

Corpus Umanistico

CINECA

Lo Scopo

Rendere accessibile e analizzabile un corpus di manoscritti rinascimentali in scrittura umanistica attraverso trascrizioni automatiche generate da un modello di intelligenza artificiale specializzato.

CINECA

Le Fonti

- **Corpus composto da 10 codici manoscritti microfilmati** (di cui 2 supervisionati)
- **Datazione:** 1401–1515
- **Consistenza:**
 - 966 carte (cartulazione)
 - 1932 facciate (paginazione)
- **Scrittura:** Umanistica
- **Elementi decorativi:** fregi, iniziali ornate, miniature, oro, mappe

Le Fonti

- **Biblioteca Universitaria di Pavia, Fondo Aldini:**
34, 86, 360
- **Biblioteca Nazionale Braidense di Brera:**
AD.IX.1, AC.IX.31, AC.IX.33, AC.X.37,
AC.IXV.44
- **Biblioteca Medicea Laurenziana di
Firenze, Fondo Acquisti e Doni:** 326, 429

Trascrizione dei 8 manoscritti non supervisionati

Pipeline di processamento:

- **Layout Analysis e Object Detection:**
 - **Medieval YOLO** per riconoscere linee di testo ed elementi decorativi
- **Handwritten Character Recognition (HCR)**
 - TrOCR_umanistica per trascrizione automatica (specializzato su scrittura umanistica rinascimentale)

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

- Modello specializzato: **TrOCR_umanistica**
- Layout analysis: **Medieval Yolo** - soon -

Trascrizione dei 2 manoscritti supervisionati

Pipeline di processamento:

- **Layout Analysis:**
 - **P2PaLA** per individuazione delle linee di testo
- **Handwritten Character Recognition (HCR)**
 - **TrOCR_umanistica** per trascrizione automatica (specializzato su scrittura umanistica rinascimentale)
- **Supervisione:**
 - Supervisione e correzione delle trascrizioni da parte di un esperto di dominio

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

- Modello specializzato: **TrOCR_umanistica**
- Layout analysis: **P2PaLA**

Il Dataset

- Trascrizioni generate dal modello TrOCR_umanistica (8 manoscritti)
- Trascrizioni generate dal modello TrOCR_umanistica e supervisionate da un esperto di dominio (2 manoscritti)

Il Dataset in numeri

- **10 manoscritti** in scrittura umanistica trascritti
- 1045 carte
- **2089** di facciate trascritte