

Università degli Studi di Torino
Dipartimento di Informatica
Corso di Laurea Magistrale in Informatica



ESTRAZIONE DI CONOSCENZA PER UNA STORIOGRAFIA DIGITALE

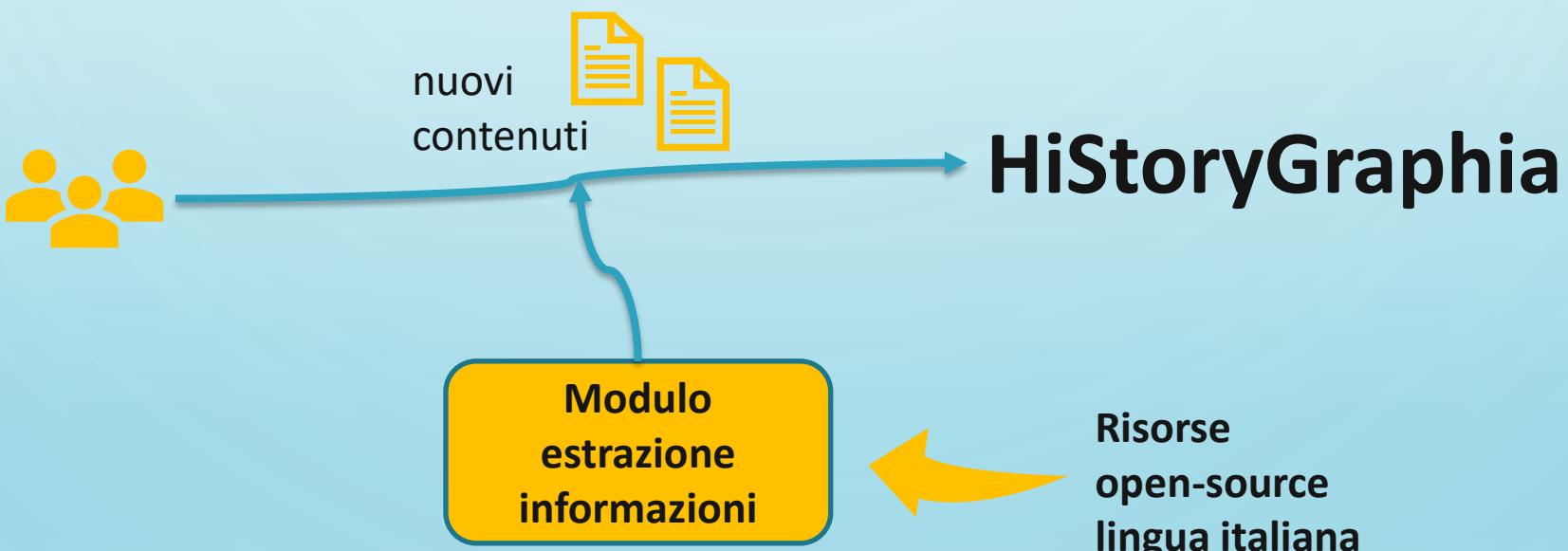
Candidato:
Biondi Giuseppe

Relatore:
Vincenzo Lombardo

Anno Accademico 2020 - 2021

INTRODUZIONE

HiStoryGraphia è un progetto di ricerca per la **raccolta di contenuti storici**.



CONTENUTI

- **Estrazione di informazioni:**
 - Wrapping
 - Estrazione di entità
 - Estrazione di relazioni
- Dominio storico
- HiStoryGraphia
- Il modulo Storytelling2Knowledge
- Sperimentazione e risultati

ESTRAZIONE DI INFORMAZIONI

- WRAPPING
- ESTRAZIONE DI ENTITÀ
- ESTRAZIONE DI RELAZIONI

WRAPPING

Una procedura per estrarre un particolare contenuto da una risorsa dotata di struttura

WRAPPING

Articolo: San Bernardo Castelletto Stura

[Diocesi]: Diocesi di Asti; dal 1430 Diocesi di Mondovì e in seguito alla riorganizzazione post-napoleonica (1817) passa sotto la Diocesi di Cuneo [fonte: L. Berra, *Riordinamento delle Diocesi di Mondovì, Saluzzo, Alba e Fossano ed erezione della Diocesi di Cuneo nel 1817*, in "Bollettino della Società per gli Studi Storici, Archeologici e Artistici della Provincia di Cuneo", 36, 1955, p. 51]

[Dipendenze]: Dal 1430 il territorio di Castelletto è assorbito in quello di più vasta pertinenza della "villanova" di Cuneo [fonte: R. Comba, *Due resoconti inediti della castellania di Cuneo (1388-1409)*, in "Bollettino della Società per gli Studi Storici, Archeologici e Artistici della Provincia di Cuneo", 67, 1972, pp. 32-33]. Nel 1619 Castelletto è infeudato ad Amedeo Ponte di Scarnafigi; nel 1661 passa a Francesco Bartolomeo Sandri Trottì, marchese di Montanera, che nel 1668 lo vende a Giovanni Battista Lamberti, famiglia a cui apparterrà fino al periodo della conquista francese [fonte: M. Ristorto, *Castelletto Stura. Storia civile e religiosa*, Cuneo 1977, pp. 56-73; G. Comino, *Castelletto Stura*, 1998 <https://www.archiviocasalis.it/localized-install/content/castelletto-stura>].

WRAPPING

Articolo: San Bernardo Castelletto Stura

Diocesi

[Diocesi] Diocesi di Asti; dal 1430 Diocesi di Mondovì e in seguito alla riorganizzazione post-napoleonica (1817) passa sotto la Diocesi di Cuneo [fonte: L. Berra, *Riordinamento delle Diocesi di Mondovì, Saluzzo, Alba e Fossano ed erezione della Diocesi di Cuneo nel 1817*, in "Bollettino della Società per gli Studi Storici, Archeologici e Artistici della Provincia di Cuneo", 36, 1955, p. 51]

Dipendenze

[Dipendenze] Dal 1430 il territorio di Castelletto è assorbito in quello di più vasta pertinenza della "villanova" di Cuneo [fonte: R. Comba, *Due resoconti inediti della castellania di Cuneo (1388-1409)*, in "Bollettino della Società per gli Studi Storici, Archeologici e Artistici della Provincia di Cuneo". 67, 1972, pp. 32-33] Nel 1619 Castelletto è infeudato ad Amedeo Ponte di Scarnafigi; nel 1661 passa a Francesco Bartolomeo Sandri Trottì, marchese di Montanera, che nel 1668 lo vende a Giovanni Battista Lamberti, famiglia a cui apparterrà fino al periodo della conquista francese [fonte: M. Ristorto, *Castelletto Stura. Storia civile e religiosa*, Cuneo 1977, pp. 56-73; G. Comino, *Castelletto Stura*, 1998 <https://www.archiviocasalis.it/localized-install/content/castelletto-stura>].

Fonti

ESTRAZIONE DI ENTITÀ

TASKS:

- Named Entity Recognition
- Entity Linking

TECNICHE:

- Gazetteer

TASK

NAMED ENTITY RECOGNITION

...

I confratelli di Santa Croce commissionano (1658 - 1660) un ciclo di dodici tele a Lorenzo Gastaldi.

...

Il modello decorativo di riferimento è quello della Confraternita di Santa Croce a Cuneo, dove nel 1626 sono allestite le tele con i Miracoli della Vera Croce dipinti nel 1626 da Giulio e Giovanni Battista Bruno

...

