

The background is a complex collage. On the left, a historical figure with a beard is shown writing with a quill on a document. In the center, there's a large, dark, swirling shape resembling a globe or a celestial body. On the right, a network diagram with nodes and connecting lines is visible, along with a circular radar-like pattern. The entire image is overlaid with a dark blue diagonal band that contains the title text.

Studio dell'evoluzione storica del lemma “Piovra” nei periodici

The logo consists of a large, white, stylized letter 'C' that is open on the right side. The word 'CINECA' is written in white, uppercase, sans-serif font inside the curve of the 'C'.

CINECA



CINECA

Lo Scopo

- Studiare l'evoluzione del lemma "**piovra**", adottato dal vocabolario italiano sul calco del francese *pieuvre*, creato da Victor Hugo nel 1866 ne *Les travailleurs de la mer* (1866-1900).
- Rendere facilmente analizzabile e studiabile dagli storici un ampio dataset estratto da quotidiani storici



Le Fonti

- Digitalizzazioni di due **importanti quotidiano storici italiani**:
 - **“Il Corriere della Sera”** negli anni 1876 – 1944, 70 anni e 25.886 uscite;
 - **“Il Popolo d'Italia”** negli anni 1914-43, per un totale di 30 anni e 9.456 uscite

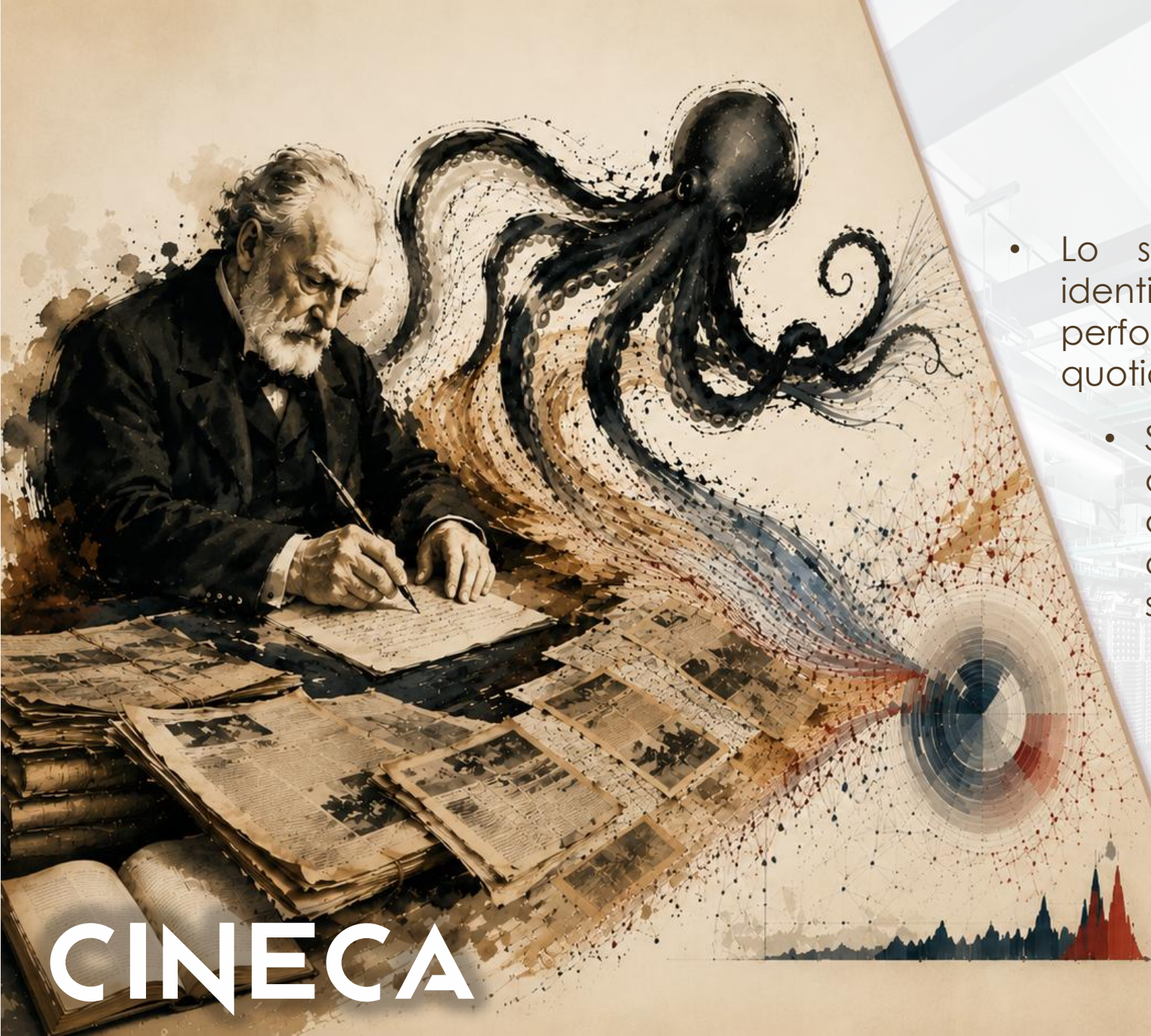




L'elaborazione

- A partire dalle digitalizzazioni, si è analizzato il **layout del quotidiano** identificando:
 - Titoli
 - Testi degli articoli
 - Fotografie
 - Grafici e disegni
 - Annunci
- **Trascrizione tramite OCR**
- **Identificazione del lemma**

CINECA



CINECA

Layout Analysis

- Lo strumento **open** allo stato dell'arte identificato, LayoutParser, seppur performante, poco si adatta all'analisi di quotidiani storici italiani.
- Si è sviluppata una **piattaforma** per aiutare gli esperti di dominio nella creazione di un dataset annotato, base che si è utilizzata per il **fine-tuning** dello strumento.

Entrambi gli strumenti sono a disposizione fra i servizi richiedibili all'interno di DPaaS:

- *Piattaforma:* **LWP Data Preparation**
- *Modello specializzato:* **lp_carlino40**



CINECA

Trascrizione

- Tutti i testi sono stati trascritti tramite uno strumento **open** di OCR

➤ OCR: **TESSERACT**



Dataset

- Il dataset è composto da tutte le **occorrenze della parola piovra** trovate all'interno dei quotidiani
- Sono presenti sia **le immagini** che tutti i **dettagli all'articolo originale contenente la parola**: nome del quotidiano, data e numero di pagina, posizione (in pixel) del testo
- Ogni articolo è corredato dalla sua trascrizione.



CINECA



Il Dataset in numeri

- Analizzate interamente due importanti testate nazionali:
 - “Il Popolo d’Italia” (1914-43)
 - “Il Corriere della Sera” (1876 – 1944)
- **35.342 uscite** processate
- **295 occorrenze** della parola piovra identificate

CINECA