Toolkit to Assess OCR’ed Historical Text in the Era of Big Data

Grant DOI: https://doi.org/10.54514/id4hs

Grantee: University of Utah

Lead Investigator: Harish Maringanti (https://orcid.org/0000-0002-1669-011X)

Keywords:

Year Awarded: 2020

Amount: $30,100

Related Items:

- Final Report: https://doi.org/10.48609/tk2e-rr32
- Presentation Slides:

Description: Cultural heritage institutions have been using Optical Character Recognition (OCR) to extract full text from scanned page images, for decades. however, the quality of extracted text is low for historical texts. In this era of big data, such historical texts will be left behind, both in search rankings and their use through computational tools. This project developed a set of guidelines, and tools that will assist organizations in improving their existing OCRed collections.