TASK

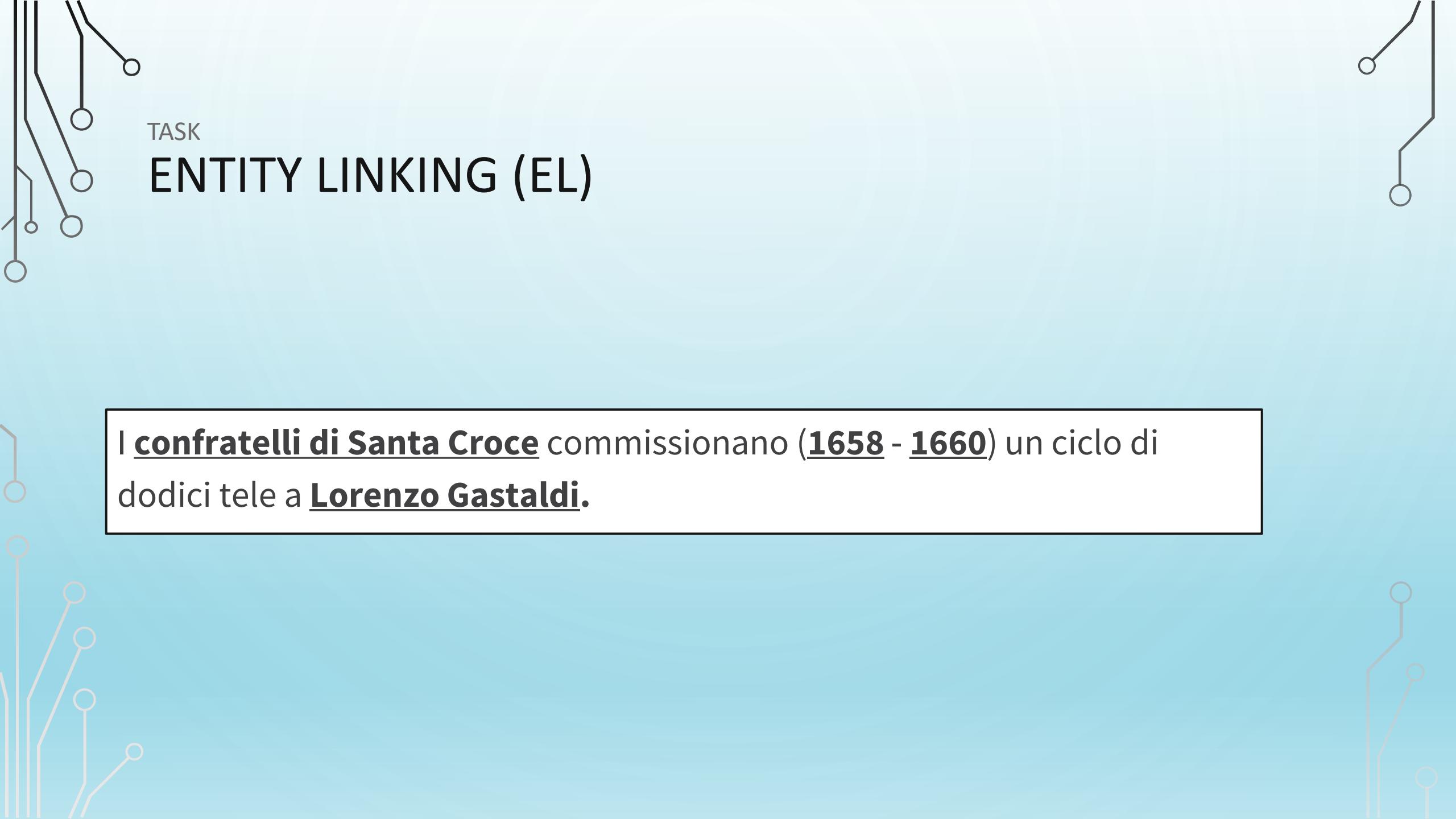
NAMED ENTITY RECOGNITION (NER)

...
I confratelli di Santa Croce commissionano
(1658 - 1660) un ciclo di dodici tele a Lorenzo
Gastaldi.

...
Il modello decorativo di riferimento è quello
della Confraternita di Santa Croce a Cuneo,
dove nel 1626 sono allestite le tele con i
Miracoli della Vera Croce dipinti nel 1626 da
Giulio e Giovanni Battista Bruno

classi

Persone
Località
Date
Organizzazioni
Misto



TASK

ENTITY LINKING (EL)

I **confratelli di Santa Croce** commissionano (**1658 - 1660**) un ciclo di dodici tele a **Lorenzo Gastaldi**.

TASK

ENTITY LINKING (EL)

<https://www.wikidata.org/wiki/Q81175539>

<https://www.wikidata.org/wiki/Q6991>

<https://www.wikidata.org/wiki/Q7001>

I confratelli di Santa Croce commissionano (1658 - 1660) un ciclo di dodici tele a Lorenzo Gastaldi.

The screenshot shows a Wikidata item page for Lorenzo Gastaldi (Q112111539). The page title is "Lorenzo Gastaldi". Below the title, it says "Italian painter (1625-1690)". There are links for "In more languages", "Gargano", "Language", "Label", "Description", and "Also in". The "Statements" section shows one statement: "instance of" with a value of "human".

<https://www.wikidata.org/wiki/Q112111539>

TECNICA

GAZETTEER

NER GAZETTEER

Utile per riconoscere classi specifiche del dominio

La prima data utile per ricostruire l'attività di Giovanni Mazzucco è il 1481, quando firma gli affreschi

EL GAZETTEER

Utile per riconoscere entità ambigue ma che in un determinato dominio assumono un preciso significato.

La prima data utile per ricostruire l'attività di Giovanni Mazzucco è il 1481, quando firma gli affreschi

TECNICA

GAZETTEER

NER GAZETTEER

Utile per riconoscere classi specifiche del dominio

La prima data utile per ricostruire l'attività di
Giovanni Mazzucco è il 1481, quando firma
gli affreschi

Gazetteer per la classe Pittore:

- Giovanni Mazzucco
- Lorenzo Gastaldi
- Leonardo

EL GAZETTEER

Utile per riconoscere entità ambigue ma che in un determinato dominio assumono un preciso significato.

La prima data utile per ricostruire l'attività di
Giovanni Mazzucco è il 1481, quando firma gli
affreschi

TECNICA

GAZETTEER

NER GAZETTEER

Utile per riconoscere classi specifiche del dominio

La prima data utile per ricostruire l'attività di
Giovanni Mazzucco è il 1481, quando firma
gli affreschi

Gazetteer per la classe **Pittore**:

- Giovanni Mazzucco
- Lorenzo Gastaldi
- Leonardo

EL GAZETTEER

Utile per riconoscere entità ambigue ma che in un determinato dominio assumono un preciso significato.

La prima data utile per ricostruire l'attività di
Giovanni Mazzucco è il 1481, quando firma gli
affreschi

Gazetteer dei concetti presenti in HSG:

- *Giovanni Mazzucco* > https://www.hsg.org/giovanni_mazzucco
- *Entracque* > <https://www.hsg.org/entracque>

ESTRAZIONE DI RELAZIONI

TASKS:

- OPEN IE

TASK

OPEN IE

Durante il Seicento gli entracquesi vendono ovini e bovini ai macelli di Grasse, Nizza e Genova

TASK

OPEN IE

Durante il Seicento gli entracquesi vendono ovini e bovini ai macelli di Grasse, Nizza e Genova

(entracquesi; vendono; ovini)

(entracquesi; vendono; bovini)

