
Distributed Proofreaders:

Online Collaborative Production of Accurate Digitized Texts

William A. Tozier

Industrial & Operations Engineering, University of Michigan
wtozier@umich.edu

Juliet Sutherland

Director, Distributed Proofreaders Foundation
juliets@pgdp.net

Submitted as 20-minute presentation; subject matter is not appropriate for software demonstration.

Summary: We present an overview of the Distributed Proofreaders (DP) online community, its structure and performance, and a demonstration of its web-based collaborative interface. We discuss plans for future development of the community, and the use of metadata and complex document structural analysis, emphasizing the scalability and robustness of the distributed workflow structure. Finally, we sketch the role DP can play as an interface between mass digitization efforts and the more stringent needs of scholarly text production.

Distributed Proofreaders (<http://pgdp.net>) is an online community of lay volunteers devoted to the digitization and production of high-quality electronic texts of public-domain works, and is by far the largest of its kind. Unlike mass scanning and digitization efforts underway elsewhere, the work product of DP is not a searchable index, but rather a corpus of accurate electronic *transcriptions* of scanned pages, of quality comparable to that produced by a professional publishing house. As of this writing, more than 9000 books and other texts have been scanned, proofread and formatted by almost 50000 volunteers, and released online through the Project Gutenberg Archive (<http://gutenberg.org>).

The innovative approach used by DP applies a collaborative divide-and-conquer strategy to two daunting challenges: the correction of errors made by optical character recognition (OCR) software, and the creation and validation of ASCII and HTML editions of scanned works.

This work is entirely volunteer-driven, and collectively managed from within by the community, with support from the nonprofit Distributed Proofreaders Foundation. In the DP workflow, physical texts are scanned primarily by the volunteers, and shrinkwrapped OCR software is used to generate text files with these page images. Rather than approaching the work of proofreading each scanned work as a monolithic whole, once the page images and OCR text files have been uploaded to the website, thousands of volunteers work on individual pages se-

lected from a pool of active projects. A volunteer chooses any page that interests them, checks it out of the pool of work, makes the corrections needed to improve the match between the text file and its associated page image, and returns the improved text to the workflow. The text transcription of every page image in every project is sequentially checked *four or more times* by volunteers with increasing qualification and experience. When all pages of a single work have been validated page-wise for spelling, punctuation and structural errors, they are stitched back together to form a complete plain text and HTML edition—which is often subjected to further rounds of checking.

It is clear from our experiences that a community-based resource like DP can support production of very stringent electronic versions of texts, of quality sufficient for scholarly needs. OCR algorithm developers aim for “six-nines” quality, but the sheer scale of the modern digitized corpus could easily overwhelm even this unattainable error rate. Further, many older texts diverge significantly from the expectations held by OCR software developers, and contain archaic spelling and typographic variations, diacritical anomalies, rough and broken type, fraktur and inline language shifts, interlinear texts, or decorative type. Beyond these intrinsic textual traits, the physical effects of aging of the original document can interfere with OCR software’s ability to *perceive* a character, let alone recognize it correctly.

For the foreseeable future, the production of trustworthy electronic editions of texts will demand intervention by experienced humans—not just for the validation of content, but also for the annotation of electronic editions with metadata and hyperlinks. By engaging human proofreaders, the DP process can do far more than merely correct recognition errors: it can also overcome the semantic challenges of structurally complex documents.

The work-in-process at DP includes non-English language texts (about 20% of active work), profusely illustrated volumes, transcribed manuscripts, complex scholarly treatises, mathematical monographs and little-known periodicals. We present some examples of note, and discuss recent measurements of error rates, showing that they are comparable if not superior to those of professional printers (including, in many cases, the *original* printers).

A number of philosophical issues are raised by the current wave of digitization, not least those involving the nature of authority and the proliferation of variants of digitized works. We explore the relation between DP, mass digitization, and the Academy’s desire to produce and maintain orderly, authoritative texts. We argue that the DP community (and related efforts) play a powerful role as a *validation service* between automated processing and trustworthy editions.