Supporting Historical Research and Education with Crowdsourced Analysis of Primary Sources

Nai-Ching Wang

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science & Application

Kurt Luther, Chair
Edward A Fox
Gang Wang
Paul Quigley
Matt Lease

December 7, 2018
Blacksburg, Virginia

Keywords: Crowdsourcing, Historical Research, History Education

Supporting Historical Research and Education with Crowdsourced Analysis of Primary Sources

Nai-Ching Wang

(ABSTRACT)

Historians, like many types of scholars, are often researchers and educators, and both roles involve significant interaction with primary sources. Primary sources are not only direct evidence for historical arguments but also important materials for teaching historical thinking skills to students in classrooms, and engaging the broader public. However, finding high-quality primary sources that are relevant to a historian's specialized topics of interest remains a significant challenge. Automated approaches to text analysis struggle to provide relevant results for these "long tail" searches with long semantic distances from the source material. Consequently, historians are often frustrated at spending so much time on manually the relevance of the contents of these archives other than writing and analysis. To overcome these challenges, my dissertation explores the use of crowdsourcing to support historians in analysis of primary sources. In four studies, I first proposed a class-sourcing model where historians outsource historical analysis to students as a teaching method and students learn historical thinking and gain authentic research experience while doing these analysis tasks. Incite, a realization of this model, deployed in 15 classrooms with positive feedback. Second, I expanded the class-sourcing model to a broader audience, novice (paid) crowds and developed the Read-agree-predict (RAP) technique to accurately evaluate relevance between primary sources and research topics. Third, I presented a set of design principles for crowdsourcing complex historical documents via the American Soldier project on Zooniverse. Finally, I developed CrowdSCIM to help crowds learn historical thinking and evaluated the trade-offs between quality, learning and efficiency. The outcomes of the studies provide systems, techniques and design guidelines to 1) support historians in their research and teaching

practices, 2) help crowd workers learn historical thinking and 3) suggest implications for the design of future crowdsourcing systems

Supporting Historical Research and Education with Crowdsourced Analysis of Primary Sources

Nai-Ching Wang

(GENERAL AUDIENCE ABSTRACT)

Historians, like many types of scholars, are often researchers and educators, and both roles involve significant interaction with primary sources. Primary sources are not only direct evidence for historical arguments but also important materials for teaching historical thinking skills to students in classrooms, and engaging the broader public. However, finding high-quality primary sources that are relevant to a historian's specialized topics of interest remains a significant challenge. Automated approaches to text analysis struggle to provide relevant results for these "long tail" searches with long semantic distances from the source material. Consequently, historians are often frustrated at spending so much time on manually the relevance of the contents of these archives other than writing and analysis. To overcome these challenges, my dissertation explores the use of crowdsourcing to support historians in analysis of primary sources. In four studies, I first proposed a class-sourcing model where historians outsource historical analysis to students as a teaching method and students learn historical thinking and gain authentic research experience while doing these analysis tasks. Incite, a realization of this model, deployed in 15 classrooms with positive feedback. Second, I expanded the class-sourcing model to a broader audience, novice (paid) crowds and developed the Read-agree-predict (RAP) technique to accurately evaluate relevance between primary sources and research topics. Third, I presented a set of design principles for crowdsourcing complex historical documents via the American Soldier project on Zooniverse. Finally, I developed CrowdSCIM to help crowds learn historical thinking and evaluated the trade-offs between quality, learning and efficiency. The outcomes of the studies provide systems, techniques and design guidelines to 1) support historians in their research and teaching

practices, 2) help crowd workers learn historical thinking and 3) suggest implications for the design of future crowdsourcing systems

Acknowledgments

I am deeply grateful to the following people, and many others not specifically named, without whom the work described in the following pages would not have been possible.

First, I would like to thank my parents, brother and sister. Their endless reserves of enthusiasm helped to replenish my own when it mattered most. Jingzi, with her warm encouragement and company, not only helped me finish the last mile but also made it pleasant.

My advisor, Kurt Luther, who gave me a chance to work with me and became my trusted mentor. He embodies many of the qualities I admire in a scholar, such as putting his students first, valuing teaching, research quality, and exemplifying the highest standards for ethical behavior and research excellence.

My dissertation committee, Ed Fox, Paul Quigley, Gang Wang and Matt Lease, kept me focused on the questions that mattered when my research took unexpected turns. They challenged me to do my best work, and always quick to provide feedback and encouragement to help me me that challenge.

Mapping the Fourth project collaborators, especially David Hicks, Daniel Newcomb, Kevin Caprice, helped contribute ideas and resources to the digital archive.

The American Soldier project collaborators, especially Ed Gitre, Bradley Nichols, and Amanda French, helped contribute their expertise to the American Soldier.

Contents

List of Figures

List of Tables

1

Introduction

It is essential to support historians for their research because advancements in historical research help us understand ourselves by knowing how people and societies have functioned and provide guidance how societies could evolve by showing examples in the past. In addition, well-told history also acts as art and entertainment that bring aesthetics to our lives.

Historians, like many types of scholars, are often researchers as well as educators, and both roles involve significant interaction with primary sources. Primary sources are not only direct evidence for historical arguments [114] but also important materials for teaching historical thinking skills to students in classrooms, and engaging the broader public [129, 130]. However, finding high-quality primary sources that are relevant to a historian's specialized topics of interest remains a significant challenge. Manual indexing by professional metadata librarians is cost-prohibitive for many organizations, but when it is available, annotations are often provided to serve the broadest possible audience. Automated approaches to text analysis also struggle to provide relevant results for these "long tail" searches with long semantic distances from the source material.

Many research materials (e.g. primary sources mentioned earlier) historians use for their research are from galleries, libraries, archives and museums (GLAMs) which are also important sources for other humanities researchers. These institutions contain many (first-hand) primary sources that historians need to support their arguments and answer their research questions. Although these cultural institutions have often invested much effort on digitiz-

ing their collections of historical resources, researchers may not discover relevant materials in these archives because expert transcription and metadata generation is expensive and time-consuming. Taking transcribing digitized images for example, institutions often use automated techniques, such as Optical Character Recognition (OCR), to turn images into texts. Unfortunately, the resulting texts usually are not perfect and require manual corrections from humans. According to a report from the Trove project of National Library of Australia in 2013, it might have cost the library $12 million dollars to correct 78 million newspaper lines [12]. Consequently, due to the limited and declined funding in humanities [5, 19], institutions have explored crowdsourcing as a way to simultaneously engage the public in history while generating scalable transcriptions and tags [13, 15]. These efforts have already saved much time and money and yielded valuable results for scholarly research and successfully engaged many "citizen archivists" for public history such as history education.

Archival activities are preserving historic materials and making them available for use [14]. Citizen archivists are non-professionals (voluntarily) involved in these archival activities. Common examples of the activities include scanning, transcribing, tagging, uploading, and editing documents. With the help of these citizen archivists, many more records have been made accessible and searchable to historians (and the broader public) who may need these records for their research. Public history is a synonym of applied history, that is, how history is used in diverse ways in the world and how history is applied to real-world issues [1]. Therefore, public history is an important means for historians to apply their research to contemporary issues and to demonstrate the value and impact of their research on the general public and societies. It is also a way to return the favor of contributions made by citizen archivists to their research and to encourage more citizens to participate in these activities that help with their research. These results make crowdsourcing a promising approach to help with the two supports historians look for. However, there are at least three challenges

to integrating crowdsourcing into historical research.

1.1 Challenges to Integrating Crowdsourcing into Historical Research

While crowdsourcing has been shown to be a potentially powerful tool for research either through many citizen science projects [16] and/or the use of crowdsourcing market places (e.g., [28, 32, 97]) , prior work [87] shows that there are several potential uncertainties (knowledge, quality, process, data, and delegation) preventing crowdsourcing from being more widely adopted. For example, knowledge uncertainty is that doing complex tasks generally require extensive domain knowledge where there are ambiguities in the tasks, lacking well-defined rules to complete the tasks. Quality uncertainty is that crowdsourced data may require additional examination which may actually increase research workload.

Unfortunately, historical research and history education is one of the under-studied fields in crowdsourcing literature. Therefore, the first challenge is that beyond crowdsourced transcriptions, backed by many successful projects, it is hard for historians to adopt crowdsourcing in their research and teaching practices in more complex tasks with these uncertainties.

A second challenge is that even if historians are willing to take the risk of trying crowdsourcing for their research and teaching practices, it is unclear how to design a crowdsourcing system to scale up crowdsourced tasks for history domain from the literature although there is reflection of crowdsourcing projects (e.g., [112, 113]). From crowdsourcing literature, designing an effective crowdsourcing system generally requires additional technical and design expertise that historians may not have. It also take much effort in designing an engaging crowdsourcing system or costs much money to attract enough crowds to make substantial

contributions.

A third challenge is that while much of the crowdsourcing research focuses on productivity, there are few studies exploring how to help the crowd learn skills to help crowds develop their career while doing crowdsourced tasks [56, 88, 109, 156]. Moreover, results of these few studies are mixed and scattered in different domains. The introduction of learning in a crowdsourcing workflow may have no, positive or negative impacts on work quality depending on what is learned, what the crowdsourced task is and in which domain.

1.2 Research Questions

I propose the following research questions of the four studies whose answers will help address the three challenges to integrating crowdsourcing in historical research and history education.

1.2.1 Incite and class-sourcing

RQ 1a: How can we design a crowdsourcing system to both support historical research and history education?

RQ 1b: How do teachers and students use such a system?

1.2.2 RAP: Scaling up Crowdsourced Historical Connections

RQ 2a: How well does the novice crowd make connections between historical primary sources and high-level research topics?

RQ 2b: How can crowds connect related primary sources to scholarly topics as accurately as historians?

RQ 2c: How can crowdsourcing systems identify opportunities for public history interventions?

1.2.3 Zooniverse: Scaling up complex crowdsourced transcriptions

RQ 3: How can we scale up the analysis of complex historical documents while minimizing cost and effort?

1.2.4 CrowdSCIM: Scaling up learning

RQ 4a: How can we help crowd workers learn domain expertise in history domain (historical thinking)?

RQ 4b: How does the introduction of learning affect work quality and efficiency of crowdsourcing?

1.3 Study Overview

To address these three challenges to supporting historical research and history education via crowdsourcing, in my first study (Chapter 3), I first proposed a class-sourcing model to enhance historians' existing practices instead of introducing radical new tools or changes. To realize the class-sourcing model, I designed and developed Inicte, a plugin to a well-known open source content management system, Omeka, already used by many cultural institutions such as libraries and museums to build their own digital archives and showcase their digital collections. Incite helps crowdsource digital collections on Omeka by introducing several common crowdsourcing tasks including transcribing, summarizing, tagging, connecting and

discussing.

I then deployed Incite with two historical digital archive projects, Mapping the Fourth and the American Soldier. I used them as case studies to see how Incite may help support historical research and history education in a classroom setting with students as the crowd. Mapping the Fourth is a collaborative project between Computer Science, History, Education and University Libraries at Virginia Tech to create a freely-accessible crowdsourced digital archive about American Civil War and Independence Day. Mapping the Fourth consists of more than 4000 raw primary sources (still growing), various high-level topics historians currently use for their research such as Racial Equality, and several common crowdsourced tasks (transcribe, summarize, tag, and connect) on digital archives. The American Soldier is another collaborative project between Computer Science, History and University Libraries at Virginia Tech to transcribe research surveys of American soldiers during World War II era.

With the generally positive feedback of the class-sourcing model from both case studies, my second study (Chapter 4) was to expand student crowds to more general novice crowds to see how the crowd perform beyond transcription tasks in history domain. I first conducted an experimental study with novice crowd from Amazon Mechanical Turk on a customized version of Incite using data from a real historical archive to understand how different crowdsourcing tasks (reading, tagging keywords and summarizing documents) may affect crowdsourced production for scholarly and public history. Second, from the experimental results, I developed a crowdsourcing algorithm, read-agree-predict (RAP), to accurately collect related sources for scholarly history and detect high-impact opportunities for public history. Third, I validated read-agree-predict with another historian and a new dataset. It also serves as an example how the developed crowdsourcing algorithm can be used for other similar crowdsourcing projects and systems.

The experimental results indicated that crowdsourced connections filtered by majority vote could save up to 75% of the time historians spend on unrelated documents from the dataset. The results also showed that read-agree-predict predicted exact matches based on the historian's pattern of making connections and was able to help identify high-impact learning opportunities while summarizing helps disclose details of misconceptions. The validation showed read-agree-predict had identical predictions of historian's answers and one missed high-impact opportunity.

After demonstrating crowdsourcing can be a useful research tool in history domain, my third study (Chapter 5) explored how to design a crowdsourcing system to scale up crowdsourced transcriptions of complex documents while minimizing required cost and effort for historians. While there are many crowdsourced transcription projects, there were no clear design guidelines to help historians design workflows of a crowdsourcing system to crowdsource their archives. From crowdsourcing literature, we know that workflows play a very important role in the success of crowdsourcing systems (e.g., [29, 44, 90, 93]). In addition, the documents of many crowdsourcing projects were primarily from single source (e.g., diaries of the same person) in the same format (e.g., only handwritten). It was not clear how to design a crowdsourcing system that has to deal with a collection with a lot of document variants (29 in our case study) from different respondents (more than 65,000 soldiers). To fill in this gap, I went through an iterative design process with subject experts, students and public users to design the American Soldier project, an official and featured project on Zooniverse [17], the largest citizen science platform in the world. I documented the design process and proposed a set of generalizable design guidelines that will help design of future crowdsourcing projects.

To address the third challenge about supporting history education at scale, my fourth study (Chapter 6) investigated how to scale up learning core domain expertise in history domain (historical thinking). I first designed CrowdSCIM, a crowdsourcing system that can help

crowd workers learning historical thinking while doing crowdsourced tasks, via a series of four pilot studies. I then evaluated CrowdSCIM by comparing it with a baseline and two other state-of-the-art crowdsourcing techniques in terms of learning, quality and efficiency.

The results show that CrowdSCIM was the only one of the four tested techniques that can help crowd workers learn historical thinking without impeding quality of crowdsourced tasks. Even better, CrowdSCIM is the only one of the techniques that improves quality of summary. With these benefits, CrowdSCIM takes more time to finish the tasks. I then discussed the potential trade-offs between these techniques in terms of the three measures and implications for future crowdsourcing research.

1.3.1 Thesis statement

Crowdsourcing can support historical research with more complex tasks beyond transcriptions such as summarizing, tagging and connecting primary sources in an archive to high-level topics/themes historians use for research and as a by-product, identify opportunities for learning interventions for classroom and public history (RQ 1 and RQ 2). A set of design guidelines explaining potential trade-offs and suggesting sweet spots helps minimize cost and effort required for designing highly scalable crowdsourced projects (RQ 3). While doing crowdsourced tasks, crowd can learn domain expertise with an appropriate design (RQ 4).

2

Review of Literature

In this chapter, I review related work on historical research, crowdsourcing, educational psychology, and history education.