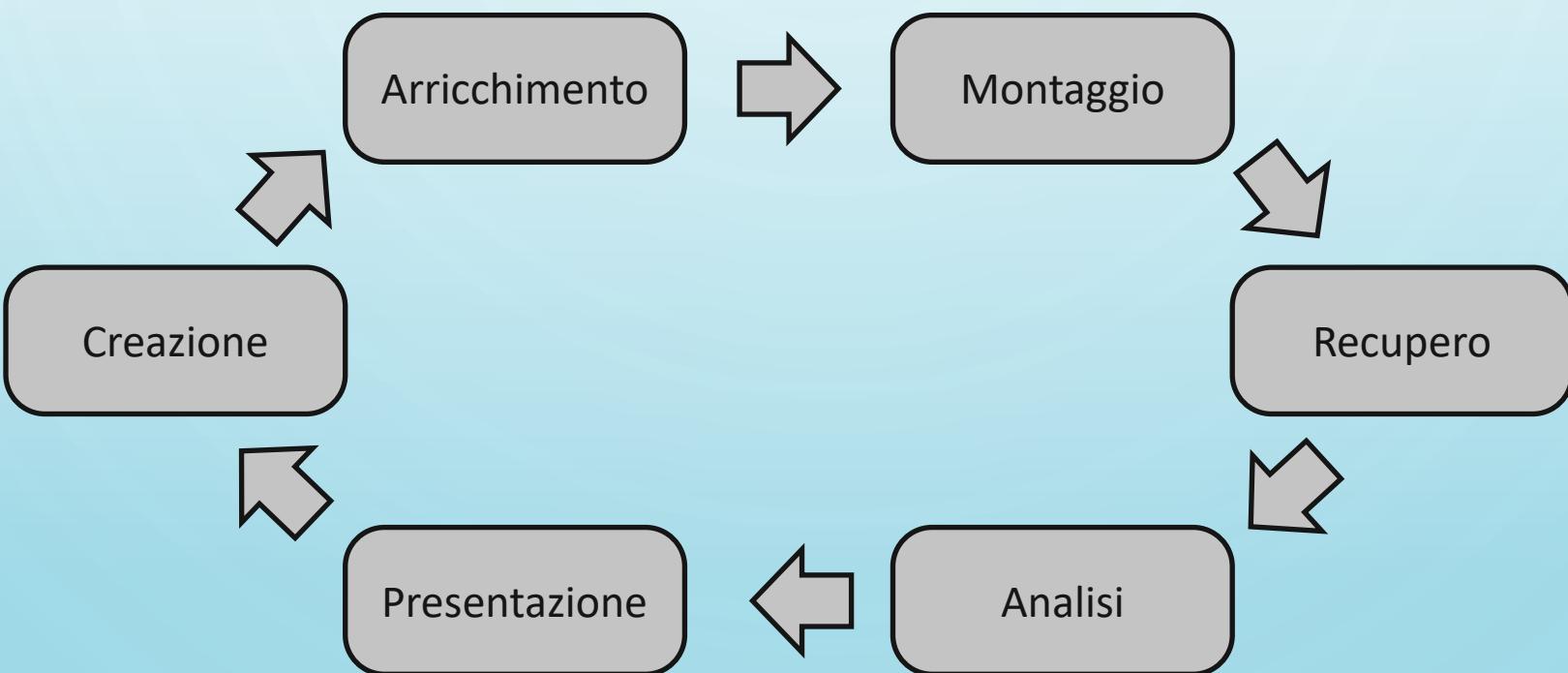
CONTENUTI

- Estrazione di informazioni:
 - Wrapping
 - Estrazione di entità
 - Estrazione di relazioni
- **Dominio storico**
 - HiStoryGraphia
 - Il modulo Storytelling2Knowledge
 - Sperimentazione e risultati

DOMINIO STORICO



CICLO DI VITA DELL'INFORMAZIONE STORICA



HISTORYGRAPHIA WORKFLOW

Estrazione conoscenza
(NER, EL, Gazetteers..)
e Annotazione con metadati

Estrazione Knowledge
Graph



Montaggio

Using KG to update HSG
database

Arricchimento

Creazione

Recupero

Analisi



Narrazioni create
da esperti



Presentazione

CONTENUTI

- Estrazione di informazioni:
 - Wrapping
 - Estrazione di entità
 - Estrazione di relazioni
- Dominio storico
- **HiStoryGraphia**
- Il modulo Storytelling2Knowledge
- Sperimentazione e risultati

HISTORYGRAPHIA

CONTESTO DEL PROGETTO

HISTORYGRAPHIA

Progetto di ricerca dell'Università degli studi di Torino,
finalizzato alla raccolta di contenuti storici per:

- Preservarli nel tempo
- Diffonderli (ricerca, turismo)
- Interconnetterli

Tramite partecipazione attiva dell'utente

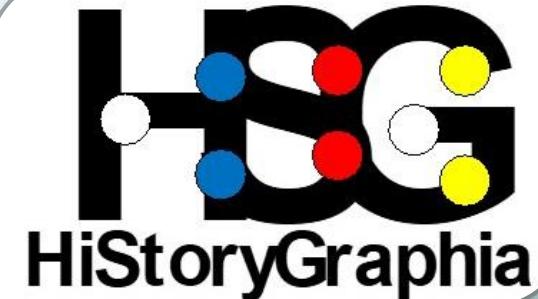
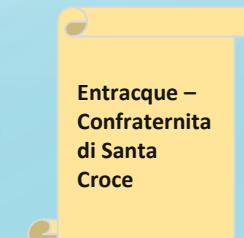


Utente

Può effettuare una
ricerca

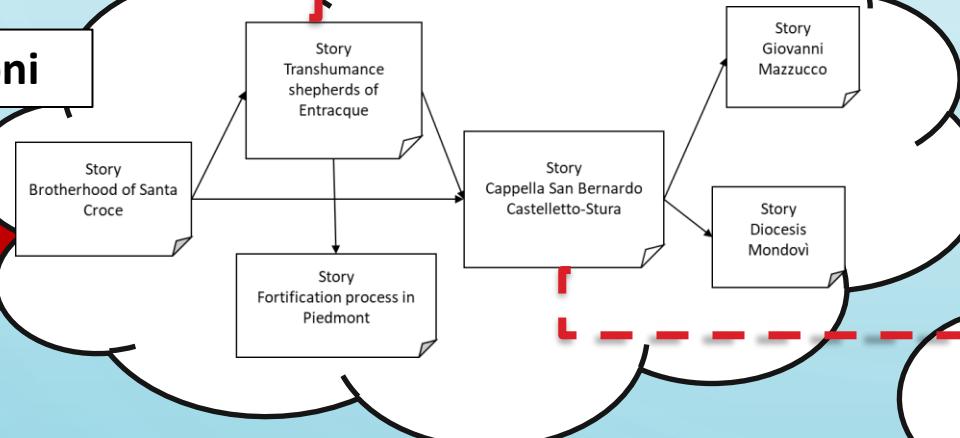


Può inserire una
nuova narrazione

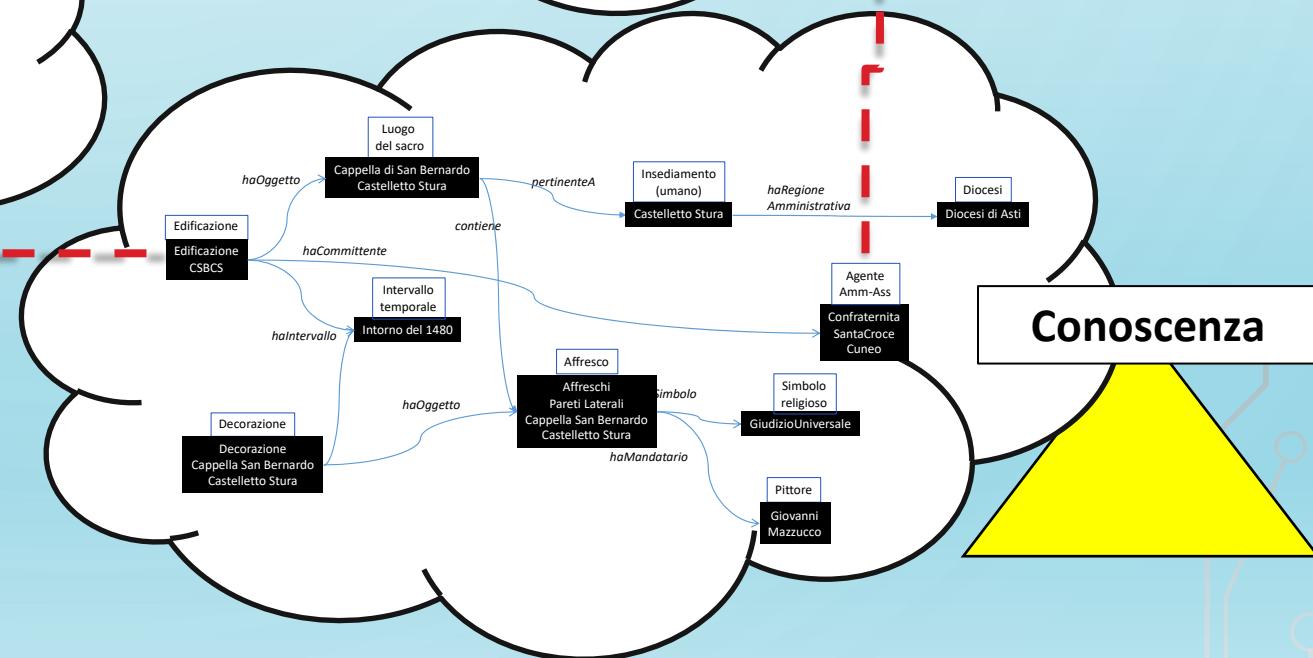


La piattaforma si divide in tre livelli di rappresentazione

Narrazioni

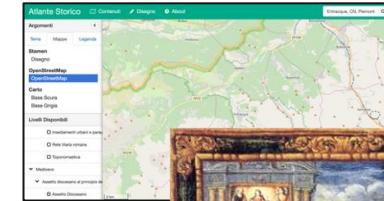


Conoscenza

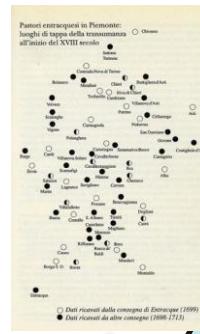


Fonti

Mappe



Schermata



Esempio: Confraternita di Santa Croce a Entracque

Narrazione (+illustrazione + riferimento alla mappa di Entracque) connessa alla narrazione del pittore Gastaldi



