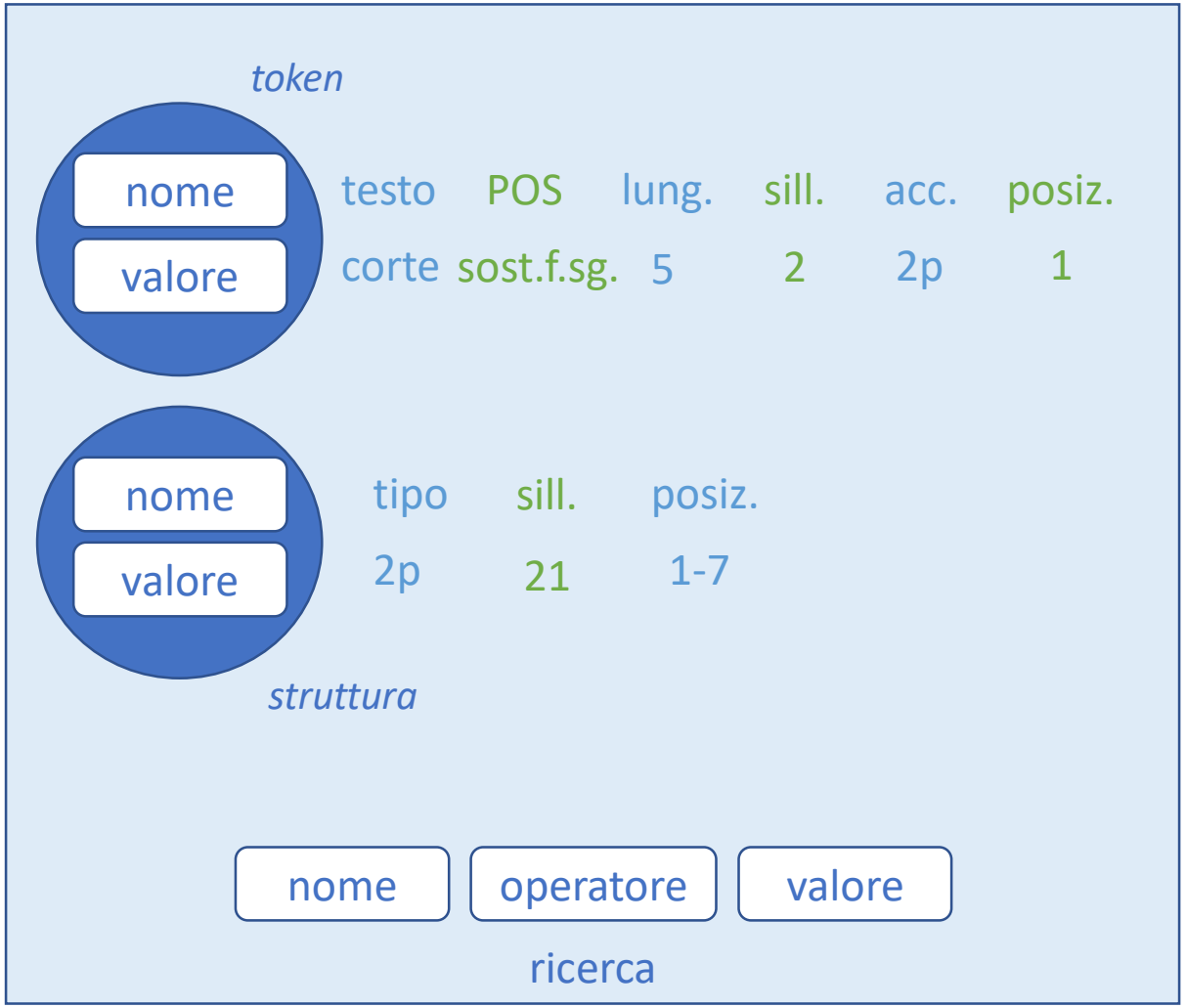
documento

corpus

con posizione

senza posizione

oggetti



Operatori

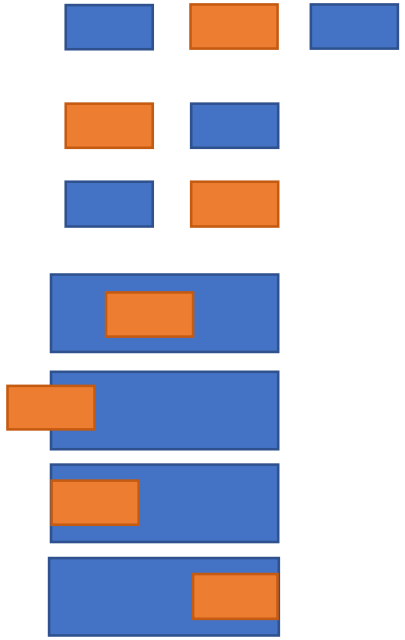
n = distanza minima
m = distanza massimo
s = nome della struttura di contesto
ns = distanza minima dall'inizio struttura
ms = distanza massima dall'inizio struttura
ne = distanza minimima dalla fine struttura
me = distanza massimo dalla fine struttura

Base

- = uguale a
- <> non uguale a
- *= contiene
- ^= inizia per
- $= finisce per
- ?= wildcard ? e *
- ~= espressione regolare
- %= fuzzy matching (con soglia)
- numerici: == != < > <= >=

Posizionali

- NEAR(n,m,s)
- BEFORE(n,m,s)
- AFTER(n,m,s)