2.1 Historical Research

Historians, like researchers in many disciplines, regularly engage in sensemaking processes involving large amounts of complex data in order to better understand the past. History (i.e., the past) is the way how we understand how we have arrived in the present and how we may proceed towards the future [100, 128]. History helps us understand how people and societies have operated and evolved and thus how we define our identities and societal settings such as ethics and politics. It also provides guidance how societies might progress by showing examples in the past. Although the past is often unclear and ambiguous, progression in historical research helps clarify the foggy past and reveal what has been behind the curtain. In addition, history has been a good source of artful and entertaining stories that bring aesthetics and joy to our lives [128]. In order to identify services to better support research practices of academic historians in the U.S., Ithaka S+R, a consulting agency, conducted a study with 39 historians and 14 research support professionals [92]. The study and others found that interacting with primary sources, such as gathering, discovering, and organizing, is still central to historical research [104, 114, 144]. While gathering sources, historians identify

related sources to address their research questions and support arguments, and organize the sources dominantly based on topics of interest so that they are able to later (re)use these topics to find the sources again via a top-down approach.

The same study emphasized that even with modern search engines, historians still spend a large part of their daily work gathering and discovering sources related to their research topics [114]. With modern search engines, historians often cannot search by topics of interest directly. Instead, they may try many different search terms find resources related to topics of interest, and filter out many irrelevant search results caused by wrong search terms. However, organizing sources into topics of interest (e.g., American Nationalism) takes considerable time. Some historians even regret spending the time on organizing instead of other research activities. As one historian notes: "Once it's organized, it's up to me to think about it and write. But I do resent the time that's spent organizing and managing everything" [114]. In addition to scholarly historical research, the same study further mentioned the trend of "engaging the public" is inevitable in the discipline. With the adoption of digital methods, many historians are motivated to have conversations with the public and to help the public be better informed about the world with their special expertise in various topics [107, 114]. In particular, historians who leverage public contributions such as crowdsourcing or publicly-generated sources, often feel committed to share their work with public. However, identifying promising opportunities for high-impact public history interventions that leverage the historian's expertise can be challenging.

2.2 Crowdsourcing and Digital Archives

Crowdsourcing in digital archives has been classified into six broad types: 1) correction and transcription tasks; 2) contextualization; 3) complementing collection; 4) classification;

5) co-curation; and 6) crowdfunding [108]. Since historical research still involves much in text processing, we only focus on the ones that are directly related, that is, classification via tagging/labeling, and contextualization via summarizing by assuming transcriptions are available because of many successful projects (e.g., [6, 7, 8, 13, 15]). In addition to basic transcribing tasks, many projects also include tagging or labeling tasks to recognize named entities, provide alternative search terms based on folksonomy and better reflect user's perception of the resources [9, 10, 11, 134]. A large number of tags has been collected via many successful projects such as Library of Congress' Flickr Commons Projects (about 70 thousand tags for about 4615 resources within first 10 months) [18], steve.museum (about 43 thousand tags for 1784 resources in about first 2 years) [134], and National Library of Australia's Trove project (1.7 million tags in about first 5 years) [13]. Many crowdsourcing projects have also started to ask users to provide more high-level metadata such as providing brief description [106], synthesizing and editing articles [3], providing notes [8], and providing summary [4]. Some research has begun to investigate how crowdsourced production can be used for scientific research which requires high quality data [121, 141]. Although research on crowdsourced digital humanities efforts has considered how crowd-generated metadata (e.g. a folksonomy) could be useful for archives [96, 134] and how crowdsourcing could engage the public [112, 113], little research addresses how crowdsourced production can be directly used for and integrated into scholarly humanities such as historical research. Disciplines such as history, researchers often already have topics of interest in mind, and use a top-down approach to find relevant primary sources and answer their research questions [114, 144].

2.3 Crowdsourced Clustering

There have been several studies exploring the use of crowdsourcing for clustering and categorizing textual data. Some studies produce taxonomies by introducing workflows to combine crowd workers' results based on predefined algorithms or processes [24, 44] and may also combine results from machine learning techniques to increase accuracy and efficiency [23, 34, 38]. Some studies introduce scaffolds and systematic guidance to cluster unstructured text [24] and merge structured information [93]. These studies focus on crowdsourcing the generation of categories and taxonomies from unprocessed texts and compare the results between crowd workers and experts. In contrast, my dissertation work focuses on crowdsourcing the generation of (multiple) connections between existing topics and raw texts by following the same historians' patterns to expand their work. This is closer to real situations and creates the opportunities to motivate citizen scientists/archivists by engaging them in real research.

2.4 Crowdsourced Assessment

Crowdsourcing has also been used for assessing relevance of documents to given topics or queries, especially in information retrieval domain (e.g. [20, 65, 77, 98]) and credibility of claims or statements from given texts (e.g. [57]). Early studies focused on the assessment task (i.e., directly ask crowd workers to assess desired attributes of the given materials) and how to improve accuracy with various techniques such as simple majority vote [127] and expectation maximization [53]. To assess the relevance of web content to a given query, recently a two-stage rationale task was proposed to provide near-perfect accuracy (96%) for relevance judgements by adding to the existing assessment task an additional rationale step along with two filtering algorithms [98]. To assess credibility of claims about relations

derived from given texts, Microtalk [57] introduces a argumentation process to select "discerning" crowd workers to complete extra "justify" and/or "reconsider" tasks in addition to the existing assessment task and is able to achieves 84% accuracy. This is an improvement over both simple aggregation with majority (58% accuracy) vote and simple aggregation with expectation maximization (64% accuracy). As seen in previous work, accuracy can vary a lot depending on the application domain. My dissertation work differs from these studies in that my work deals with historical documents and high-level topics historians use for their scholarly research, while these studies focus on general web contents/queries or relation extraction tasks.

2.5 Crowd Workflows

Many crowdsourced tasks are designed to be at micro scale in order to overcome some common challenges of crowdsourcing such as workers being transient, lacking expertise, or providing conflicting responses [24]. These crowdsourcing systems often need well-designed workflows which usually involves decomposition of a large, complex task into many micro-tasks, arrangements and assignments of these micro-tasks to workers, and aggregation of output from various micro-tasks to produce final results (e.g. [24, 29, 44, 81, 90]). Crowdsourcing research shows that the arrangement and design of (multiple) crowdsourced (micro-)tasks of a crowdsourcing system may have impact on the work efficiency, output quality and crowd's learning of these crowdsourced tasks [37, 40, 90, 93]. Different orders of a chain of micro-tasks in terms of complexity, operations and content may affect work efficiency, mental load and perceived difficulty at sentence-level in the writing domain [37]. Context and structure may also affect quality of results and work efficiency [93].

Prior research has also explored some trade-offs between various crowd workflows such as

iterative (or say, sequential, usually in a linear form), parallel (usually in a tournament form) , and simultaneous workflows [25, 44, 90, 93]. Depending on the applications and purposes, different workflows may have different benefits. For example, an iterative process may help increase average response quality but may suffer from biases introduced in early iterations. A parallel process may help avoid passing biases and the priming effect but may suffer from duplicate work. My dissertation work differs these prior studies in that these prior studies focus on the same task (e.g. limerick writing and image description) with different workflows, while my work involves different tasks (transcribing, tagging, and connecting) with different workflows in the history domain.

Another branch of crowd workflow research focuses on building tools for design complex crowdsourcing workflows such as Turkit [91], CrowdForge [81], CrowdWeaver [82], Turkomatic [85] and ReTool [39]. However, these tools require technical expertise such as programming scripts to create effective workflows and/or were designed for crowdsourcing markets. My dissertation work (Chapter 5) differs from these tools in that a set of design guidelines was proposed and can be easily applied to existing volunteer-based crowdsourcing platforms such as Zooniverse without the requirement of programming expertise.

2.6 Automated Techniques for Text Classification and Topic Modeling

The Machine Learning, Information Retrieval and Natural Language Processing research communities have developed automated techniques to help cluster related documents and/or model the relationships between documents and topics based on semantic similarity. Techniques such as TF-IDF use the concept of bag-of-words and word frequency in the document

[115]. They model documents as vectors so similarity between documents can then easily be calculated by measuring distance between vectors. Other, more complex techniques introduce the concept of latent semantic structures in the document (topic modeling) such as latent semantic indexing (LSI) [52], probabilistic latent semantic indexing (pLSI) [72] and latent Dirichlet allocation (LDA) [30]. Topic modeling techniques such as LDA see documents as having some underlying latent semantic structure (distribution of topics), which may be inferred from word-document co-occurrences. Documents can therefore be clustered based on similarity or different topics (latent semantic structure). Unfortunately, these clustering techniques are not able to generate meaningful labels for the clusters, let alone to understand existing labels and follow the same patterns that historians use. Recent studies show that for using LDA-based techniques for textual datasets similar to our dataset, there were still a big portion (one-third in [69] and two-thirds in [24]) of topics (i.e., clusters or groupings) that were not interpretable or helpful. On the other hand, my dissertation work (Chapter 4) focuses on understanding existing topics and clustering documents by producing connections between existing topics and historical documents like historians do. Techniques for multi-label classification can also be used to help connect documents to high-level topics historians use for research because each topic can be seen as a label that might be related to multiple documents. These techniques can be generally classified into two categories: 1) problem transformation and 2) algorithm adaptation (see [155] and [118] for a detailed review). Problem transformation is a technique that transforms a multi-label problem into multiple single-label problems which can then be solved by an algorithm for single-label classification. For example, Binary Relevance (BR) [33] divides a multi-label classification problem into multiple independent binary classification problems, each of which corresponds to a label in the label space. Algorithm adaptation is a technique that adapts algorithms for single-label classifications. For example, Multi-Label k-Nearest Neighbor (ML-kNN) [154] is an adaptation of k-Nearest Neighbor (k-NN) algorithm that is for single-label problems.

Recent studies show that performance (i.e., accuracy) of multi-label classification algorithms alone are far from perfect [92, 135]. Therefore, my dissertation work (Chapter 4) first focuses on leveraging human intelligence via crowdsourcing to achieve (near-)perfect accuracy and these connections can be used as gold standard to help develop better or train existing machine models.

2.7 Psychology of Learning with Semantic Processing

Making good connections between documents and related topics requires humans to have a good understanding of the documents and topics after reading the documents. Reading comprehension has been modeled as a complex cognitive process involving different levels of lexical and semantic processing [80]. Research on levels of processing suggests that deeper elaboration leads to better recall and understanding (e.g. [49, 50]). The action of tagging, labeling, and finding keywords is very similar to underlining while providing a note or summary is very similar to summarization. Underlining (i.e., highlighting or identifying important words) has been reported to be the most frequently used study technique by college students to improve their reading comprehension [1]. Some studies argue there is no performance improvement using underlining techniques and if there is, the improvement is only on underlined parts caused by von Restorff effect (i.e., a salient isolated item is more likely to be attended and remembered than the rest of the items) [41, 42, 43, 105, 110, 123]. At the same time, several other studies have shown that people have better reading comprehension while reading and underlining important words from prose than people reading without underlining [31, 117, 153], for both intentional (underlined) and incidental (non-underlined) materials [31, 105], and attribute the performance improvement to deeper level of processing. According to generative model of learning [148], reading comprehension requires that readers

relate their prior knowledge and past experience from their memories to the meanings of the text. The text in a document serves as retrieval cues that trigger semantic processing of stored information in memory. With the semantic processing, the reader produces meanings for the text and the meanings then become the reader's comprehension of the text. By actively constructing meaningful elaborations of the text (e.g. generating a summary), the reader's comprehension of the text is enhanced. Several studies have confirmed this model by showing that participants reading with summarization perform significantly better than those read without summarization in terms of (immediate) memory recall and comprehension tests [89, 149, 150] (see [150] for a detailed review). Another study from levels of processing also confirms that participants read with summarizing or paraphrasing activities outperform those read without these activities [36]. From the literature, there are mixed results as far as which technique is better, but both seem promising. In my dissertation work (Chapter 4), I directly compare them in the context of crowdsourced document analysis. I hypothesize that summarization will have stronger effect on performance than underlining because a recent study suggested that summarization requires deepest level of process among tasks in writing domain [37]. A better understanding of the potential effects on the crowdsourced production can help better organize these tasks. For example, if these tasks increase the quality of the production, then we should try to integrate these tasks into the crowdsourcing workflow but if the tasks decrease the quality of the production, then we might want to avoid or separate the tasks.

2.8 Historical Thinking and History Education

A Library of Congress publication colorfully depicts historians as detectives searching for evidence among primary sources [145]. Learning history is more than merely memorizing facts

from various sources. Although historians have expertise in different periods or topics of history, they share some common way of thinking history and analyzing historical documents [146]. The analysis of sources includes identifying factual information, evaluating reliability of sources, understanding multiple perspectives, contextualizing sources in time and space, reasoning and inferences, cor- roborating across multiple sources, and generating possible understandings and interpretations [27, 46, 143, 144]. Several learning approaches have been proposed to help students learn history through historical thinking, such as learning through authorship [67], apprenticeships (or guidance) [47, 116], and confronting questions [147]. These strategies generally require substantive interactions between a human instructor and a student. Some studies have shown that the use of hypertext scaffolding may support some historical thinking processing [74, 75]. Building on these studies, Hicks et al. developed SCIM-C [71], a strategy that can scaffold the historical thinking process when analyzing historical primary sources. It includes five phases: 1) Summarizing information and evidence from the source, 2) Contextualizing the source in time and space, 3) Inferring from subtexts and hints in the source, 4) Monitoring initial assumptions and overall focus, and 5) Corroborating under- standing across multiple sources. Evaluations showed that SCIM-C is an effective strategy to help students learn historical thinking through multimedia embedded scaffolding [70, 99]. While SCIM-C has been shown to be effective in the classroom with collocated students, experienced teachers, and multi-day training sessions, its applicability for novice crowd workers is unknown. My dissertation work (Chapter 6) explores how SCIM-C can be adapted for a micro-tasking context, providing just-in-time domain expertise for workers completing tasks requiring historical thinking skills.

2.9 Crowd Learning

2.9.1 Learner-sourcing

Some research has begun exploring the use of crowdsourcing in classroom-related settings. This body of work focuses on improving learning with collective learner activity or receiving feedback from other (paid) crowds, including creating crowdsourced subgoals in how-to videos [78, 79, 140], crowdsourced assessments or exercises [102, 122], personalized hints for problem-solving [62] , receiving design critiques [66, 151], collaborative discussion [45], identifying students' confusions [61], and generating explanations for solving problems [142]. These studies try to address the issue of low ratios of expert teachers to learners, especially in MOOCs. This body of research is also termed learner-sourcing because it focuses on how learners can collectively generate useful learning ma- terials for future learners. CrowdSCIM in my dissertation work differs from these studies in that CrowdSCIM is built on top of a scaffolding technique to be used without the need of other peer learners or crowds. While these learner-sourcing techniques require additional (learner) crowds' participation or content production (e.g., sub-goals in how-to videos, design critiques, and explanations) to facilitate learning, a CrowdSCIM user can learn historical thinking while doing tasks without feedback or participation from others. Further, CrowdSCIM is designed for paid crowd workers, a population with greater limitations of time, interest, and expertise, than students in MOOCs or traditional classrooms.