... “La chiesa della confraternita viene edificata nel 1538, segno dell'affermazione di un notabilato che controlla la vita economica e sociale della comunità e sviluppatosi...” ...



Fonte



Entracque: una comunità alpina tra Medioevo ed Età moderna [Volume 12 di Storia e storiografia](#), [Rinaldo Comba, Mario Cordero](#). Società per gli studi storici, archeologici ed artistici della provincia di Cuneo, 1997

Grafo di conoscenza

Erection
Erezione

Erezione Chiesa della Confraternita di Santa Croce a Entracque

Place of sacred things
Luogo sacro

Chiesa della Confraternita di Santa Croce a Entracque

Time interval
Intervallo temporale

1538

has_object

has_time_interval

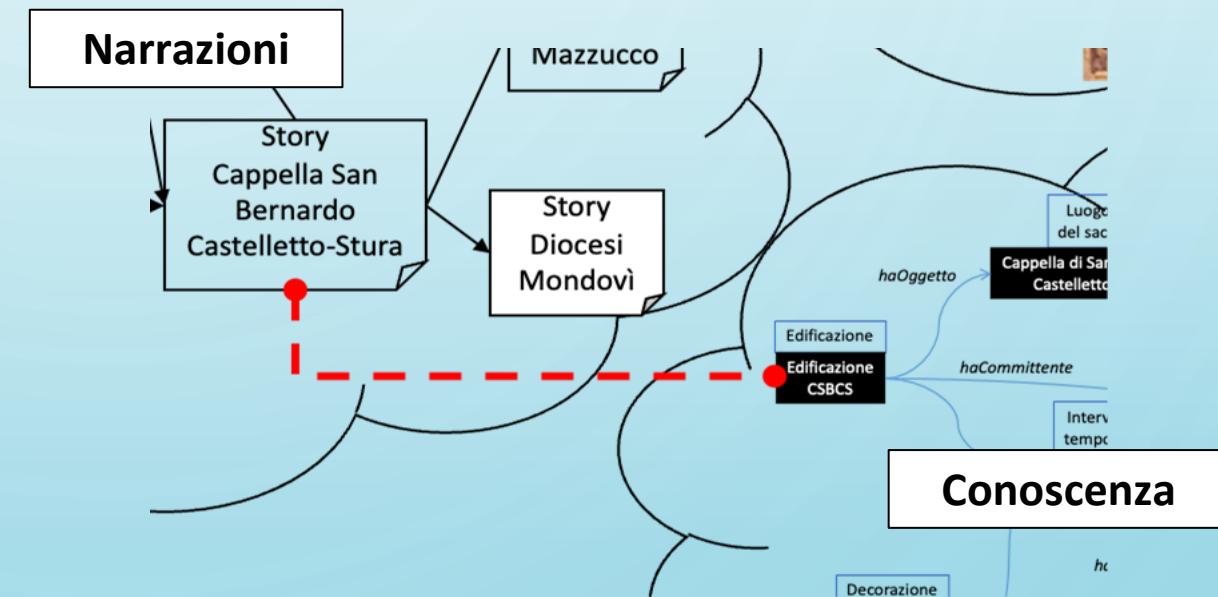
PROBLEMA DI RICERCA

Goal:

Automatizzare il processo di estrazione della conoscenza

Output:

Dato un testo, si vuole estrarre un grafo di conoscenza con le principali entità e relazioni



ESEMPIO

CAPPELLA DI SAN BERNARDO

...

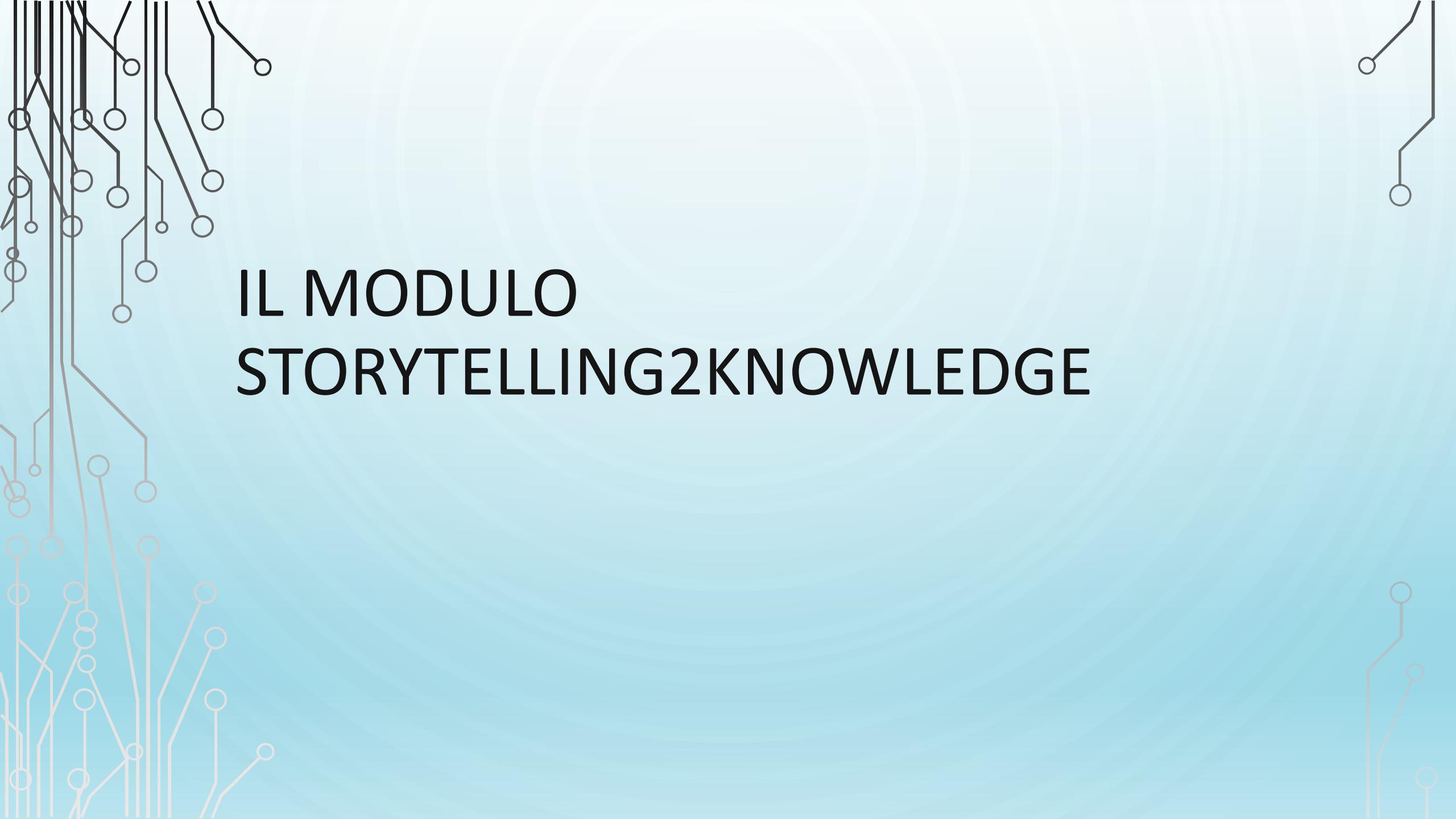
[Committenza] : L'**edificazione della cappella e la sua decorazione** ad affresco sono realizzati su **commissione del parroco del paese** con la partecipazione di alcuni notabili (che identifichiamo nelle figure affrescate ai margini della lunetta con l'Incoronazione della Vergine, abbigliate in vesti moderne e più caratterizzate nella fisionomia rispetto agli altri personaggi), illustri **membri della confraternita di Santa Croce**, da poco istituita in paese come filiazione della più importante congregazione cuneese [fonte: la confraternita di Santa Croce di Cuneo concede un aiuto in denaro il 7 marzo 1473].