- INSIDE(ns,ms,ne,me,s)
- OVERLAPS(n,m,s)
- LALIGN(n,m,s)
- RALIGN(n,m,s)



Esempio

- documento e metadati (da header)
- token e metadati (da testo + analizzatore prosodico + POS tagger)
- strutture e metadati (da tag + analisi algoritmiche)

Esempio: query

```
<lg type="eleg" n="3">
<l n="5" type="h">credo, sic mater, sic liber avunculus eius</l>
<l n="6" type="p">sic maternus avus dixerat atque avia.</l>
```

@[author="catullus"] AND ([category="poetry"] OR [datevalue<"0"]);
[value="sic"] BEFORE(m=0,s=l) [value="mater"]

documento: autore=catullus

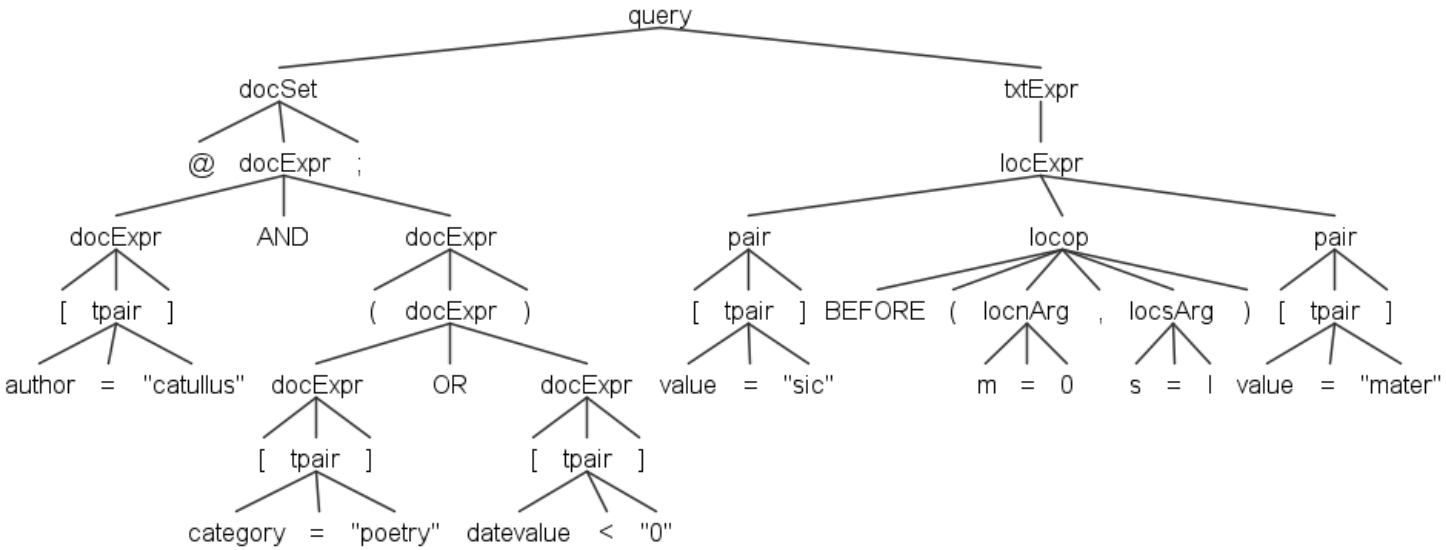
documento: categoria=poetry

documento: valore data < 0 (=a.C.)