2.9.2 Crowd learning on citizen research platforms

While citizen research platforms like Zooniverse have attracted many non-professionals to contribute to major discoveries, these projects are also considered a means of engagement and

outreach, such as citizen science and public history. Yet, recent studies show that learning often happens outside the context of the crowdsourced tasks [76]. Most relevant to my work, Crowdclass [88] was among the first efforts to design in-task learning modules for citizen science. Similar to my work, Crowdclass focuses on paid novice crowd workers and uses pre- and post-tests to measure learning. However, unlike my approach, Crowdclass focuses on learning factual knowledge; workers correctly answer multiple-choice questions (and "hybrid questions" synthesizing facts across multiple lessons) to demonstrate mastery and advance in a hierarchy of learning modules. In contrast, CrowdSCIM teaches workers to consider the meaning of a document from multiple perspectives by reflecting on a set of generalized questions. Although direct comparisons are complicated by differences in task and domain (i.e., analyzing historical documents vs. classifying galaxies), CrowdSCIM has the benefit of being content-agnostic within a domain. Crowdclass may require experts to design new questions and answers to teach and test each new type of fact, whereas CrowdSCIM does not require expert intervention when new documents are presented. CrowdSCIM builds on Incite (see Chapter 3 for summary, tone rating, tag and theme rating in the history domain. I chose Incite as our target crowdsourcing platform for a few reasons. First, it includes a variety of higher-level tasks to support historical research, including summaries, tone ratings, tags, and theme ratings, in contrast to most crowdsourced history platforms that focus on simple transcription (e.g., [6, 8]). Incite groups these tasks into three steps: Transcribe (transcribe, summarize, and rate tone), Tag (tag entities), and Connect (rate theme). This selection of tasks is consistent with what prior work suggested in supporting historical research [58, 114]. Second, Incite has been used by real digital archive projects to support historical research. Third, it is open-source and easy to plug into existing digital archives. While Incite may also be used to support history education, CrowdSCIM differs in that Incite is optimized for students in classrooms with instructors' intervention over a multi-day time span, while CrowdSCIM is designed as a standalone system with novice, paid crowd workers and micro-

tasks. Also, CrowdSCIM excludes the simple transcription task and focuses on higher-level Summary-tone, Tag, and Connect tasks.

2.9.3 Crowd learning and work quality

While most crowdsourcing studies focus on work quality, some research considers both worker learning and work quality [54, 55, 56, 156]. While most of these studies show that learning can help improve quality, others do not. Pandey et al. [109] found that workers who had access to MOOC-style learning materials about microbiomes scored higher on a subject matter test, yet produced similar work quality (i.e., generating creative ideas about microbiome influences) compared to workers without access to the learning materials. Crowdclass [88] shows that a workflow designed for learning may actually lower the work quality. These mixed results motivate our current study. To understand CrowdSCIM's potential trade-offs between learning, quality and efficiency, I selected two of the most similar approaches from prior work, Reviewing vs. Doing (RvD) [156] and Shepherd [56] as comparison conditions. In RvD, Zhu et al. [156] found that workers who review others' work perform better on subsequent tasks than workers who simply performed more tasks. They theorize that reviewers experience learning benefits seen in offline studies of mentorship. In Shepherd, Dow et al. [56] compare the performance of workers receiving no feedback to workers who either perform a self-assessment using a rubric, or receive an external assessment from an expert. Self-assessment was as effective as expert assessment in improving work quality. Aiding the comparison to CrowdSCIM, both prior studies reported work quality and learning in detail, and both included some type of writing or summarization tasks, though in different domains. However, it is not clear which technique works better and how they are applicable to other types of tasks and domains. Moreover, both RvD and Shepherd focus on learning the task through provided rubrics, while CrowdSCIM focuses on learning domain expertise through

analytical thinking skills. This comparison supports a close examination of which type of learning is more effective for gaining domain expertise (i.e., historical thinking).

2.10 Chapter Summary

In this chapter, I first described historical research and supports historians look for. I then described how crowdsourcing has been a potential approach to support historical research and history education followed by different aspects of crowdsourcing research including crowd assessment, clustering, workflows and learning. I also discussed related automated techniques and how results of crowdsourced production can be used to improve these automated techniques.

3

Incite and Class-sourcing

> *"Thanks to you all! Really appreciate this and I have to say, some of my students were SO into this. It was fun."*
>
> – A beta user (a college history professor) of Incite

This chapter is based on work presented at the Annual Conference of the National Council for the Social Studies (NCSS 2016) [111] and the Annual Meeting of the American Historical Association (AHA 2018) [138].

3.1 Motivation and Research Questions

From studies on how to better support historians, there is a clear gap between historians' goals for conducting research and getting value from digital archives of primary sources, and historians' educational practice in the classroom.

As discussed in Chapter 2, many crowdsourcing projects in digital humanities have been successful collecting meaningful and useful data (e.g. transcriptions and tags) from the crowd. That encourages more archives, either from institutions or scholars, to be open to the crowd. However, it is not clear how crowdsourcing can make these archives more usable and valuable to a wide variety of scholars. Therefore, in addition to some fundamental tasks such as transcribing to make primary sources searchable, Incite is designed to also collect

metadata from the crowd in a more meaningful way so that the primary sources can be more broadly useful and valuable for scholarly research. For example, researchers like historians are also invited to provide topical directions for collecting metadata from the crowd. In addition to default topics, historians can specify their research topics of interest and the crowd can focus on them so that the crowdsourced production will contribute directly to real scholarly research goals.

In spite of much contribution from the crowd in various crowdsourcing projects mentioned in Chapter 2, it is not clear if and how the crowd learns the subject matter by participating in crowdsourcing projects in the digital humanities. Recent research from citizen science shows that it is possible for the "citizen scientists" to learn domain knowledge while working on well-designed crowdsourced tasks [88]. However, there is little research on how crowdsourcing systems might be designed to help the crowd learn domain expertise and knowledge in the humanities.

Due to the uncertainties of crowdsourcing in history domain (mentioned in Chapter 1.1), it might be too risky for historians to adopt, let alone take advantage of crowdsourcing to support their research and education practices. To address these uncertainties, I proposed a class-sourcing model and developed a crowdsourcing system, Incite, to support historians' existing practices in classroom settings instead of proposing any radical changes. The concept of class-sourcing model is that while a historian teaches history in a classroom, the historian also crowdsources the analysis of primary sources to the students. As for the students, while the students learn how to analyze primary sources, they also contribute to historical research. To realize this class-sourcing idea, I asked the following research questions:

RQ 1a: How can we design a classroom-based software tool that helps students learn history while generating valuable historical analysis for instructors?

RQ 1b: How do teachers and students use such a system?

3.2 Method

To answer the research questions, I used design-based research [26] to go through an iterative process with several historians and educators, deploy Incite in several history classrooms and gather feedback from the educator historians about how well Incite supports history education.

3.3 Design Process

The design process of Incite can be broadly divided into three periods including 1) initial design; 2) initial feedback from domain experts (e.g. history professors and teachers); and 3) iterative development with feedback from collaborators and volunteers. While Incite is designed to be a more general crowdsourcing platform for digital archives, Incite uses a digital archive project, Mapping the Fourth of July, as a concrete example to envision how Incite may be useful for historical research and education. Mapping the Fourth is a collaborative project between Computer Science, History, Education and University Libraries at Virginia Tech to create a freely-accessible crowdsourced digital archive about the American Civil War and Independence Day. Mapping the Fourth also has an advisory board, core users and several volunteer history professors and teachers who give feedback and/or test Incite in classrooms.

3.3.1 Initial design

To support historical research and education via crowdsourcing, Incite was initially designed based on a literature review of history education, psychology, and crowdsourcing. We first collected what tasks (i.e. activities) are appropriate to be crowdsourced and then how these tasks might be related to history education. We identified four activities, including transcribing, tagging, connecting and discussing, that were likely to produce metadata that could be used support historical research, and could be mapped to steps of historical thinking.

3.3.2 Initial feedback from domain experts

The initial design of Incite was presented to the advisory board and core users in a workshop hosted for Mapping the Fourth. The advisory board and core users include 15 high school and college history instructors and experts from several other domains including Human-Computer Interaction, Library and Information Science, and History Education. These experts' feedback has been used to understand their research and pedagogical goals to improve and extend the design of Incite. For example, the group feature of Incite was inspired by discussions about classroom use.

3.3.3 Iterative development

After the workshop, our development process has been iterative. After a major feature from the design is implemented, the feature is demonstrated to project collaborators for feedback in regular project meetings. After the feature becomes stable, the feature is released to core users and other volunteer history professors and teachers to be used for their classes. Feedback from these teachers after class uses is then used to revise Incite. For example, the

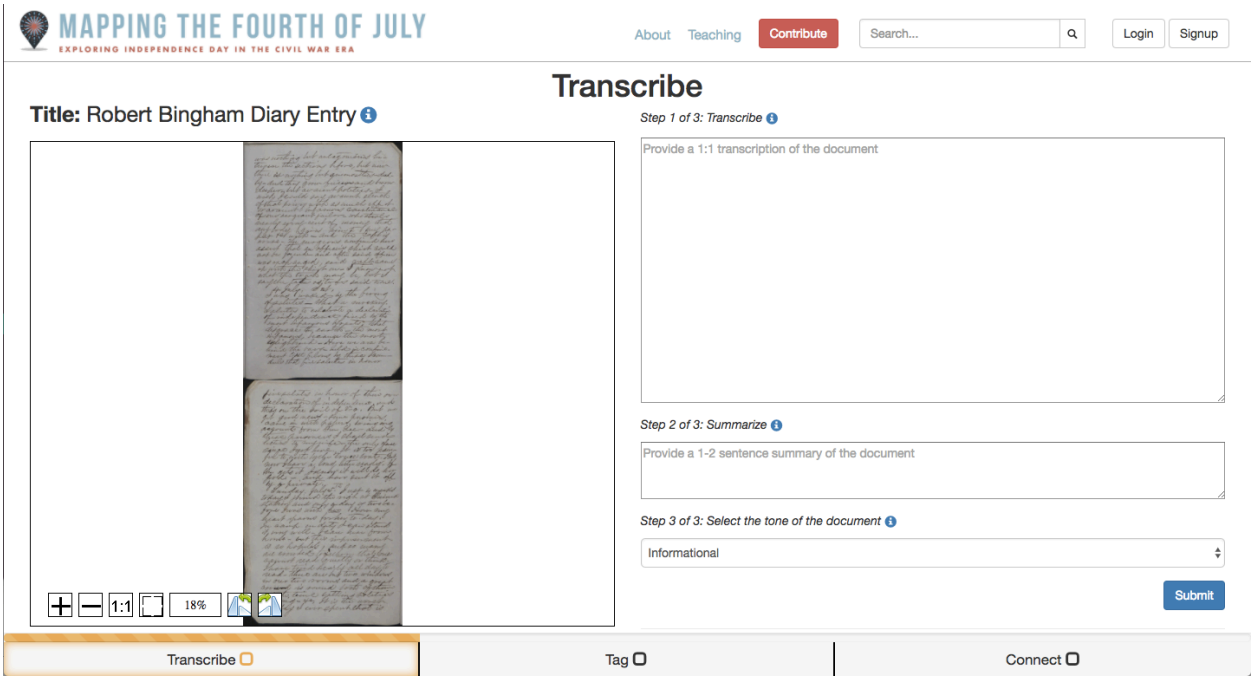


Figure 3.1: A screenshot of the transcribe task

feedback resulted in two features: 1) a step-by-step tutorial to guide the work process of each task and 2) group management for the teacher to track students' progress.

3.4 System: Incite

3.4.1 Transcribe

The main production goal of the transcribe task is to transform primary sources from images into texts which can be searched later and support further text analysis. The main education goal of the transcribe task is to examine the documentary aspects of the text such as type, subject and specific details of the historical document (a primary source) and this task corresponds to the first step, Summarize, of the SCIM-C strategy.

A screenshot of the transcribe task is shown in Figure 3.1. The left panel of the transcribe

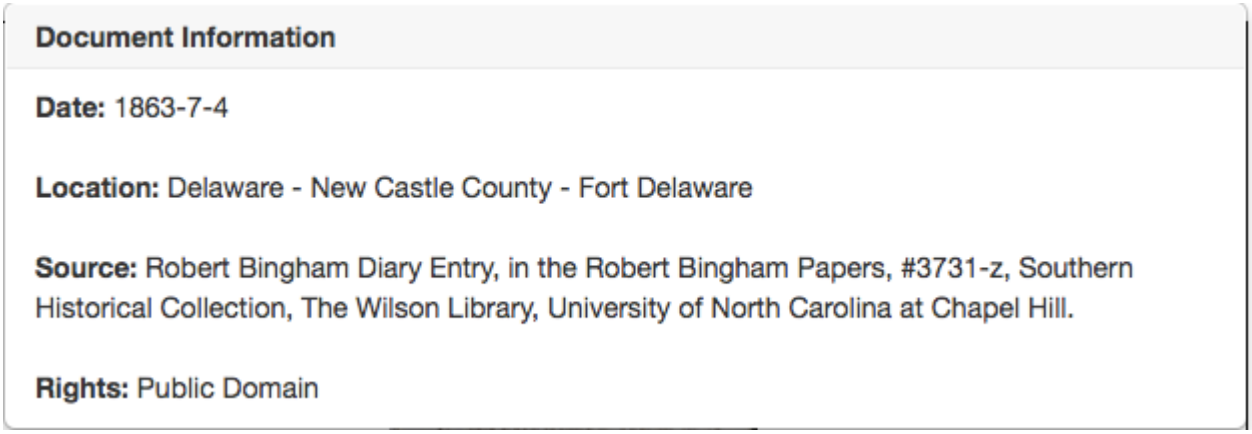


Figure 3.2: A screenshot of document information

task contains title, and other source metadata (via mousing over the information icon), and an image viewer for a document.

The metadata as shown in Figure 3.2 includes date, location, source and right information if available. The image viewer provides several widgets (e.g. zoom, rotate) to help the user to view the image document with a preferred resolution and orientation while transcribing. By default, the document is scaled to the size of the viewer. The right panel of the transcribe task is the working area that contains three steps (transcription sub-tasks) to complete the transcribe task. Step 1 is to transform a digitized image document into a searchable textual document by providing a 1 to 1 transcription. Step 2 is to provide a short (1-2 sentence) summary which can be useful when people browse and navigate a list of documents. Step 3 is to identify the tone for the document because tone helps convey an author's higher-level meaning and intent beyond the raw information in the summary. Six possible tones — Aggression, Anxiety, Informative, Optimism, Pride, and Sarcasm — are available, based on recommendations from the historians who constructed the initial Mapping the Fourth archive. Steps 2 and 3 are included here because this is the first crowdsourced task, so if the information is collected at this stage, it can be used for the rest of the process of searching, browsing and navigating. In addition, these two steps also help ensure the user reads the

Figure 3.3: A screenshot of the tag task

document and extracts important details from the document.

3.4.2 Tag

The main production goal of the tag task is to identify meaningful entities and entity types from the primary sources such as person, location, organization and so on so that this information can be used for clustering and analysis later. The main education goal of the tag task is to understand the context of the historical document such as when and where the source was produced and this task corresponds to the second step, Contextualize, of the SCIM-C strategy.

A screenshot of the Tag task is shown in Figure 3.3. As in Transcribe, there are two panels in the Tag task. The left panel of the Tag task contains the title, metadata and a two-tab document viewer. The default tab of the document viewer is the transcription of the target

document obtained from the Transcribe task. The transcription has been pre-tagged with a widely-used automated Named Entity Recognition (NER) tool [59] . The other tab of the document viewer is the original image of the document.

The right panel of the Tag task consists of three steps (sub-tasks) to complete the task. Step 1 shows tags that have been automatically recognized by the NER tool which recognizes tags along with their categories. The user is asked to verify if the automatically-recognized categories are correct for the tags, expand tags by adding information about subcategories and details, and delete tags that are incorrectly recognized. Step 2 is to add tags that are not recognized by the NER technique along with information about the tags. If the user mouses over a colorful tag on the left panel, the right panel will scroll to an appropriate position to show the corresponding tag on the right to help the user locate where the target tag is in the working area. The NER technique is introduced to save time for the user so that the user can focus on work that automated techniques cannot do well. Step 3 asks the user to contextualize the target document by identifying time and locations about the target document. In addition to the purpose of providing scaffolding, this step is also helpful for assessing how well the user learns the content and for collecting extra metadata for historical research use.

