

Nordic AI Meet 2026

P1: AI, Society, Governance, Education & Working Life

14. A modular machine-mediated pipeline for historical document digitisation and humanist-corpus interaction

Authors: David Rosson¹; Mikko Tolonen²

Affiliations: 1. Department of Computer Science, University of Helsinki, Finland. 2. Department of Digital Humanities, University of Helsinki, Finland

Abstract: Preparing a critical edition of a historical work requires comparing textual variations across editions. For large works with many editions, the task may become too laborious to tackle manually. This paper presents a modular digitisation and automation pipeline, illustrated with a working prototype, the Edition Reader, for undertaking such scholarly projects in the humanities. The pipeline ingests upstream page data containing layout geometry and OCR text, maps display to page images, supports various annotation tasks for page inventory, layout geometry, and text correction, automates logical structure assignment, and shortlists textual changes across editions for examination while retracing the digital text segments to page-image evidence. This paper describes its implementation in an ongoing critical-edition project of substantial scope: David Hume's "History of England", with field experts as users, and demonstrates that the modular pipeline approach can help historians get to interpretive work sooner while preserving traceable bibliography to source pages, and that the adaptable conceptual design can generalise to similar document-centric digital humanities projects.