token: valore=sic

token: valore=mater

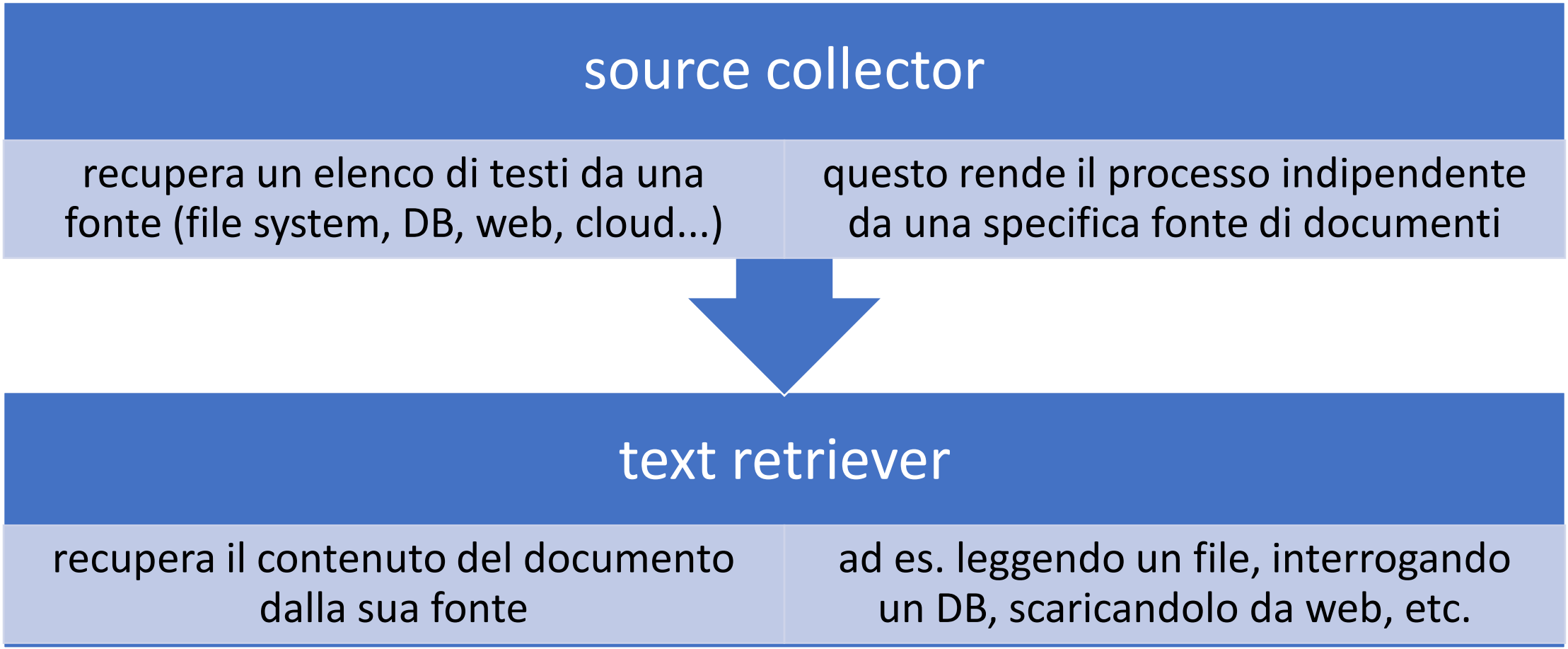
= trova tutte le occorrenze di “sic” immediatamente precedente “mater”, ove sia “sic” che “mater” appartengano allo stesso verso, e solo nei documenti scritti da Catullo e o di genere poetico o datati avanti Cristo.