3.4.3 Connect

The main production goal of the connect task is to connect a primary source to related topics that interest historians so that historians may use these connections to find related primary source for their topics of interest. The main education goal is to make inferences and reflect from the historical document such as what is suggested by the source and the task corresponds to third and fourth steps, Infer and Monitor, of the SCIM-C strategy.

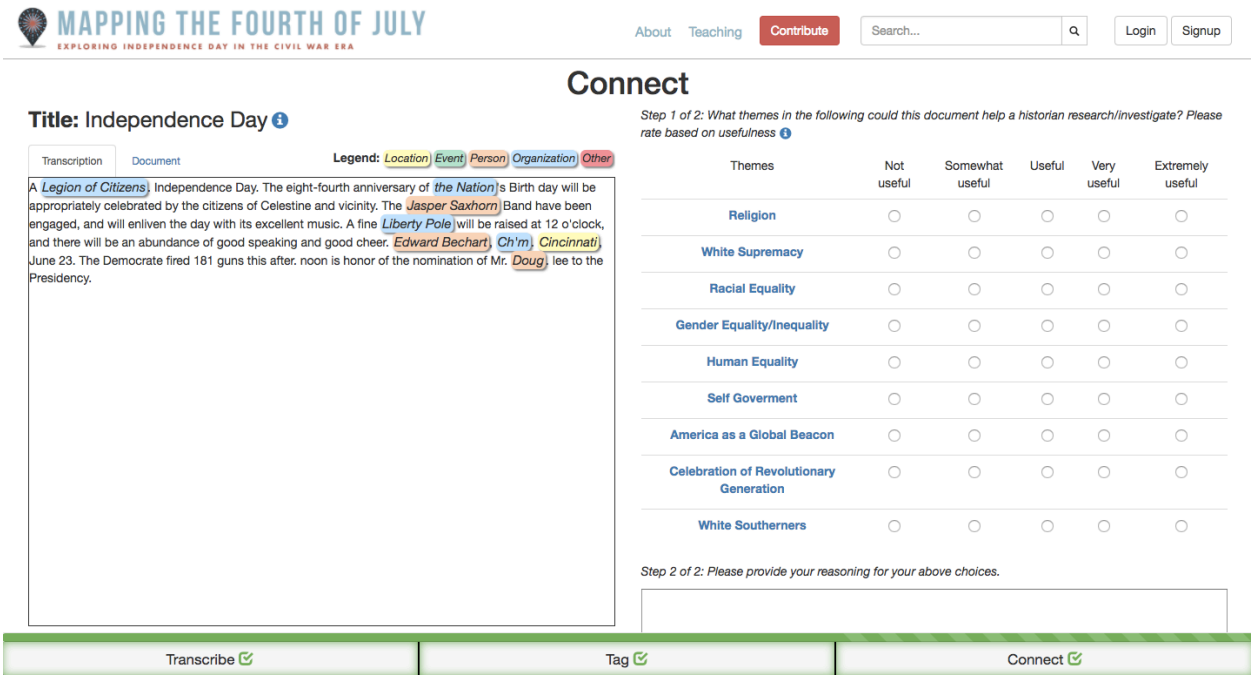


Figure 3.4: A screenshot of the connect task

A screenshot of the Connect task is shown in Figure 3.4. The left panel of the interface is the same as that of the default but the right panel includes two steps to complete the Connect task. Step 1 asks the user to rate the usefulness of different topics that could help a historian research and investigate in a 5-level rating scale. This multi-level scale is more appropriate for assessing how similarly an individual user and a historian think, instead of a binary decision. Step 2 asks the user to provide reasoning for the decisions. The reasoning could be a useful resource for history educators to understand the reasoning process of the user, especially when the ratings are unexpected. In addition, the prompt in Step 2 serves as a place for the user to reflect on the relationships between the document and inferences made in Step 1. The prompt also provides a place for the user to give alternative answers. The concepts in the Connect task are customizable via the default Omeka setting page. That is, the user can add/remove/modify concepts that will appear in the Connect task.

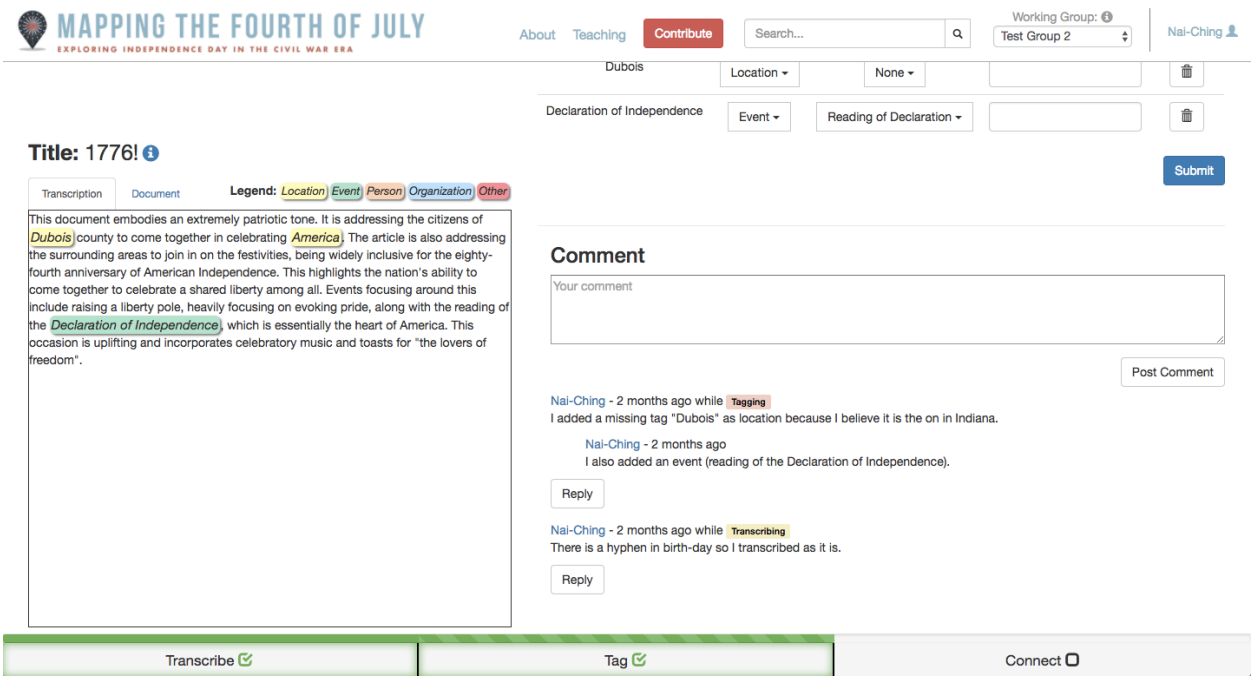


Figure 3.5: A screenshot of the comment feature in the tag task

3.4.4 Comment

The Comment feature is supported in within-document tasks (i.e., transcribe, tag and connect) to provide a place where the user can discuss about the document, especially about the task the user is working on. The Comment feature is only open for users who have logged into the system to prevent potential malicious use. There is a comment section at the bottom of the working area of the within-document tasks. An example Comment section of the Tag task is shown in Figure 3.5. There is also a label for each comment to show during what task the comment is made to provide more context about the comment.

3.4.5 Discussion

While the previous three crowdsourced tasks (transcribe, tag and connect) are within document participation, Discussion provides a way for users to participate in a between documents

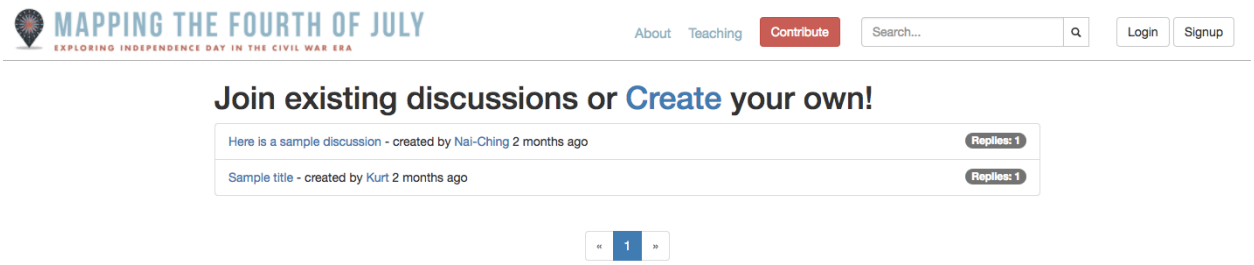


Figure 3.6: A screenshot of the discussion list

way. This also serves as the Corroborate step in the SCIM-C strategy for between-documents analyses where the user can start and/or participate in a discussion topic by analyzing multiple sources and receiving feedback from others. A user can view existing discussions, participate in a specific discussion and create a new discussion. As shown in Figure 3.6, the user can see all current discussions in a list view sorted by hotness (i.e., number of replies). If the user is interested in a discussion, the user can click on the discussion and view details on the discussion page as shown in Figure 3.7.

Similar to other crowdsourced tasks, there are two panels in a discussion page. The left panel is a tabbed document viewer (similar to other crowdsourced tasks) to show documents related to the discussion. There are tabs both at the top and the bottom of the document viewer. The top tabs of the document viewer are used to switch between image and transcription views of the document if the transcription of the document is available. The bottom tabs of the document viewer contain up to three most-recently-opened documents and are used to switch between different documents. The right panel includes title, comments or replies, and related documents of a discussion. The user can mouse over a document icon to see its title and summary. If the user clicks on a document icon, the document will be opened in the document viewer on the left panel. The user can also participate in the discussion by replying to the discussion.

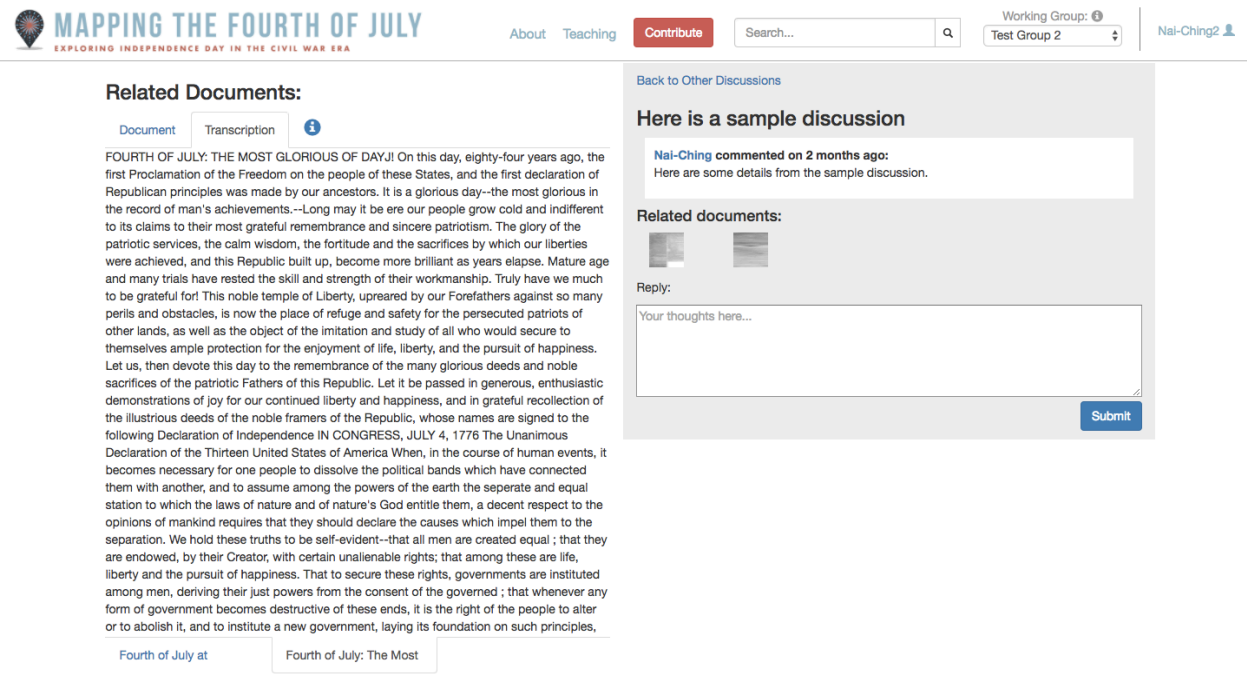


Figure 3.7: A screenshot of the discussion page

The user can also create a new discussion by clicking the link “Create” on the Discussion List page as shown in Figure 3.6. The create discussion page is shown in Figure 3.8. Similarly, there are two panels on the page. The left panel is used for searching related documents. The user can use keywords to search documents and attach documents as references to the discussion by selecting the checkboxes and clicking the “Add Selected as Reference(s)” button. By mousing over the string, “(summary)”, beside a document, the user can see the summary of the document like the one in Figure 3.8. The right panel provides places for the user to fill in the title and content of the new discussion. The user also sees documents that have been added as references and is able to remove references with the “Delete Selected Reference(s)” button

