Reading the HathiTrust Digital Library: An Exploration of Readability, Relevance, and Acceptance

Introduction: Research related to readability and reading on screens is rooted in comparisons between paper and screen, drawing on the long history of empirical research into print legibility and typography. Over the last twenty-five years this research has made constant use of print as a reference point for reading on screens, from user expectations of look and feel to navigation strategies to effects of legibility on performance. Human factors and HCI research related to readability and reading on screens focuses most closely on understanding the relationship between users and interactive technologies in particular task settings in order to inform the design of future interfaces and content delivery systems. Dillon et al (2004) argue that no single variable may fully account for reading on screens, and that “reading is best understood in user experience terms as a combination of image processing, document manipulation, and modeling of structure in pursuit of a task goal in a given context.” Several reviews have explored and synthesized human factors research into screen reading from the 1980s and 1990s (Dillon 1992, Dyson 2005, Mills & Wheldon 1987). Recent work has begun to synthesize human factors questions of readability with broader concerns with the changing nature of reading in an environment in which screen reading is becoming the default mode of text presentation for many users (Marshall 2010, O’Hara & Sellen 1997).

My research at the University of Michigan School of Information focuses on the nature, extent, and distribution of error in the HathiTrust Digital Library and the impact of page-image and full-text error on the acceptance of digital surrogates of books for a variety of purposes (Conway 2013). HathiTrust is an international preservation repository collaborative – based at the University of Michigan – preserving and providing access to the output of large-scale book digitization projects, such as those undertaken by Google, the Internet Archives, and Microsoft. A principal component of the research assesses the impact of intrinsic digital file characteristics (supplied by error data) by gaining an understanding of end-users’ evaluative and decision-making related to using digitized text for reading. Research on the use-case of scholarly research conducted using corpora of digitized books will lead to new knowledge on the ways in which large-scale data and networked digital technologies shape research methodologies. The research program contributes to user requirements research and more broadly will inform our understanding of humanities scholars’ research practices and evaluative processes in online environments.

Summer 2015 Research Project: A systematic qualitative study of information use among digital humanities scholars, exploring how faculty and doctoral students map their specific information requirements to a large corpus of digitized books, make judgments regarding use of digitized page-images and text, and assess the impact of image and text quality – and specifically error – in negative use decisions. This study will investigate connections between humanists’ evaluations of the image quality of digitized resources and their willingness to use these resources to support specific research activities. When a printed text has been transformed into a digital representation, at what point do page image errors become debilitating for use by humanists undertaking specific research activities? What digitized page image errors (in terms of frequency and/or severity) present a barrier to use by humanists for a given research activity? The population of study will be faculty and/or doctoral students in literary studies and history specializing in some aspect of the long nineteenth century. Participants will be recruited using email lists, blogs, and other resources targeted to Victorianists by the North American Victorian Studies Association. Utilizing a filtering questionnaire, we will select a sample that reflects a sufficient range of topical scholarly research interests and research activities.

Data Collection: This end user study will have two main components. We will utilize semi-structured interviews to identify end users’ needs, desires and expectations for using digitized documents. Second,
we will engage study participants in a set of structured activities, using samples of digitized books with known quality issues, to evaluate quality in use. Activities may include sorting and rating quality errors and contextualizing digitization error in terms of readability requirements for digital humanities scholarship. These activities will allow us to further specify the decision-making and evaluative processes by which end users determine a threshold for acceptable error for use of digitized texts and page-images.

**Phase 1: Preliminary Questionnaire**
- Questionnaire used to screen potential participants and gather information to be drawn on in subsequent stages. Questions include: demographic information, job title or role, frequency of use of types of digitized resources in recent scholarly research, research activities for which digitized resources are used, and familiarity/use of Google Books and HathiTrust.

**Phase 2: Structured Task**
- Introduce range of errors identified in the HT corpus to participants
- Ask participants to engage with and evaluate digitized resources that contain a known quantity and severity of error. Such engagement serves as a catalyst for Phase 3 interview.

**Phase 3: Post-Task Interview**
- Used to follow up, build on, or deepen understanding of page image error characteristics/evaluations identified in the first two phases, connect to research activities
- Semi-structured interview protocol: Identification of user needs or expectations (page image characteristics) for using digitized resources within specific activities; use of digitized resources instead of/alongside print; barriers to use and acceptable error threshold; impressionistic feedback.

**Expected Outcomes:** Students will map the use of digitized resources by Victorianists in two fields (literary studies, history) to types of research activities, and connect this mapping to evaluations of different types of page image error frequency and severity. Students will broaden their understanding of the concept of quality evaluation, moving from being conceptualized as a single moment of judgment to an ongoing, contextually bound evaluative process. Students will gain experience in crafting a specific research plan of work. Students will learn how to systematically gather, code, analyze, and interpret qualitative research data and then associate qualitative and quantitative findings on a single issue.