...



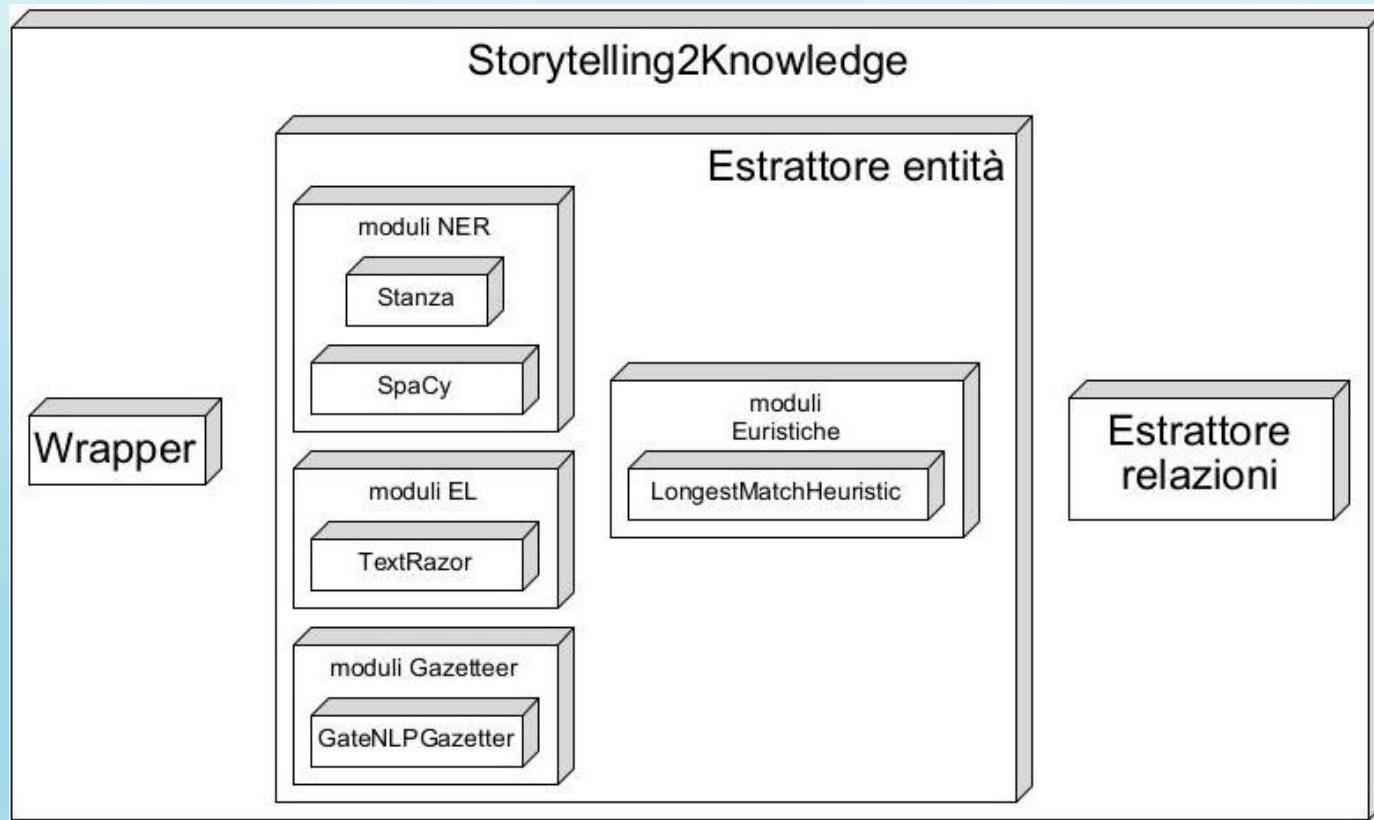
CONTENUTI

- HiStoryGraphia – contesto del progetto
- Estrazione di informazioni:
 - Wrapping
 - Estrazione di entità
 - Estrazione di relazioni
- Il modulo Storytelling2Knowledge
- Sperimentazione e risultati