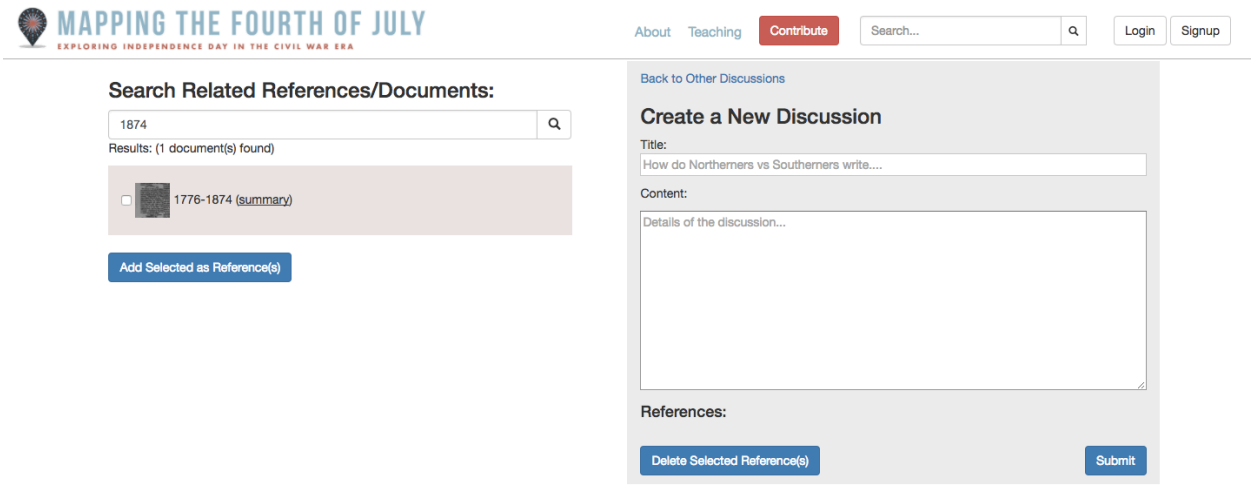


Figure 3.8: A screenshot of the create discussion page

Figure 3.9: A screenshot of the search box on the navigation bar

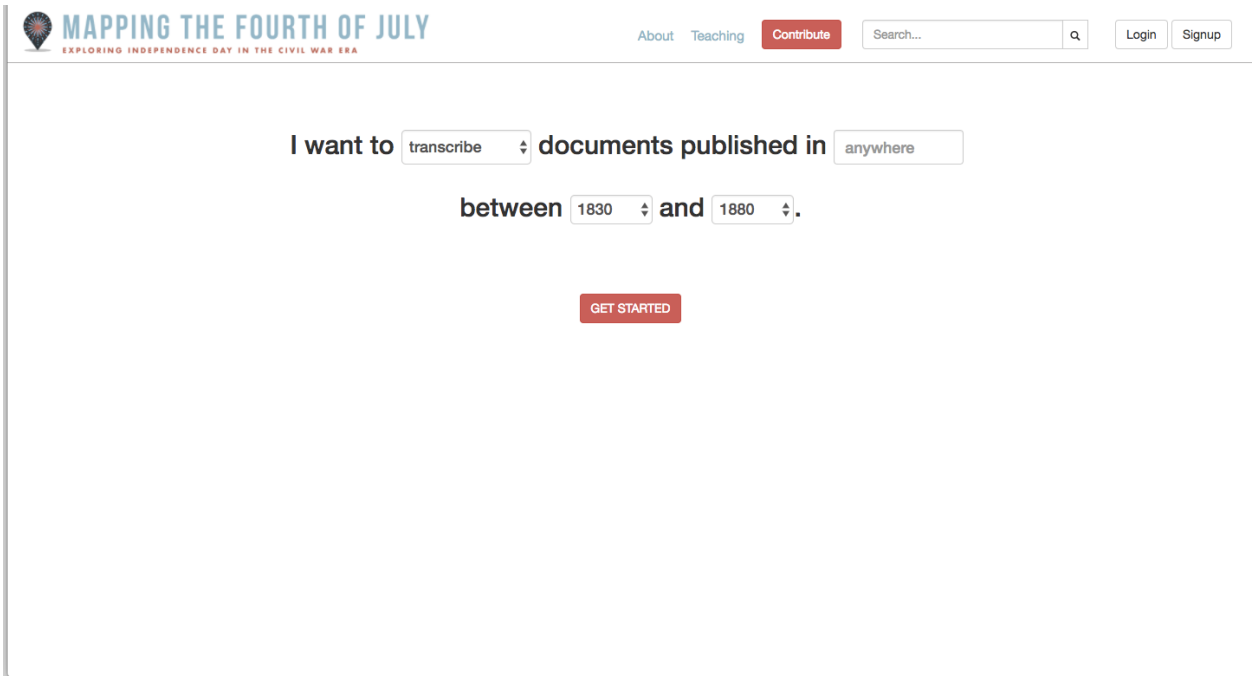


Figure 3.10: A screenshot of the contribute page

3.4.6 Search

Incite provides two ways for the user to search for documents of interest. One is the search box on the navigation bar as shown in Figure 3.9. This provides the user a quick access to the search function because the search box stays on the navigation bar on all pages. The search function is keyword-based. In addition, once the search box is focused or moused over, a filter dialog appears with default, or previously-used if available, parameters and provides more options for the user to specify more parameters such as task type, location and date range. This interface is designed for frequent users or users who are comfortable with (advanced) search interfaces. The other way to search is through the red "Contribute" button on the navigation bar to land on Contribute page as shown in Figure 3.10. This interface arranges search options into a simple and full sentence so that these search options are clear and meaningful for first-time users or users who are less comfortable with advanced search interfaces. The default search options may change based on user's information or the

need of the system. For example, the default location can be automatically set to user's location by using IP address geolocation services and the default task type can be set to be the most needed one. The major difference between the two ways of search is that the search box on the navigation bar allows keyword search and search for all types of tasks, while on the Contribute page, the two features are not available because they are not as appropriate in the context of making contributions.

3.4.7 Search results

An example of a search result pages is shown in Figure 3.11. There are two panels on a search result page. The left panel contains a zoomable map view of the search results while the right panel includes a list view of the search results and pagination if appropriate. Each pin on the map corresponds to a document and each item on the list view also corresponds to a document. Each item on the list includes a thumbnail, title, year, location and work status of a document. Three glyphicons show the work status of a document including glyphicon-pencil as "need to be transcribed", glyphicon-tag as "need to be tagged", and glyphicon-tasks as "need to be connected". The black version of a glyphicon for a document means the corresponding task of the document has been worked on before, while the gray version of the glyphicon means the task of the document is still available. Once a pin on the map or an item on the list is moused over, more details (in a popover fashion) about the item show up both on the map and beside the list to help the user match the same document on the two views as shown in Figure 3.11

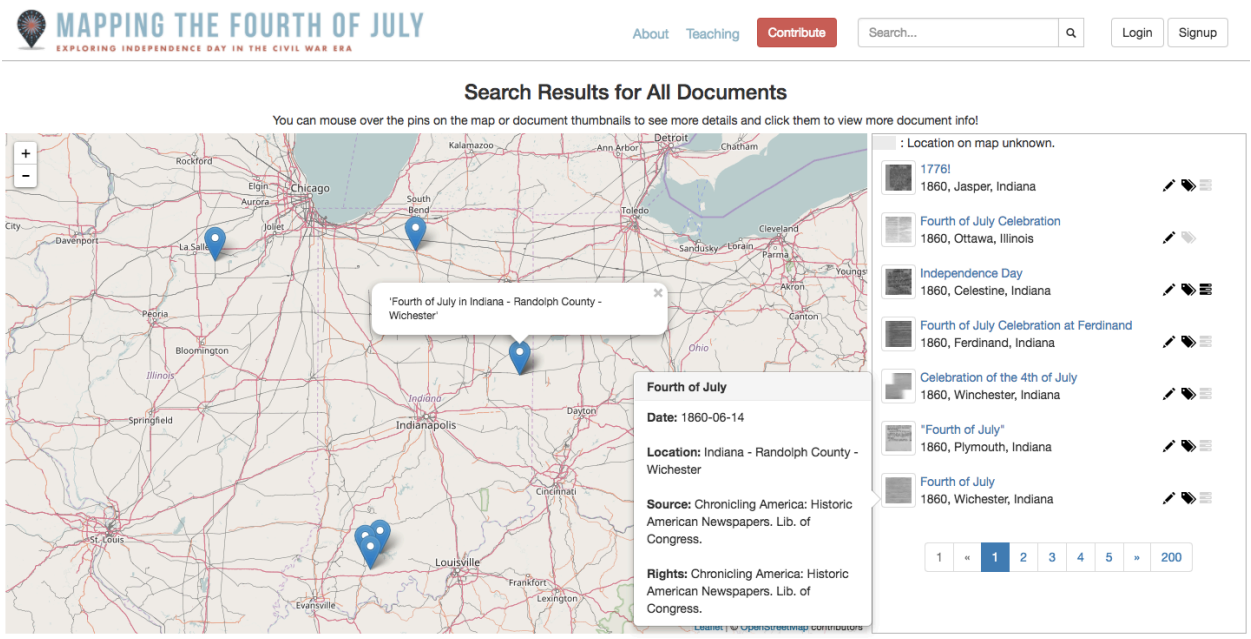


Figure 3.11: A screenshot of the search results page (all task type)

3.4.8 User account

A user can create a personal account anytime by clicking the "Signup" button on the navigation bar and entering required information. The user can also log in his or her account anytime by clicking the "Login" button on the navigation bar and entering required credentials. Once the user is logged in, the user's group information and name appear on the rightmost of the navigation bar as shown in Figure 3.12 and by mousing over the name, the user can edit the profile, manage group information and view activities. The group feature will be fully discussed in a later section (Chapter 3.4.9). On the Activity page as shown in Figure 3.12, the user can view all documents and tasks the user has contributed to. Activity Overview provides basic contribution numbers for documents under each task or activity. Activity Feed provides a detailed list of the user's contribution history. The user can also click on the four tasks under Activity Overview to filter the Activity Feed based on task type.

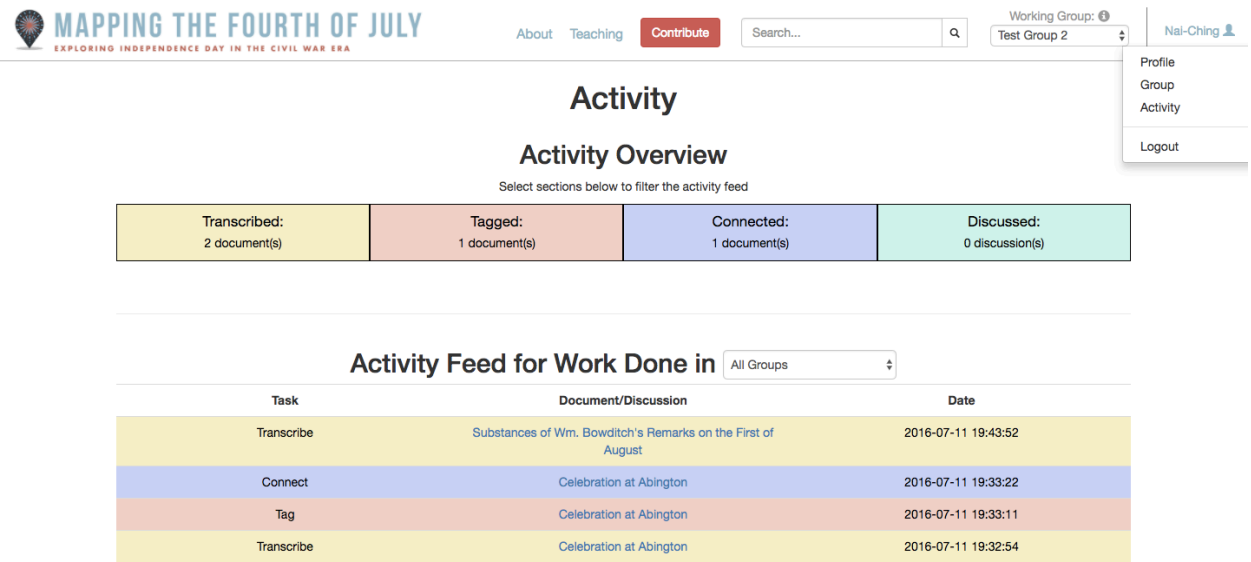


Figure 3.12: A screenshot of the user activity page

3.4.9 Group and class

The group feature of Incite is designed to provide support for collaboration and administration of group work such as classroom use. The user can participate in multiple groups with approval from owners of those groups. The user can also create multiple groups and invite other users. On the Group page shown in Figure 3.13, the user can search existing groups or create a new group. By clicking on a group name from the search results, the user is redirected to the homepage of the group (as shown in Figure 3.14) where the user can see overview information about that group and send a join request to the owner of that group if the user is not yet a member of that group.

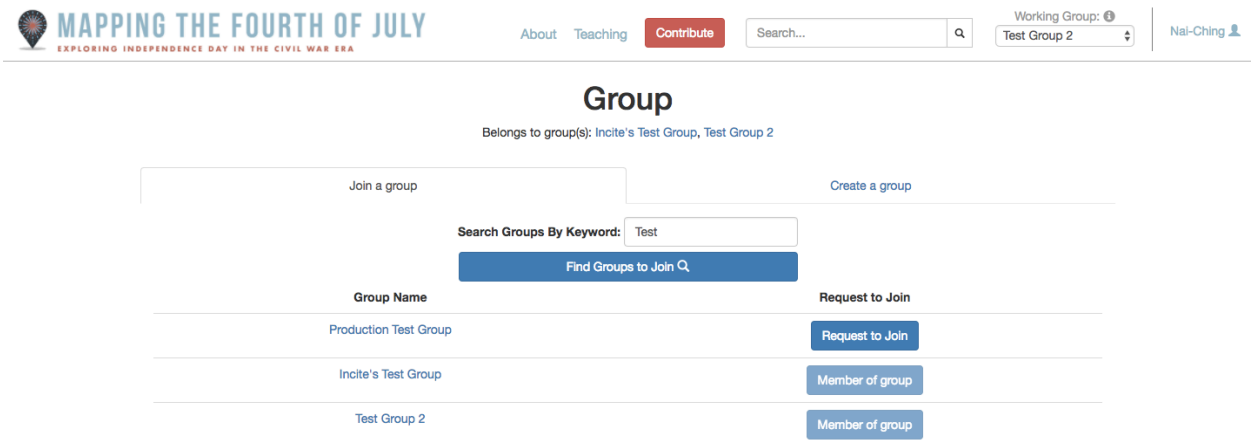


Figure 3.13: A screenshot of the group search page

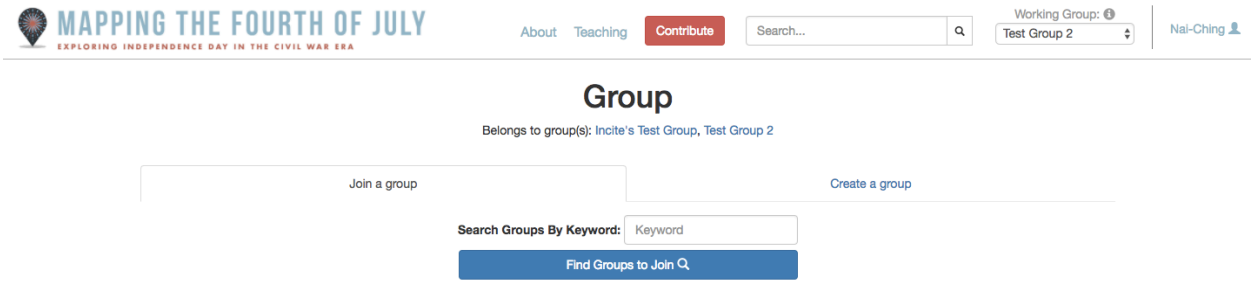


Figure 3.14: A screenshot of a group's home page

Group work

After joining groups, the user can define which group a contribution is made to by checking into a group. Once the user is logged in, on the navigation bar, there is a group section telling the user which working group is currently selected, as shown on the top right of Figure 3.5. Once a group is selected, all the contributions the user make will be associated with the group. The user can later check group activities by selecting a particular group on the Activity page shown in Figure 3.12. The user can also check out of the selected working group by selecting "Reset Working Group" from the dropdown menu on the navigation bar.

Group management

The user can also create a group on the Group page by entering a unique group name and will be redirected to the Owner's Page of the group as shown in Figure 3.15. On the Owner's Page, the user can invite new members by sending an invitation email to potential members, set customized group instructions that will be displayed to all group members when they check in, view group activity via an overview of each member's activity in the group, and manage group members

3.4.10 Deployment

I deployed Incite with two classroom-based case studies, Mapping the Fourth of July (M4J) and The American Soldier (TAS). I evaluated these deployments with a mixed-methods approach. First, I gathered and analyzed qualitative feedback from the instructors (historians and educators) who used Incite in their classrooms (Section 3.5). Second, I quantitatively analyzed the data produced by crowd workers (primarily students) who used Incite during these deployments (Section 3.6).