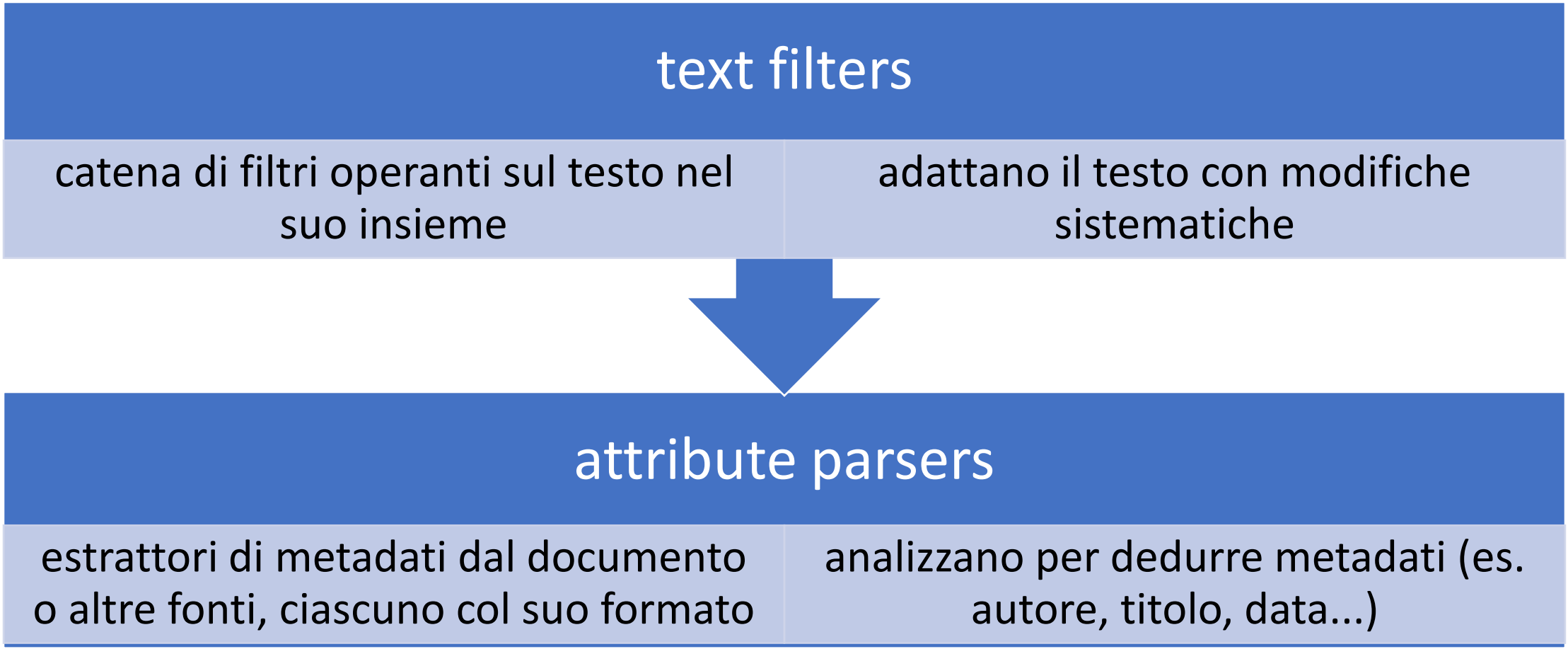
Indicizzazione modulare

Adattabilità del sistema

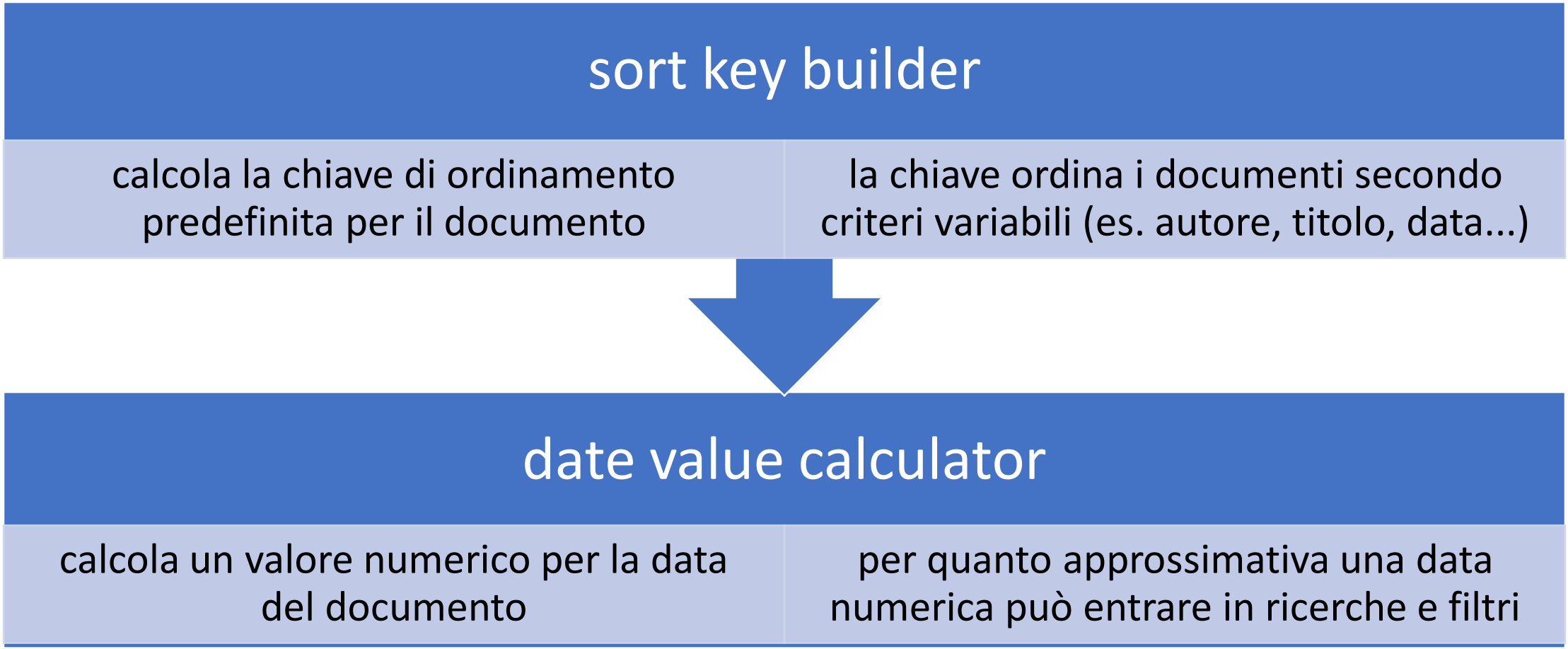
Recupero dei documenti



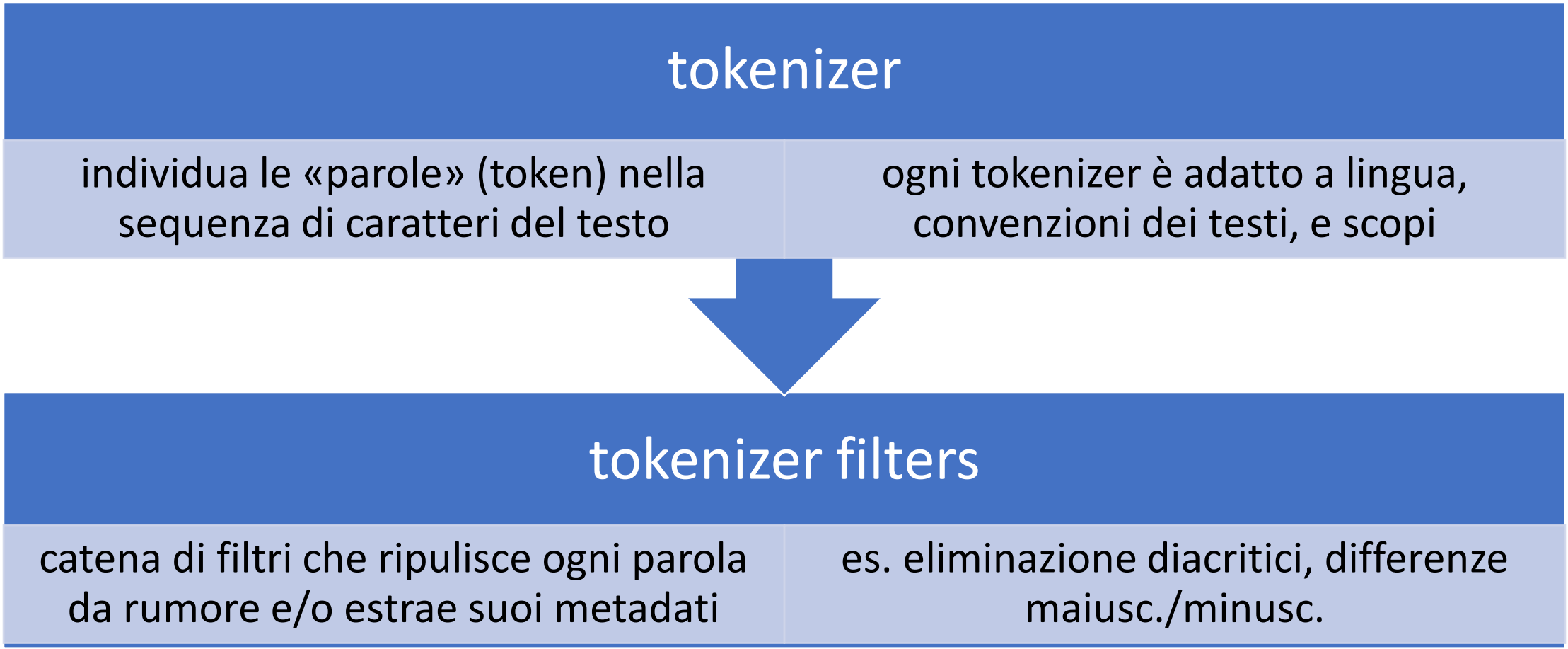
Pretrattamento del documento



Calcolo di metadati del documento



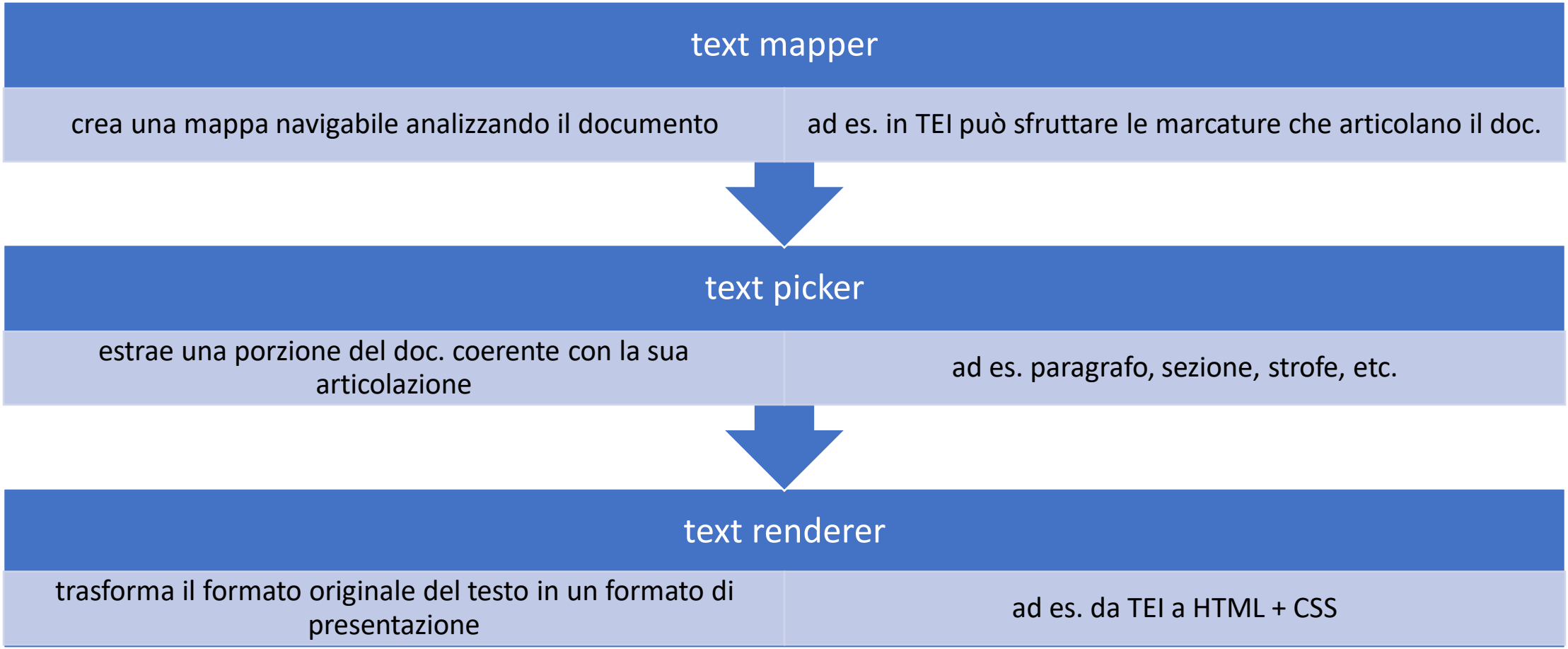
Estrazione e filtro delle parole



Estrazione delle strutture testuali

structure parsers		
analizzano il documento per estrarne strutture testuali di qs tipo	strutture possono essere desunte da marcature (l) o per via algoritmica (es. frase), e sono spesso sovrapposte	alcune strutture sono individuate solo per attribuire ulteriori metadati alle parole che contengono: es. persName