IL MODULO STORYTELLING2KNOWLEDGE

ARCHITETTURA



ARCHITETTURA

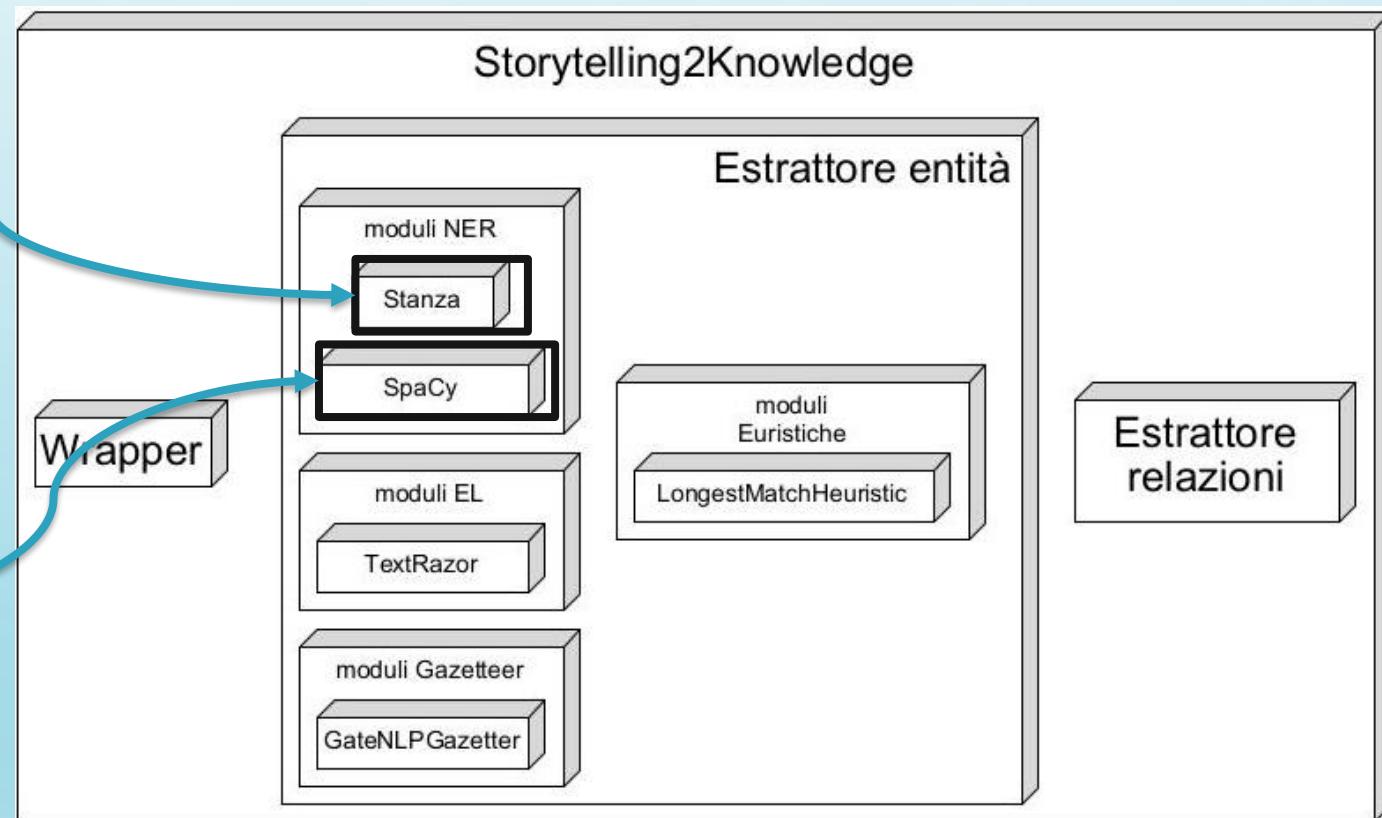
Moduli NER

spaCy

LOC – Località
PER – Persona
ORG – Organizzazioni
MISC – Misto

 **Stanza**

LOC – Località
PER – Persona
ORG – Organizzazioni

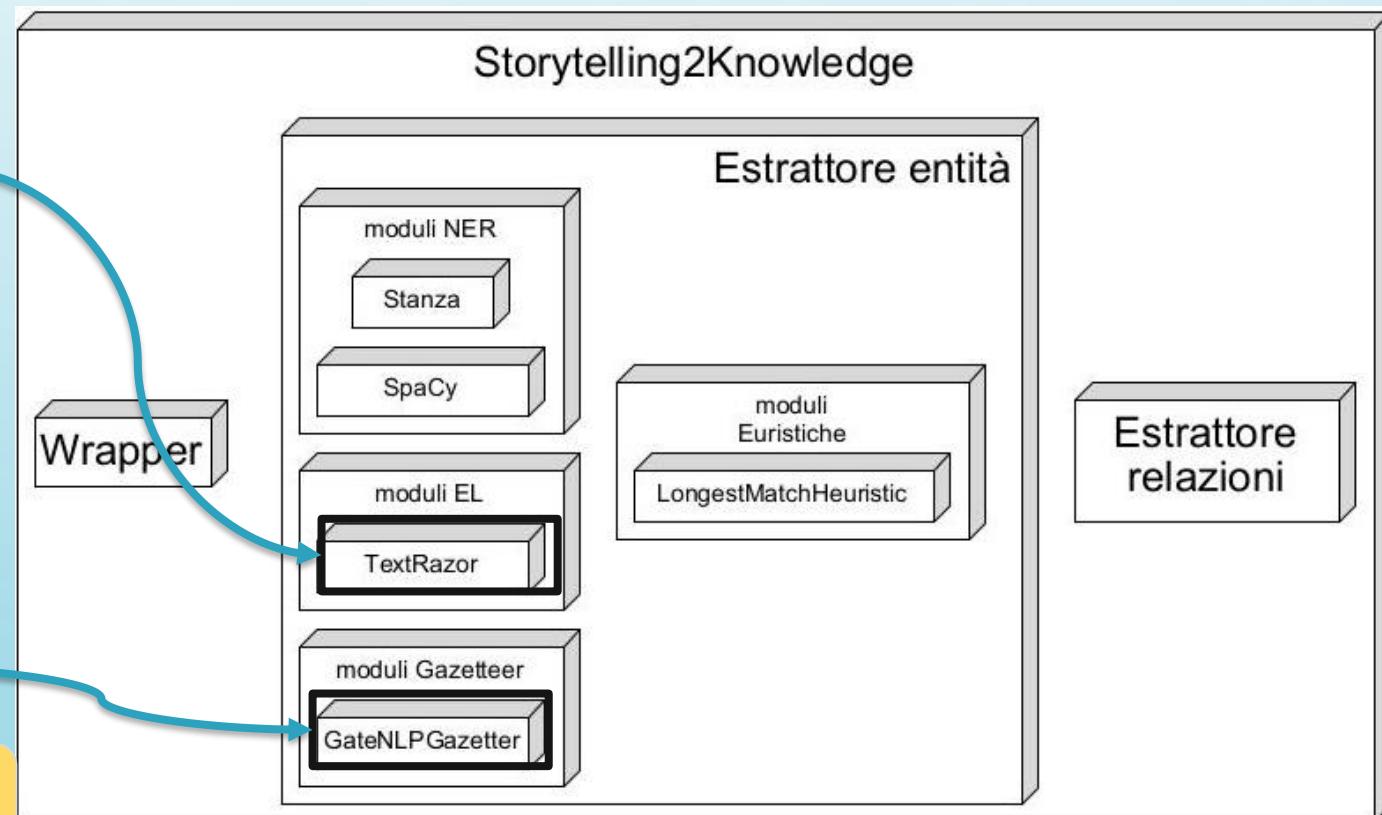


ARCHITETTURA

Moduli EL & Gazetteer

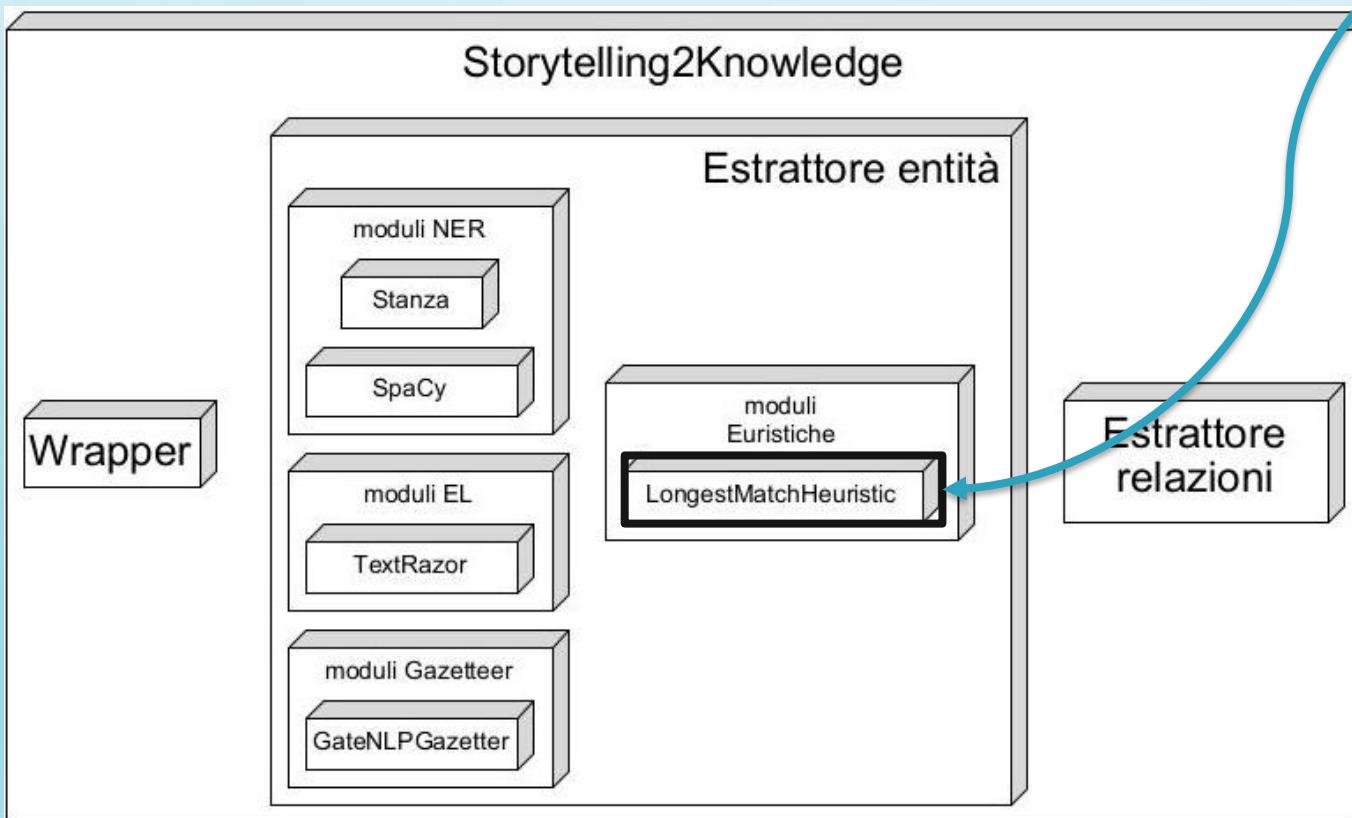


GATE NLP
Predisposto per EL e
NER



ARCHITETTURA

Modulo Euristiche



Tra le sovrapposizioni
sceglie la menzione più
lunga

Esempio di testo:

«Il **Ritratto di Dante** è
stato venduto per..»