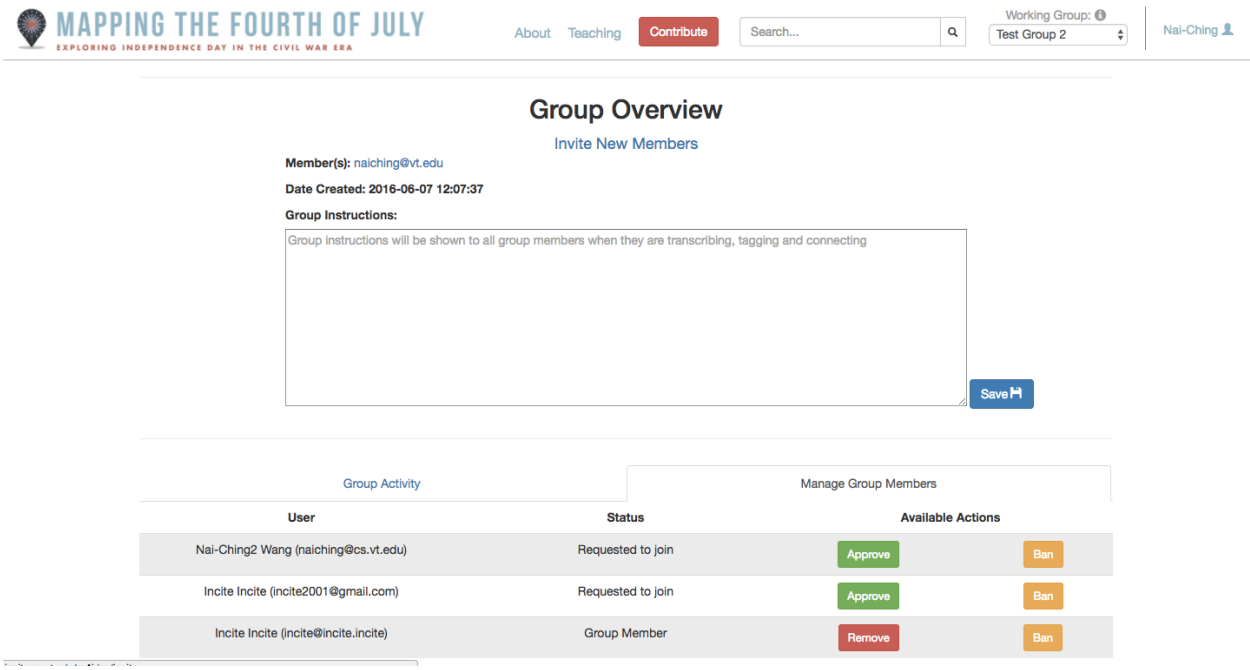


Figure 3.15: A screenshot of a owner's page of group

3.5 Evaluating Incite: Instructor's Perspective

3.5.1 Mapping the Fourth

Participants

The first case study for Incite was historian Paul Quigley's "Mapping the Fourth of July in the Civil War Era," (M4J) an archive of more than 4,000 American Civil War-era primary sources such as newspapers, diaries, and letters. I deployed M4J in 12 classrooms at various schools including Virginia Tech, University of Central Arkansas, University of Alabama, Blacksburg High School, University of Georgia, DePaul University, Bedford County (VA) public school, Arlington County (VA) public school and eventually online.

Method

After the classroom use, each instructor was asked to fill out a survey about how they use Incite in their classrooms in terms of the following topics. The classrooms included middle school, high school and college/university levels.

Results

How Incite was used in class Incite was used in various ways including extra credit, part of final exam to transcribe a manuscript collection on Incite, regular assignments, semester-long engagement and demonstration, data sources for class activities, discussions and comparative reports.

Some sample responses include:

"I had students complete a transcribe and tag assignment as well as use three documents for a comparative paper on the changing meanings of the holiday for two different groups of Americans." - HG.

"I offered the students extra credit for working through a source from transcribing, tagging, and contextualizing. They were then required to write a couple of paragraphs letting me know about not only the source, but also the experience. ..." - EL.

Ease of integrating with curriculum and standards All college/university instructors indicate that it's easy to do so because they generally have more flexibility deciding their own curricula.

In general, middle/high school teachers were able to use Incite in their classes. Some say Mapping the Fourth and Incite fit their curricula and standards while some mentioned some

extra effort and stretch were required to do so. The match between class schedule, content of primary sources and site availability was key to the integration.

Other potential uses Most of the instructors indicate that Incite would be useful for a variety of classes and activities such as enrichment of experiences for students, methods classes (how historians work), topical classes (due to the available sources on Incite), developing documentaries or digital exhibit presentation.

Understanding of primary sources For college level students, instructors reported that the system provided a nice introduction to primary sources and laid the foundation for in-class research. Students generally enjoyed using primary sources because they are real pieces of history.

For middle and high school level students, while the use may be somehow limited possibly due to the inflexibility of curricula, the system helped provide more insights and diversity of points of view. These insights may spark curiosity and lead to more meaningful class discussions.

A few sample responses include:

"Loved the diversity of sources. POV was very evident and drove some great class discussions" - BK.

"I think it was quite effective in this respect" - MS.

Things that worked well and future improvements Things that worked well include the intuitive interface and task designs for transcribing and tagging.

Suggested improvements include more flexibility in transcribing and tagging such as font

types in transcriptions and more categories in tagging. More specific and clear instructions are needed.

Suggested features include a folder to save works-in-progress, undo activities, more guidelines for transcriptions, and more bibliographic information of the sources. Adding primary sources and creating "playlists" of primary sources are also desired.

Willingness to use in the future Only one of the teachers said "probably not" due to the instability of first version. Others show their interest in using Incite with future classes, especially higher-level, methods classes.

3.5.2 The American Soldier

3.5.3 Participants

For a second case study, historians Ed Gitre and Bradley Nichols used Incite for a World War II transcription project, "The American Soldier", with 4 history courses at Virginia Tech.

The process of how Incite was used in these classes was as follows

1. proofread existing results from previous batch if any,
2. work on some assigned documents,
3. peer-review results from 2,
4. work on a new set of documents.

Methods

After the classroom use, instructors were asked to provide feedback about how they use Incite in their classrooms in terms of the following topics.

Results

Ease of use In general, it was easy for the students to follow.

Some sample responses include:

"The students' reactions were uniformly positive. They enjoyed the assignment as a whole, and found the software both helpful and easy to navigate." - BN.

"A central digital platform to engage students with primary sources with having to do anything on paper." - BN.

Usefulness in teaching history The instructors liked Incite and thought Incite was an effective tool to help them teach history.

A sample response was *"The site was incredibly effective in teaching students to identify significant historical patterns by parsing and comparing first-hand accounts"* - BN.

Things that worked well and future improvements The major complaint was about the dual logins with Omeka and Incite. Due to the private nature of the collection, a student needed to log into Omeka to have access to the collection and log into Incite to keep track of their progress for the classroom use. It was confusing for users especially first time users.

3.6 Evaluating Incite: Crowdsourced Production

3.6.1 Descriptive statistics

In the Mapping the Fourth case study, the crowd helped transcribe more than 1,000 documents; add more than 7,000 tags; and make more than 4,000 connections to help historians conduct research. In the American Soldier case study, the crowd helped transcribe more than 3,000 documents; add more than 3,000 tags; and make more than 29,000 connections to help historians conduct research.

3.6.2 Analyzing quality of Mapping the Fourth data

To see how the collected crowdsourced work can be used for historical research, I closely worked with historian Dr. Paul Quigley using data collected in the Mapping the Fourth project.

To assess crowd's performance, my evaluation was two-fold: 1) crowd vs. expert and 2) crowd vs. crowd. Crowd vs. expert was used to measure the quality of crowdsourced production. Crowd vs. crowd was used to measure the consistency of crowdsourced production.

Due to the large amount of collected data, I focused on a representative subset for more detailed analysis. I first selected documents with more than three submissions for a task (whenever possible) so that we can have both results for the two measures.

One hand, more duplicate submissions per document can provide more accurate crowd vs. crowd results, but that limits the number of qualified documents. On the other hand, the smaller the number of submissions is, the more documents can be used, but we might not be able to assess crowd vs. crowd results. After discussing with the expert, I determined that

Doc	Doc. Len. (Char)	Distance 1	Distance 2	Distance 3	Avg. Distance/Char
1	640	32	15	0	0.02
2	379	148*	24	3	0.15
3	652	52	89	2	0.07
4	159	6	1	0	0.01
5	3364	94	66	66	0.02
6	293	2	2	2	0.01
7	790	59	52	73	0.08
8	619	72	73	0	0.08
9	1854	13	16	16	0.01
10	2208	2	2	0	0.00
Overall	10958	480	340	162	0.03

Table 3.1: Crowd vs. expert: edit Distance between expert's and crowd's transcriptions

three seemed to be a good trade-off.

There were 17 documents with three or more submissions for the Transcribe and Tag tasks. I randomly chose 10 of them for further analysis. For the Connect task, there were not enough documents with multiple crowd submissions for a crowd vs. crowd evaluation, so I randomly chose 10 documents out of 87 documents (each with one or more submissions per document) for further analysis.

In the following analysis, I will present the results in terms of types of production tasks, including transcription, tone, summary, tag and connection.

3.6.3 Transcription

To assess the quality of a transcription submission, I calculated edit distance with the commonly-used Levenshtein distance.

The results of crowd vs. expert are shown in Table 3.1. I used the expert's transcription as the gold standard and calculated document length for each document in terms of number of characters.

Doc	Doc. Len. (Char)	Distance 1: Crowd 1 vs 2	Distance 2: Crowd 2 vs. 3	Distance 3: Crowd 1 vs. 3	Avg. Dist./Char
1	640	33	14	32	0.04
2	379	159	8	154	0.28
3	652	110	90	52	0.13
4	159	5	1	6	0.03
5	3364	287	0	287	0.06
6	293	0	0	0	0.00
7	790	1	31	32	0.03
8	619	7	72	73	0.08
9	1854	3	0	3	0.00
10	2208	0	2	2	0.00
Overall	10958	605	218	641	0.04

Table 3.2: Crowd vs. crowd: edit distance among the crowd's transcriptions

As we can see from the results of crowd vs. expert, the average edit distance per character (approximation of error rate) is about 3%. A few of the transcriptions achieve perfect match with the expert's transcriptions. There was one transcription with extremely large edit distance (39% distance per character) for document 2. It turned out the digitized image contained some fraction of another document so the worker transcribed that fraction as part of the transcription for the document.

As for the results of crowd vs. crowd shown in Table 3.2, the average edit distance per character (approximation of error rate) is about 4% similar to crowd vs. expert. The error rate of another crowdsourcing project about US Census data was about 16-17% for more variable texts such as names and was less than 5% for less variable texts such as marital status and gender [68]. The documents from M4J project contain variable texts so 3-4% error rate seems to be very good comparing to prior work. Prior work also showed that the word error rate was about 22% for OCR for the 19th Century Newspaper project [131]. The results indicated that crowdsourcing-based method could be a good source to train automated OCR techniques.

Doc	Expert	Crowd 1	Crowd 2	Crowd 3	Hit Rate	RAI
1	Pride	Pride	Optimism	Optimism	0.33	0.33
2	Informational	Informational	Informational	Informational	1.00	1.00
3	Informational	Informational	Informational	Pride	0.66	0.33
4	Informational	Informational	Informational	Informational	1.00	1.00
5	Optimism	Pride	Optimism	Informational	0.33	0
6	Anxiety	Informational	Anxiety	Anxiety	0.66	0.33
7	Aggression	Informational	Informational	Aggression	0.33	0.33
8	Pride	Informational	Informational	Pride	0.33	0.33
9	Anxiety	Informational	Informational	Informational	0.00	1.00
10	Optimism	Aggression	Aggression	Aggression	0.00	1.00
Overall					0.47	0.3

Table 3.3: Crowd vs. expert (hit rate) and crowd vs. crowd (RAI) for tone selecting

3.6.4 Tone

To evaluate results of tone of crowd vs. expert, I used hit/miss since the user was asked to pick most appropriate one out of a list of six potential tones. The expert provided the gold standard tones for each document. For crowd vs. crowd, I used raw agreement indices (RAI) to calculate the agreement among multiple crowd workers. An RAI value ranges from 0 (no agreement) to 1 (perfect agreement).

Both results of crowd vs. expert and crowd vs. crowd are shown in Table 3.3. The Informational tone was the most common choice among both the expert and the crowd, whereas the Sarcasm tone was never chosen by either group. The average hit rate is about 47% which means that the crowd chooses the same tone as the expert (out of total six available tones) 47% of the time. While these results suggest the crowd is not a very reliable judge of document tone, their hit rate is nearly three times better than chance (17%), and may be sufficient for certain use cases, such as providing a rough first pass on the data.

For both documents 9 and 10, we can see that the hit rate and RAI value are at two opposite extremes. While none of the crowd had the right answer (hit rate = 0), the RAI

value shows there is a perfect agreement among the crowd. In other words, all of the crowd workers agreed with each other for these documents, yet none of them agreed with the gold standard answer. These results suggest at least two follow-up actions: either that this document presents a common confusion that history educators could prioritize for a high-impact learning intervention, or the gold standard itself may be worth reconsidering. This phenomenon is also observed in the preliminary study of Chapter 4 and more discussion can be found in that Chapter.

3.6.5 Summary

To measure the quality of each crowd-generated summary, I graded them with the rubric developed from previous work and guidelines gathered from school writing centers, and approved by a history professor, Historian B (see Appendix B for details). Quality was divided into three categories: low (0–3), medium (4–6), and high (7–10) based on a 10-point scale. A high-quality summary contains no or minor issues, and these do not affect reading. A medium-quality summary misses some important information, detail or context. A low-quality summary misses substantial important information, detail and/or context.

The quality results are shown in Table 3.4. The results show that the average quality score across all documents is 5.9, which can be interpreted as on the threshold between medium and high quality. Additionally, the average quality per document was always medium or high.

Individual summaries were also almost always medium or high quality. However, there was a 0 score summary for document 9. After investigating, it was caused by a nearly empty summary. This shows that while most of the crowd workers (students) are diligent, there are still occasional bad submissions, some of which could be prevented with automated quality

Doc	Crowd 1	Crowd 2	Crowd 3	Average	SD
1	7	3	3	4.3	2.3
2	8	8	8	8	0.0
3	8	5	5	6	1.7
4	8	6	7	7.0	1.0
5	3	7	7	5.7	2.3
6	5	5	5	5	0.0
7	6	6	6	6.0	0.0
8	8	3	6	5.7	2.5
9	7	6	0*	4.3	3.7
10	7	7	7	7	0.0
Overall				5.9	1.4

Table 3.4: Summary scores (0-10 scale)

control measures (e.g. minimum length requirements). The standard deviation is about 1.4 that indicates the quality is relatively stable ranging less than half of the range of a quality bin.

3.6.6 Tag

To evaluate the results, I used precision and recall for crowd vs. expert and Fleiss' kappa for crowd vs. crowd. The expert provided gold standard tags for each of the documents.

The results of tags are shown in Table 3.5. Average precision and recall for crowd vs. expert are 0.87 and 0.79, respectively. The results shown that the crowd does not provide many inappropriate tags (with 0.87 precision), and captures a large portion of the appropriate tags (with 0.79 recall). For the results for crowd vs. crowd, the average Fleiss' kappa value is 0.77, which is considered to be substantial agreement [86]. This agreement level suggests that different crowd workers tend to provide the same tags for a given document.