Ambiente di lettura



valore=abbandonato

trova gli oggetti con valore=abbandonato

ogni oggetto ha una posizione, da cui si ricava il contesto immediato

[value="abbandonato"]

search history

	author	title	5	4	3	2	1	value	1	2	3	4	5
	-	*civ-le-app-cit342-201202_01*	rinvenimento	al	proprietario	del	mezzo	**abbandonato**	e	ferma	restando	la	necessita'
	-	*civ-le-app-cit342-201202_01*	che	rinvengono	il	veicolo	presuntivamente	**abbandonato**	o	incidentato	devono	redigere	il
	-	*civ-le-app-cit342-201202_01*	raccolta	il	veicolo	si	intende	**abbandonato**	ex	art	r	gli	agenti
	-	*civ-le-app-cit342-201202_01*	fine	di	rendere	il	veicolo	**abbandonato**	un	rifiuto	da	eliminare	solo
	-	*civ-le-app-cit342-201202_01*	solo	dopo	che	il	veicolo	**abbandonato**	e	non	reclamato	dal	proprietario

Items per page: 20 1 – 5 of 5

dalla posizione si risale a una porzione di testo o al testo intero

act

TRIBUNALE DI LOMAGNA
Atto di Appello
DITTA SOLANGE DI DESIMINE EUSANIO, titolare della depositeri...
- APPELLANTE
CONTRO
CRATONE , con l'avv.avvocato Ezechiela Norcini
- APPELLATO
La Ditta Solange di Desimine Eusanio, come sopra domiciliata...
PROPONE

- - civ-le-app-cit342-201202_01

Prima di procedere alla demolizione del mezzo, dunque, la normativa stabilisce chiaramente che il gestore del centro di raccolta deve procedere alla cancellazione (o radiazione) del veicolo affidato dal PRA, trascorsi sessanta giorni dalla notificazione del verbale di rinvenimento al proprietario del mezzo abbandonato e *ferma restando la necessita' di comunicazione da parte degli organi di polizia di tutti i dati necessari per la presentazione, da parte del centro di raccolta, della formalita' di radiazione.*

il testo (qui TEI) viene convertito in HTML

l'atto è leggibile per paragrafi o per intero navigando la sua mappa

Daniele Fusi

daniele.fusi@unive.it

github.com/vedph

https://github.com/vedph/pythia

https://github.com/vedph/pythia-app

[ve]dph

Venice Centre for Digital and Public Humanities