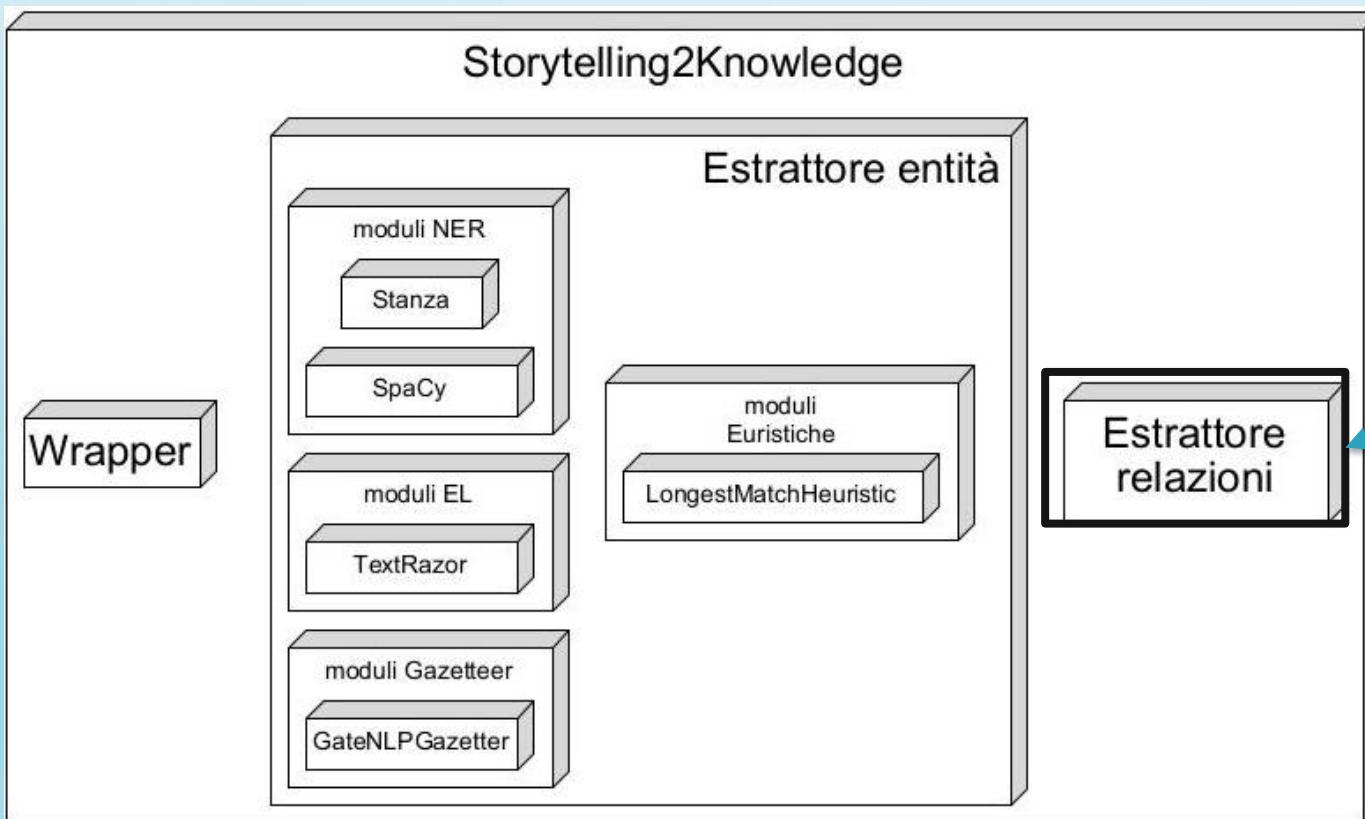
Estrattore 1:
«**Dante**» come
persona

Estrattore 2:
«**Ritratto di Dante**»
come dipinto

L'euristica sceglie:
Ritratto di Dante

ARCHITETTURA

Modulo Euristiche



spaCy

1) Albero a dipendenze
estratto da spaCy

2) Set di regole riadattate
dall'inglese all'italiano

Tratte dal
lavoro di
Schrading

3) Triplette
«soggetto, verbo, oggetto»
(entracquesi; vendono; bovini)

CONTENUTI

- HiStoryGraphia – contesto del progetto
- Estrazione di informazioni:
 - Wrapping
 - Estrazione di entità
 - Estrazione di relazioni
- Il modulo Storytelling2Knowledge
- Sperimentazione e risultati

SPERIMENTAZIONE E RISULTATI

MainWindow

File

Lista Moduli

spacy_module	ON
stanza_module	ON
text_razor_module	OFF

Scegliere una euristica

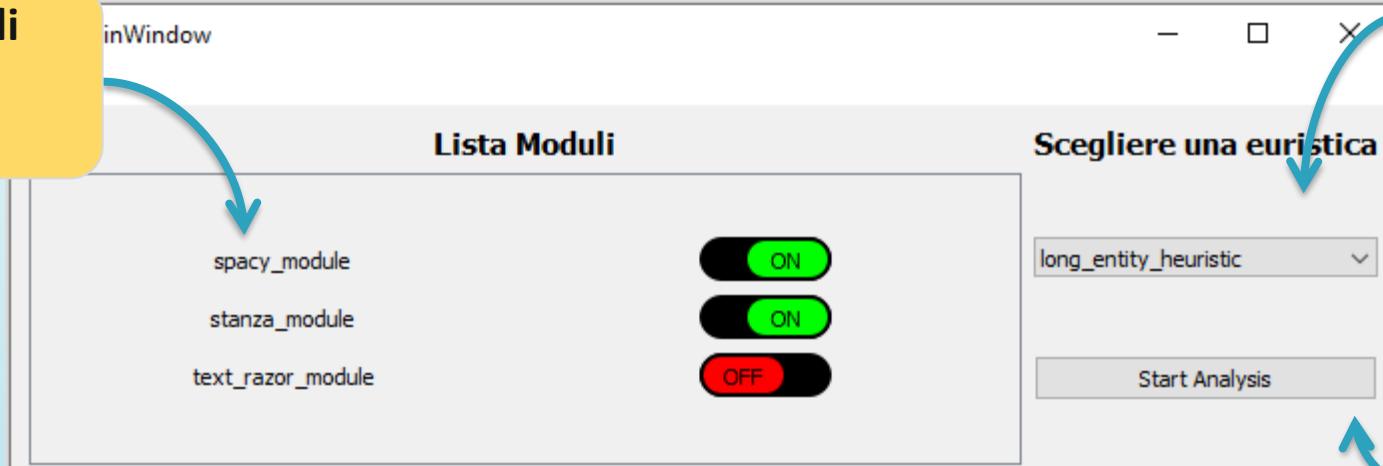
long_entity_heuristic

Start Analysis

Risultati

	NER Type	Precision	Recall	F1 score	match type	numbers	values
31	PER	0.043	0.667	0.082	exact match	tp: 2,fp: 44, fn:1	Show
32	PER	0.043	0.667	0.082	partial match	tp: 2,fp: 44, fn:1	Show
33	LOC	0.129	0.241	0.168	exact match	tp: 13,fp: 88, fn:41	Show
34	LOC	0.495	0.926	0.645	partial match	tp: 50,fp: 51, fn:4	Show
35	ORG	0	0	0	exact match		Show
36	ORG	0	0	0	partial match		Show
37	MISC	0.269	0.125	0.171	exact match	tp: 7,fp: 19, fn:49	Show
38	MISC	0.308	0.143	0.195	partial match	tp: 8,fp: 18, fn:48	Show
39	/	0.131	0.195	0.157	exact match	tp: 23,fp: 152, fn:95	Show
40	/	0.577	0.856	0.689	partial match	tp: 101,fp: 74, fn:17	Show