Doc	Crowd 1		Crowd 2		Crowd 3		Avg.		Fleiss
	Pre.	Rec.	Pre.	Rec.	Pre.	Rec.	Pre.	Rec.	Kappa
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.0	0.57	0.78	1.0	0.78	1.0	0.85	0.86	0.58
3	1.0	0.93	1.0	0.93	0.93	1.0	0.98	0.95	0.91
4	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
5	1.0	0.38	1.0	0.31	0.93	0.93	0.98	0.54	0.4
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.91	1.0	0.91	1.0	0.91	1.0	0.91	1.0	1.0
8	1.0	0.5	1.0	0.75	1.0	0.75	1.0	0.67	0.77
9	1.0	0.88	1.0	0.88	1.0	0.88	1.0	0.88	1.0
10	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Overall							0.87	0.79	0.77

Table 3.5: Crowd vs. expert (precision and recall) and crowd vs. crowd (Fleiss' kappa) for tagging

3.6.7 Connection

To assess the quality of connections, I used weighted Cohen's kappa because the relevance rating is provided in Incite using a 0–4 Likert scale. The weighted kappa means that scores that are further away from the expert's are penalized more. The concepts used in the Connect task were previously provided by the expert. The expert provided gold standard connections for each document.

Due to the lack of redundant submissions, there is no crowd vs. crowd analysis for this type of crowdsourced production.

The results of quality of connections are shown in Table 3.6. The average kappa is 0.54 which is considered to be moderate agreement between the expert and the crowd [86].

Doc	Weighted Cohen's Kappa
1	0.60
2	0.33
3	0.62
4	0.25
5	1.0
6	0.43
7	0.0
8	0.6
9	0.74
10	0.83
Overall	0.54

Table 3.6: Crowd vs. expert: agreement between expert and the crowd

3.7 Discussion

3.7.1 Incite

Incite seems to be more appropriate for focused, higher-level, methods classes that involve more in historical thinking and comparative work. While Incite may spark insights, engage students with digital activities, it may not be suitable for some classes with strict curricula and standards.

While many of the suggested improvements have been addressed through the iterative design process, there are still some features that will be helpful. Saving temporary work, allowing undo's and customizing primary sources are the most desired features.

3.7.2 Class-sourcing model

Based on the case studies, Incite seems to be a quite useful tool for history teachers to use in the classroom. We can also see that Incite has helped analyzed thousands of documents

with transcripts, tags and connections that can be used for historical research.

Due to its web-based nature, Incite may also be useful for online classes (e.g., MOOC) to support more students at the same time without physical constraints like traditional classrooms.

3.7.3 Opportunities for historical scholarship

Incite demonstrates that the class-sourcing model can help teacher historians produce valuable results (i.e., searchable transcriptions, tags and connections), saving much time that would be spent by themselves.

3.7.4 Opportunities for history education

In the two case studies with 15 classes, Incite demonstrated that the effective design building on the class-sourcing model can help students learn history via interacting with real pieces of history (primary sources). It also helps teacher historians to teach historical thinking and methods.

3.8 Chapter Summary

In this chapter, I described the idea of class-sourcing model, the design process of Incite and various features of Incite. By reviewing the literature and discussing with experts, I concluded several crowdsourced tasks that would collect useful data to support historical research. Unlike general crowdsourcing, to be useful in a classroom, some features are also very important such as group management because they will help instructors keep track of

students' progress.

The results of Incite's deployment in 15 classrooms showed that the class-sourcing model was effective in terms of collecting useful data for historical research and helping historians teach their students in the classroom.

In the next chapter, I build on these results and expand student crowds to general crowds.

4

RAP: Scaling up Crowdsourced Historical Connections

This chapter is based on work published in CHI [136] and Human Computation Journal [137].

4.1 Motivation and Research Question

While the class-sourcing model with Incite worked for many educator historians, some historians did mention some limitations of this model. For example, an existing curriculum may limit how teachers can use this model. In addition, the classroom-focused intervention limits opportunities for public history, which plays an important part in many historians' roles. Finally, bringing crowdsourced history beyond the classroom enables scaling up of document analysis for greater productivity.

To overcome these limitation, I expanded the class-sourcing model to a more general crowdsourcing model, focusing on higher level crowdsourced tasks beyond transcription. Based on the literature, historians spend significant time evaluating the relevance of primary sources that they encounter in digitized archives and through web searches. One reason this task is time-consuming is that historians' research interests are often highly abstract and specialized. These topics are unlikely to be manually indexed and are difficult to identify with

automated text analysis techniques. This *connect* task of Incite seems to be an appropriate candidate as a high-level task.

To expand the class-sourcing model, I asked the following research questions:

RQ 2a: How well does the novice crowd make connections?

RQ 2b: How can crowds connect related primary sources to scholarly topics as accurately as historians?

RQ 2c: How can we identify opportunities for public history intervention?

4.2 Method

To understand a quality baseline for crowdsourced contributions in the domain of history, I conducted a preliminary study and compared this baseline with two popular readying comprehension techniques, underlining and summarizing. Based on the results of the preliminary study, I then developed RAP that can be used to improve quality of crowdsourced connections.

4.3 Preliminary Study

4.3.1 Dataset and historian

The documents used in this study come from a digital archive1 of around 189 digitized historical primary sources (personal diaries and letters, newspaper articles, and public speeches) from the American Civil War era (ca. 1840-1870). This archive was assembled by a tenured professor of Civil War history at our institution, whom I refer to as Historian A, for a prior

research project. Historian A generated a list of six topics of interest, related to Independence Day celebrations, that he looked for in the archive. I used a subset of these documents and topics for this study, as detailed in Chapter 4.3.4.

4.3.2 Apparatus and procedure

The experiment was conducted entirely online. After completing an online IRB-approved consent form, each participant was randomly assigned to one of three conditions corresponding to one of the three semantic tasks: reading, keyword (underlining), or summary (summarizing). The participant was also assigned a topic and a document. The participant then used the web interface I developed, based on a few alternative designs in pilots, to complete a three-step process.

First, the participant filled out a short quiz in which they matched their assigned topic to its correct definition. This ensured all participants understood the topic's meaning.

Second, the participant viewed two correct examples of connections between their topic and relevant documents (provided by Historian A). Our pilots and recent work on crowd innovation [152] both suggest that by viewing good examples, people can better understand abstract concepts and analogies. The participant also practiced their assigned semantic task on these examples. The reading task involved simply reading the example documents. The keyword task involved reading the documents and selecting 4-8 important keywords or phrases for both. The summary task involved reading the documents and writing a 1-2 sentence summary for both.

Third, the participant completed the semantic task on a new document. After completing the task, the participant decided whether it was relevant to the assigned topic by clicking "Yes" or "No" and typing in a brief justification of their decision.

Figure 4.1: User interface with summary condition

A screenshot of the Summary condition is shown in Figure 4.1

4.3.3 Participants

I used Amazon Mechanical Turk to recruit novice crowd workers. I restricted to US-only workers to increase the likelihood of English language fluency, with a 95% HIT (human intelligence task) minimum acceptance rate and 50 or more completed HITs. I recruited 120 workers and randomly assigned 40 to each of the three conditions. Each worker was unique and assigned to only one HIT to ensure that the required expertise was learned within that HIT. Thus, there were five unique workers for each combination of condition (semantic task), document, and topic. I paid participants $7.25/hour based on average task times in pilots. I also paid them a 20% bonus payment if they provided a reasonable justification for their decision, even if it was wrong.

4.3.4 Experimental Design

This was a between-subjects design with one independent variable (semantic task), two covariates (topic and document), and three dependent variables (quality, agreement, and efficiency).

Independent Variable

The independent variable, semantic task type, had three levels: reading, keyword, or summary. Therefore, the experiment had three conditions.

Covariates

I controlled for two covariates: topic and document. The complexity of the topic is likely to affect crowd performance, so I selected four diverse topics — Revolutionary History and Ideals, American Nationalism, American Hypocrisy, and Anxiety — from the list generated by Historian A. Document complexity could also affect crowd performance. Therefore, I randomly selected documents that were similar in terms of length (measured by word count) and readability (college-level, according to Flesch-Kincaid readability tests). I selected two documents for each topic, one highly related and one unrelated (as judged by Historian A), for a total of eight documents. None of the documents contain the topic name verbatim.

Dependent variables

To measure quality, I compared how each worker's responses compared with gold standard responses provided prior to the study by Historian A. Specifically, I measured the accuracy, precision and recall of the connections made by the crowd, i.e. whether they indicated a document was related or unrelated to their assigned topic. I measured accuracy as the ratio of matching connections (between the crowd and Historian A) to total connections made by the crowd. I measured precision as the ratio of number of matching connections to the total crowd connections in a specific condition. I measured recall as the ratio of number of matching connections to Historian A's total connections in a specific condition.

I also measured agreement among the five workers assigned to each condition. This metric provides an indicator of reliability for crowd workers and identifies areas of confusion as potential teaching opportunities. I used two measures of agreement, Raw Agreement Indices (RAI) and Fleiss' κ. The latter provides overall agreement and there is some established interpretation for its values. In addition to overall agreement, RAI also allows finer-grained

		Topic 1		Topic 2		Topic 3		Topic 4		Avg.
Document	Condition	1	2	3	4	5	6	7	8	
Related? (Historian A)		N	Y	N	Y	N	Y	N	Y	
Related? (Majority Vote)		Y	Y	Y	Y	N	Y	N	Y	
Related? (RAP)		N	Y	N	Y	N	Y	N	Y	
Crowd Accuracy	Reading	0.4	1.0	0.4	0.8	0.6	1.0	0.6	0.8	0.70
	Keyword	0.2	0.8	0.6	1.0	0.8	0.8	0.6	0.6	0.68
	Summary	0.0	0.8	0.0	0.6	1.0	1.0	1.0	0.8	0.65
Crowd Precision	Reading									0.64
	Keyword									0.64
	Summary									0.62
Crowd Recall	Reading									0.90
	Keyword									0.80
	Summary									0.80

Table 4.1: Quality results for the preliminary study.

calculations, such as an agreement value for a particular document in a condition. Both RAI and Fleiss' κ use a 0-1 scale where 0 is no agreement and 1 is perfect agreement.

I also measured the crowd's efficiency in analyzing documents in terms of time and attempts as attrition. Time describes how long it takes for a task to be completed and is an indicator of how much effort the task requires. Attempts describes how many workers accept and return a HIT before it is completed and is an indicator of the perceived difficulty of the task.

4.3.5 Results

Individual quality similar across condition

There was no significant difference in individual quality across the three conditions in terms of accuracy, precision, or recall. The results of the quality analysis are shown in Table 4.1. The average accuracy across all conditions was 0.68 (max: 1.0). The average accuracy

		Topic 1		Topic 2		Topic 3		Topic 4		Avg.
Document	Condition	1	2	3	4	5	6	7	8	
Raw Agreement Indices (RAI)	Reading	0.4*	1.0	0.4*	0.6	0.4*	1.0	0.4*	0.6	0.60
	Keyword	0.6	0.6	0.4	1.0	0.6	0.6	0.4	0.4	0.58
	Summary	1.0*	0.6	1.0*	0.4	1.0	1.0	1.0	0.6	0.83
Fleiss' κ	Reading									0.56
	Keyword									0.54
	Summary									0.80

Table 4.2: Agreement results for the preliminary study. * indicates teaching opportunity.

per condition was reading: 0.70, keyword: 0.68, and summary: 0.65. A one-way ANOVA showed semantic activity did not have a significant effect on accuracy ($F(2, 21)=0.051$, p=n.s.). The average precision values for the reading (0.64), keyword (0.64), and summary (0.62) conditions were very similar. The average recall values were 0.90 for reading and 0.80 for both keyword and summary.

Majority vote improves quality

Since we have five unique worker results for each combination of condition, document, and topic, I also considered how an aggregated (majority vote) decision affected quality. When we used a majority vote strategy, there was only one miss for the keyword condition and two misses for each of the other two conditions, giving overall accuracy values of 0.88 and 0.75, respectively. The precision values are 0.80 for keyword and 0.67 for both reading and summary. The recall value is 1.0 for all three conditions.

Summarizing leads to higher agreement

I found that the summary condition led to higher average agreement. Table 4.2 shows intracrowd agreement. Both measures, Fleiss' κ and Raw Agreement Indices, show very similar results and trends (with r=1.0). For RAI, average agreement is highest in the summary

condition (mean=0.80). Agreement was similar in the reading (mean=0.60) and keywords (mean=0.58) conditions. For agreement in individual documents, a one-way ANOVA showed no effect of semantic task on RAI scores ($F(2, 21)=2.676$, p=n.s.).

For Fleiss' κ, average agreement in the summary condition was 0.80, interpreted as between "substantial agreement" and "almost perfect agreement". The values for the reading and keyword conditions were similar, 0.56 and 0.54 respectively, indicating "moderate agreement."

Reading is fastest

Overall, the average time to complete a task was about 11 minutes (SD = 5.3, min=1.5, max=29). Broken down by condition, the averages were reading: 7.8 min (SD = 3.7), keyword: 11 min (SD = 4.5), and summary: 13 min (SD = 6.0). A one-way ANOVA showed condition had a significant effect on time ($F(2, 117)=12.66$, $p<0.01$). Post-hoc Tukey tests showed that the reading condition was significantly faster than both the keyword and summary conditions. There was no difference between the keyword and summary conditions.

Keywords require most attempts

Overall, on average it required about 2.8 attempts (SD = 2.2, min=1, max=11) to complete a task. Average attempts per condition were reading: 2.15 (SD = 1.7), keywords: 3.80 (SD = 2.7), and summary: 2.55 (SD = 1.9). A one-way ANOVA showed that condition had a significant effect on attempts ($F(2, 117)=6.65$, $p<0.01$). Post-hoc Tukey tests showed that it took significantly more attempts to complete the keyword condition than the reading and summary condition. There was no difference between the reading and summary conditions.

4.3.6 Discussion

Quality

The results of the quality analysis showed that semantic task did not affect quality for individual workers. Across all three conditions, individuals in the crowd did better than flipping a coin but might not be good enough for scholarly work. This result supported the general assumption that a novice might not be able to produce high quality results due to lack of expertise. The imperfect numbers first indicate occurrences of crowds' confusions and reasoning provided by the crowd further reveals what the confusions are.

By investigating the keywords in the keyword condition, I found that some participants made wrong connections based on just a few keywords. For example, when some participants saw the keyword "Fourth of July", they directly connected the document to the topic American Nationalism regardless of the context for how "Fourth of July" is used. This seems to support the role of von Restorff effect in underlining or highlighting as discussed in previous studies [41, 42, 105, 110], because participants tend to remember what has been highlighted.

However, measuring aggregated crowd results using a majority vote technique showed stronger results in-line with prior work (e.g. [120]). The stronger recall and precision values suggested that the proposed crowdsourcing model in history classes was feasible because the crowd was able to find all related sources while filtering out some unrelated ones. For example, in the preliminary study's dataset, the size of search pool would have been reduced by 25% (2 misses of unrelated sources in the reading and summary conditions) or 37.5% (one miss of unrelated source in the keyword condition) for the historian. Further, 75% of time the historian spent on unrelated documents would have been saved in the keyword condition, and 50% of time the historian spent on unrelated documents would have been saved in the reading and summary conditions.

Agreement

The results of agreement are shown in Table 4.2. While there was no significant difference among the conditions based on the fine-grained RAI agreement value for each document, the overall agreement was higher at the summary condition (0.83) than at the reading (0.60) and keyword conditions (0.58). This result seems to be in line with previous studies [36, 37, 149] showing summarizing demands a deep level of semantic processing.

When multiple crowd workers make the same incorrect connection, it often means there is some shared confusion, a common misunderstanding, or both. This situation suggests an opportunity for historians to help the crowd (e.g., students in a class) better understand the material. Like most experts, historians' time is limited, so it is important to prioritize these misconceptions to help as many students as possible. Our measure of intra-crowd agreement can be a good indicator for this.

