

PUBBLICAZIONE AI SENSI DELL'ART. 19 DEL D.LGS. N. 33 DEL 14 MARZO 2013,
MODIFICATO DALL'ART. 18 DEL D.LGS N. 97 DEL 25 MAGGIO 2016
INTEGRATO DALL'ART.1 C. 145 DELLA LEGGE 27 DICEMBRE 2019 N. 160

TRACCE DELLE PROVE D'ESAME ESTRATTE A SORTE

BANDO N. 367.250 CTER (ISPC)

CONCORSO PUBBLICO PER TITOLI ED ESAMI PER L'ASSUNZIONE CON CONTRATTO DI LAVORO A TEMPO PIENO E INDETERMINATO DI UNA UNITA' DI PERSONALE PROFILO COLLABORATORE TECNICO ENTI DI RICERCA, VI LIVELLO PROFESSIONALE PRESSO L'ISTITUTO DI SCIENZE DEL PATRIMONIO CULTURALE (ISPC) DEL CONSIGLIO NAZIONALE DELLE RICERCHE - (NAPOLI)

PROVA ORALE

TRACCIA "Prova 2" (ESTRATTA)

D1: Illustrare con esempi concreti come le proprie competenze maturate durante esperienze di lavoro pregresse possano contribuire alla curatela e allestimento degli archivi digitali con strategie di mapping delle informazioni storico-archivistiche per l'estrazione della conoscenza in supporto all'attività di ricerca.

D2: Esporre quali metodi di intelligenza artificiale possono essere utilizzati per documenti manoscritti medievali;

D3: A cosa serve un software OCR e quali sono le sue caratteristiche?

D4: Valutazione della conoscenza della lingua inglese mediante la lettura di un paragrafo dell'articolo 'Hodel, Tobias, David Schoch, Christa Schneider, and Jake Purcell. 2021. "General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example." Journal of Open Humanities Data 7 (July): 13. <https://doi.org/10.5334/johd.46>. Napoli (NA), 20 gennaio 2023.

Existing text recognition engines enables to train general models to recognize not only one specific hand but a multitude of historical hands within a particular script, and from a rather large time period (more than 100 years). This paper compares different text recognition engines and their performance on a test set independent of the training and validation sets. We argue that both, test set and ground truth, should be made available by researchers as part of a shared task to allow for the comparison of engines. This will give insight into the range of possible options for institutions in need of recognition models. As a test set, we provide a data set consisting of 2,426 lines which have been randomly selected from meeting minutes of the Swiss Federal Council from 1848 to 1903. To our knowledge, neither the aforementioned text lines, which we take as ground truth, nor the multitude of different hands within this corpus have ever been used to train handwritten text recognition models. In addition, the data set used is perfect for making comparisons involving recognition engines and large training sets due to its variability and the time frame it spans. Consequently, this paper argues that both the tested engines, HTR+ and PyLaia, can handle large training sets. The resulting models have yielded very good results on a test set consisting of unknown but stylistically similar hands.

TRACCIA “Prova 1” (NON ESTRATTA)

D1: Illustrare con esempi concreti come le proprie competenze maturate durante esperienze di lavoro pregresse possano contribuire alla curatela e allestimento degli archivi digitali con strategie di mapping delle informazioni storico-archivistiche per l'estrazione della conoscenza in supporto all'attività di ricerca

D2: Qual è la differenza tra apprendimento automatico supervisionato e non supervisionato?

D3: Quali sono i tipi di memoria presenti in un calcolatore?

D4: Valutazione della conoscenza della lingua inglese mediante la lettura di un paragrafo dell'articolo 'Hodel, Tobias, David Schoch, Christa Schneider, and Jake Purcell. 2021. “General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example.” Journal of Open Humanities Data 7 (July): 13. <https://doi.org/10.5334/johd.46>.

1 THE QUEST FOR TEXT RECOGNITION

Since the early 1990s, recognition of printed text has been based on engines for optical character recognition (OCR) (Rice et al., 1993). The results have been perfected over the last fifteen years leading to satisfying results for printed material, even for printed blackletter (Neudecker et al., 2012). However, the advent of offline handwritten text recognition (HTR) lagged behind print for several decades. It was only with the implementation of deep learning, especially the cellbased neural network architecture called long-short term memory (LSTM) in the early 2010s, that handwritten text recognition achieved a quality that made such recognition processes feasible in the humanities (Graves & Schmidhuber, 2009; Leifert et al., 2016).

This paper reports on the state-of-the-art in text recognition. Its primary focus is on general models which train recognition models that are capable of recognizing not just one specific hand but similar scripts from different hands that the model has not previously seen which is one of the remaining problems in handwritten text recognition.