Scelgo quali estrattori attivare



Scegliere una euristica

Selezioni l'euristica

long_entity_heuristic

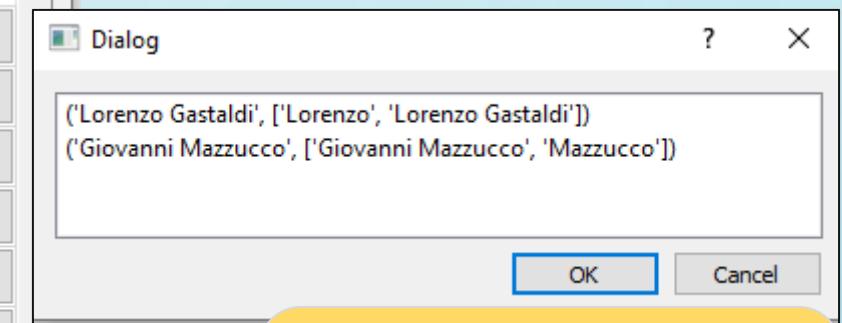
Start Analysis

Avvio test

TP: veri positivi
FP: falsi positivi
FN: falsi negativi

**Visualizza
dettagli**

	NER Type	Precision	Recall	F1 score	match type	numbers	values
31	PER	0.043	0.667	0.082	exact match	tp: 2,fp: 44, fn:1	Show
32	PER	0.043	0.667	0.082	partial match	tp: 2,fp: 44, fn:1	Show
33	LOC	0.129	0.241	0.168	exact match	tp: 13,fp: 88, fn:41	Show
34	LOC	0.495	0.926	0.645	partial match	tp: 50,fp: 51, fn:4	Show
35	ORG	0	0	0	exact match		Show
36	ORG	0	0	0	partial match		Show
37	MISC	0.269	0.125	0.171	exact match	tp: 7,fp: 19, fn:49	Show
38	MISC	0.308	0.143	0.195	partial match	tp: 8,fp: 18, fn:48	Show
39	/	0.131	0.195	0.157	exact match	tp: 23,fp: 152, fn:95	Show
40	/	0.577	0.856	0.689	partial match	tp: 101,fp: 74, fn:17	Show



**Lorenzo Gastaldi fa
match parziale con:
Lorenzo
Lorenzo Gastaldi**

Risultati per la classe Persona (PER)

PRECISIONE	RECALL	F1	MATCH	SPACY	STANZA	RAZOR	EURISTICA
0.048	0.667	0.089	exact match	on	on	on	long
0.048	0.667	0.089	partial match	on	on	on	long
0.033	0.667	0.063	exact match	on	on	on/off	/
0.033	0.667	0.063	partial match	on	on	on/off	/
0.043	0.667	0.082	exact match	on	on	off	long
0.043	0.667	0.082	partial match	on	on	off	long
0.062	0.667	0.114	exact match	off	on	on	long
0.062	0.667	0.114	partial match	off	on	on	long
0.050	0.667	0.093	exact match	off	on	on/off	/
0.050	0.667	0.093	partial match	off	on	on/off	/
0.045	0.667	0.085	exact match	on	off	on	long
0.045	0.667	0.085	partial match	on	off	on	long
0.041	0.667	0.077	exact match	on	off	on/off	/
0.041	0.667	0.077	partial match	on	off	on/off	/

Risultati per la classe Persona (PER)

PRECISIONE	RECALL	F1	MATCH	SPACY	STANZA	RAZOR	EURISTICA
0.048	0.667	0.089	exact match	on	on	on	long
0.048	0.667	0.089	partial match	on	on	on	long
0.033	0.667	0.063	exact match	on	on	on/off	/
0.033	0.667	0.063	partial match	on	on	on/off	/
0.043	0.667						ng
0.043	0.667						ng
0.062	0.667						ng
0.062	0.667						ng
0.050	0.667						/
0.050	0.667						/
0.045	0.667						long
0.045	0.667	0.085	partial match	on	off	on	long
0.041	0.667	0.077	exact match	on	off	on/off	/
0.041	0.667	0.077	partial match	on	off	on/off	/

Perché?

- Alcune **persone non sono rilevanti**
«Giulio e Giovanni Battista Bruno»
- Titoli di **opere contenenti nomi propri**
«Immacolata con i santi Sebastiano, Giuseppe, Rocco e Carlo Borromeo»

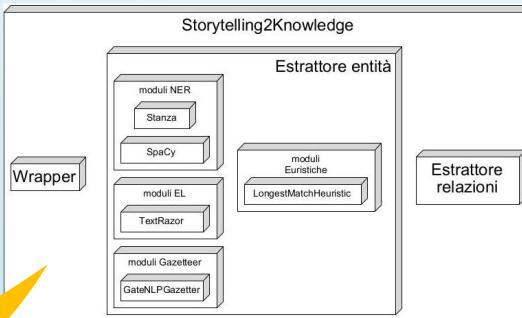
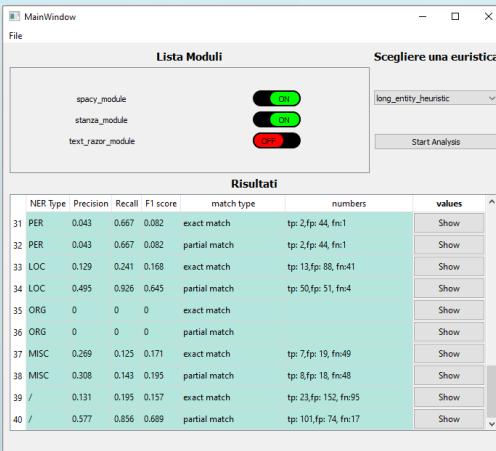
Risultati per la classe Località (LOC)

PRECISIONE	RECALL	F1	MATCH	SPACY	STANZA	RAZOR	EURISTICA
0.131	0.241	0.170	exact match	on	on	on	long
0.505	0.926	0.654	partial match	on	on	on	long
0.109	0.278	0.157	exact match	on	on	on/off	/
0.380	0.963	0.545	partial match	on	on	on/off	/
0.129	0.241	0.168	exact match	on	on	off	long
0.495	0.926	0.645	partial match	on	on	off	long
0.114	0.222	0.151	exact match	off	on	on	long
0.467	0.907	0.616	partial match	off	on	on	long
0.109	0.222	0.146	exact match	off	on	on/off	/
0.464	0.944	0.622	partial match	off	on	on/off	/
0.155	0.241	0.188	exact match	on	off	on	long
0.595	0.926	0.725	partial match	on	off	on	long
0.148	0.241	0.183	exact match	on	off	on/off	/
0.568	0.926	0.704	partial match	on	off	on/off	/

Risultati senza classificazione

PRECISIONE	RECALL	F1	MATCH	SPACY	STANZA	RAZOR	EURISTICA
0.114	0.263	0.159	exact match	on	on	on	long
0.419	0.966	0.585	partial match	on	on	on	long
0.094	0.271	0.139	exact match	on	on	on	/
0.334	0.966	0.497	partial match	on	on	on	/
0.131	0.195	0.157	exact match	on	on	off	long
0.577	0.856	0.689	partial match	on	on	off	long
0.117	0.203	0.149	exact match	on	on	off	/
0.493	0.856	0.625	partial match	on	on	off	/
0.112	0.254	0.155	exact match	off	on	on	long
0.418	0.949	0.580	partial match	off	on	on	long
0.096	0.254	0.139	exact match	off	on	on	/
0.358	0.949	0.520	partial match	off	on	on	/
0.114	0.263	0.159	exact match	on	off	on	long
0.421	0.966	0.586	partial match	on	off	on	long
0.100	0.271	0.146	exact match	on	off	on	/
0.356	0.966	0.521	partial match	on	off	on	/
0.141	0.195	0.164	exact match	on	off	off	/
0.620	0.856	0.719	partial match	on	off	off	/
0.110	0.144	0.125	exact match	off	on	off	/
0.617	0.805	0.699	partial match	off	on	off	/
0.106	0.237	0.147	exact match	off	off	on	/
0.424	0.949	0.586	partial match	off	off	on	/

CONCLUSIONI..



Architettura **modulare** e applicazione per:

- Test
- Valutazione
- Analisi risultati

.. E SVILUPPI FUTURI

#HeidelTime

#Frames

#ApprendimentoAutomatico



FINE

GRAZIE PER L'ATTENZIONE