The results from Table 4.2 showed there were two high-impact confusions in the summary condition for Topic 1 and Topic 2, and one confusion for each topic in the reading condition. In these situations, the crowd majority thought an unrelated document was related. For example, for Topic 1 (Revolutionary History and Ideals), Historian A made no connection, but crowd workers in the summary condition thought there was one.

Efficiency

The results show that the reading condition was significantly faster than the other two. However, with respect to number of attempts, there was no difference between reading and summary, while keywords required significantly more attempts than either. This latter result surprised us, as summarizing has been previously shown to be more cognitively demanding. One explanation is that our instructions were inadvertently phrased in a way that made

the keyword condition seem more laborious than it actually was. In the keyword condition, participants were asked to provide "4-8 keywords/keyphrases" while in summary condition, participants were asked to provide "1-2 sentences". By glancing the numbers shown in the semantic task instructions, there may have seemed more work to be done in the keyword condition than for the summary condition.

In the next section, we introduce a new crowd algorithm, read-agree-predict (RAP), which builds on findings from a follow-up analysis of the data from the preliminary study.

4.4 Read-Agree-Predict (RAP)

The preliminary study showed mixed results for the three semantic tasks: reading, underlining, and summarizing. Going beyond the original research questions, I made several observations in our follow-up data analysis that suggested an approach could yield better results than any one task, and better than other common quality control techniques like majority vote. I call this combined approach Read-Agree-Predict (RAP).

4.4.1 Observations from preliminary study

In the preliminary study, there were three possible levels of intra-crowd agreement: zero workers vs. five, one vs. four, and two vs. three, corresponding to RAI scores of 1.0, 0.6, and 0.4, respectively (see Table 4.2). While the first two levels were considered high agreement because the crowd had a clear majority choice, the third was considered low agreement because workers were nearly equally split. We could therefore choose 0.6 as a threshold to distinguish high (0.6) and low (<0.6) agreement.

I made two observations with respect to this agreement threshold that held for only the read-

ing condition. First, I observed that if crowd agreement was low (RAI=0.4), the document was always unrelated. In other words, confusion or disagreement among workers suggests the document is not related to the topic. These situations may reflect a lack of information or ambiguity in the source material.

Second, I observed that if crowd agreement was high (RAI 0.6), the crowd's majority-vote decision was highly accurate. In other words, when crowds converge on a single decision (related or unrelated), that decision can usually be trusted. These situations may occur when there is sufficient evidence for the crowd to make a clear yes-or-no decision.

Taken together, these observations suggest the following robust pattern could be used to be used to predict highly accurate connections between documents and topics. If a crowd reading a document reaches low agreement about its relatedness to a given topic, i.e. a nearly split vote over whether the document is or is not related, then we can predict the document is unrelated. However, if a crowd has high agreement about a document's relatedness, its majority vote decision (related or unrelated) can be trusted. I call this pattern read-agree-predict (RAP).

4.4.2 RAP vs. majority vote

RAP can be viewed as an improvement upon majority vote for crowdsourced adjudication. This improvement is two-fold: 1) it tells when to reliably use majority vote — only when crowd agreement is high (RAI 0.6), and 2) what to do when majority vote is not reliable — the given document is unrelated to the given topic. While much crowdsourcing research uses simple majority vote for adjudication or relevance assessment, RAP pushes the concept a step further by 1) demonstrating how a threshold value of majority may have strong impact on output and 2) providing a clear binary relevance judgement in all possible situations.

Overall, in the preliminary study, majority vote allowed the crowd to achieve quality scores up to 0.8 (precision) and 1.0 (recall) for certain topics and documents. These results could have helped reduce the size of a historian's search pool by up to 37% and saved up to 75% of time spent on unrelated documents in the archive.

For comparison, I applied RAP post-hoc to the preliminary study's dataset. The results in Table 1 show that RAP is a substantial improvement over majority vote, yielding perfect accuracy relative to Historian A's gold standard judgements. RAP achieved scores of 1.0 (precision) and 1.0 (recall) across all documents and topics. These results suggest a historian would not even have to search a digital archive herself, because RAP would have helped the historian find all related documents.

4.4.3 Crowd Confusion as Teaching Opportunities

Beyond producing high quality connections from noisy ones, RAP not only detects where crowds' confusions may occur but also prioritizes these confusions as a useful byproduct. If agreement in the reading condition is low and the majority thinks the unknown document is related, then RAP predicts this source-topic pair will be a high-impact confusion for teaching.

4.4.4 Usage scenario

A potential use scenario is shown in Figure 4.2. A historian would submit a topic of interest, its definition and 2 good examples to a digital archive equipped with read-agree-predict. By leveraging the crowd such as citizen archivists, read-agree-predict would go over all available sources in the archive and output sources related to the topic as well as high-impact opportunities for public history back to the historian.

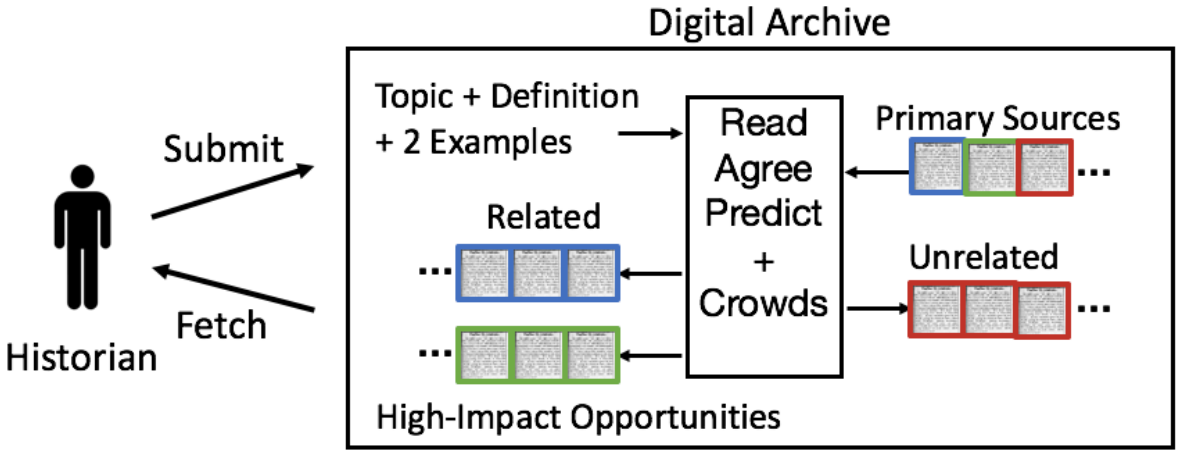


Figure 4.2: A use scenario of Read-Agree-Predict

In addition, RAP is also backward compatible with the class-sourcing model if we use student crowds. Consider the following usage scenario for how RAP could be used in a classroom setting. A historian could begin with a list of topics of interests and a collection of unprocessed primary sources. In the historian's class, she picks topics of interest related to the class and asks students to make connections between these topics and unprocessed primary sources. As the class progresses, the RAP automatically reports related primary sources to the topics and prioritizes students' confusions about sources and topics by aggregating students' connections. The historian then uses related sources for research and clears students' confusions starting with highest priority

Taking data from the preliminary study as a baseline example, with 189 primary sources and 4 topics related to the class, it would take 9828 human minutes to complete all possible connections (189 sources $\times$ 4 topics $\times$ 5 students per source-topic pair $\times$ 2.6 average reading time per source). Historian A generally has about 35 students in his class on the American Civil War, and there are about 16 weeks per semester, requiring about 17 minutes per week for each student in a semester. In reality, students should be able to analyze more sources as

they learn to improve their skills throughout the process, and five students are not always needed when there is already a high agreement.

Studies have shown that students can memorize more historical facts than historians who were not familiar with the topic, but these historians still outperformed the students in weighing competing claims and formulating reasoned interpretations for the topic [48, 146]. Therefore, authentic experiences in historical research would provide valuable opportunities for the students to learn how to apply their learned knowledge and practice historical thinking like historians [48, 130]. In addition, these semester-long class assessments regularly check if students have confusions about these topics and primary sources. Ideally, students will learn to analyze primary sources like a professional historian.

In the next section, I simulate this usage scenario with a new study to validate RAP.

4.5 Validation Study

I conducted this study to validate RAP, so the experimental design was almost identical to the preliminary study. I summarize the differences below.

4.5.1 Dataset and historian

I again used documents from an online archive of approximately 1200 primary sources from the American Civil War Era. I recruited a new expert historian from our institution, Historian B. Following the usage scenario above, I asked Historian B to generate a topic of interest, definition, and two historical documents related to that topic as good examples. Historian B, drawing on his research interests, chose the topic “Racial Equality”.

4.5.2 Apparatus and procedure

I used a very similar web-based interface and procedure as the preliminary study, with the following changes. I removed the keyword and summary conditions, which did not show clear advantages in the preliminary study and are not part of RAP. I kept the reading condition the same as before.

4.5.3 Participants

I recruited 50 participants on Amazon Mechanical Turk using the same criteria and pay rate as the preliminary study.

4.5.4 Experimental design

The experimental design mirrors that of the preliminary study, with the exception of document selection. I first randomly sampled from the archive 10 new documents of similar length and readability level that were not among the set of eight used in the preliminary study. Next,I asked Historian B to pick a topic of interest (without seeing the 10 documents). Finally, I asked Historian B to generate gold standard answers by reading each of the 10 documents and deciding whether it was related or unrelated to his topic.

I used this selection mechanism because 1) it avoided biasing our expert and 2) it reflected how RAP would be used in a real-world situation. That is, a historian approaches or acquires an unfamiliar digital archive, provides a topic, definition, and two example documents, and the (student) crowd analyzes each document from the archive to decide if it is related to that topic. After that, the historian comes back to check the sources analyzed by the (student) crowd.

		Topic 5									
	Document	9	10	11	12	13	14	15	16	17	18
Quality	Related? (Historian B)	N	N	N	Y	N	Y	Y	Y	N	N
	Related? (Majority Vote)	Y	N	N	Y	Y	Y	Y	Y	N	N
	Related? (RAP)	N	N	N	Y	N	Y	Y	Y	N	N
Agreement	RAI	0.4*	0.4*	0.6	1.0	0.4*	1.0	0.6	1.0	1.0	0.4*
	Wrong Votes (out of 5)	3	2	1	0	3	0	1	0	0	2

Table 4.3: Quality and agreement results for the validation study. * indicates teaching opportunity

4.5.5 Results and discussion

After collecting the crowd data from 50 workers, I ran the data through the RAP crowd algorithm to generate predictions of relatedness for each of the 10 documents. Table 4.3 shows that the RAP predictions exactly matched the gold standard answers provided by Historian B. Thus, in this validation study, RAP again achieved perfect accuracy for a new historian, new topic of interest, and new, random sample of documents within the same digital archive as in the preliminary study.

RAP also automatically prioritized documents with confusions based on the number of wrong votes for teacher historian's reference.

4.5.6 Simulating different crowd sizes

To further investigate the effectiveness of RAP, I ran a simulation to understand how RAP would compare to majority vote with different hypothetical crowd sizes. For each crowd size n, I resampled (with replacement) the existing crowd data to create the desired crowd size. I then calculated the average F-1 score for 1,000 resampled data points. I used the F-1

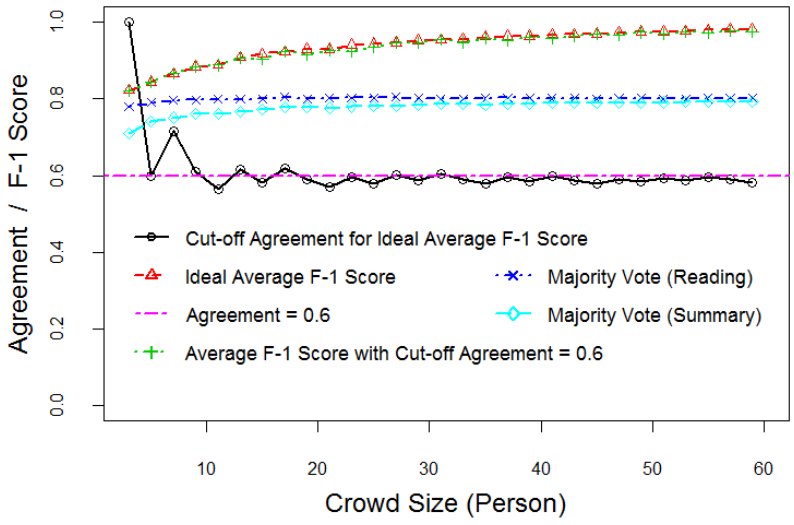


Figure 4.3: Comparison of agreement methods and recommended agreement vs. crowd size in validation

score (harmonic mean of precision and recall) because it is a widely-used measure of search performance in Information Retrieval research.

In Figure 4.3, "Ideal Average F-1 Score" is the best average F-1 score that RAP achieves for a given crowd size. "Cut-off Agreement for Ideal Average F-1 Score" is the recommended agreement threshold to achieve the ideal average F-1 score. "Average F-1 Score with Cut-off Agreement = 0.6" is the average F-1 score using a threshold of 0.6. "Majority Vote (Reading)" is an F-1 score using the majority vote from the reading condition. Crowd size is on the x-axis, and both agreement threshold and F-1 score are shown on the y-axis.

The simulation results suggest three key takeaways. First, RAP's average F-1 score is very close to the ideal average F-1 score (correlation coefficient=0.99). This suggests that the agreement threshold I used for both the preliminary and validation studies, 0.6, was an effective choice.

Third, the benefits of RAP increase with larger crowd sizes, approaching perfect accuracy.

At crowd size=5, used in the preliminary study, the average performance of RAP is already close to Historian B, with F-1=0.84. At crowd size=11, the average performance is equal to Historian B, with F-1=0.89. In contrast, the F-1 score for majority vote quickly saturates at around 0.8.

4.5.7 Historian accuracy and agreement

To complement this validation, I also sought to create a baseline of historian performance by comparing the two historians in our studies, Historian A and Historian B, to each other. I asked both historians to judge relatedness for the document set they had not seen before, i.e., Historian B judged Documents 1-8 (preliminary 8 study), and Historian A judged Documents 9-18 (validation study). Across both document sets, there was 9 substantial agreement between the two historians (Cohen's $\kappa = 0.72$). The average F-1 score across both 10 historians and document sets was 0.89. This could be interpreted as a general measure of historians' performance in finding related sources for other historians.

These results support the intuition that historians can have slightly different interpretation of documents based on their research context. RAP is able to follow individual historians' interpretation in their individual research context by achieving perfect accuracy for both historians and datasets.

4.5.8 Comparison to automated techniques

Based on their research context. RAP is able to follow individual historians' interpretation in their individual research context by achieving perfect accuracy for both historians and datasets. As a point of comparison, I also used purely automated techniques to classify the same dataset from the preliminary study, because Historian A had labeled all the primary

sources with his research topics. To maximize the power of automated techniques, I used all 189 documents with the four topics, compared to only two example documents needed for RAP. Since all the primary sources are digitized and in an image 21 format, our first step was to use an optical character recognition (OCR) system, Tesseract 4.00.00a (with LSTM) [124, 125, 126], to automatically transcribe them. Next, I preprocessed these textual documents by 23 removing stopwords and stemming words based on a Snowball algorithm.I then to transformed these preprocessed documents into TF-IDF space. Next, I chose five techniques representing five different categories of algorithms for binary text classification: 1) logistic regression, 2) kNN (k=9 to maximize available class samples), 3) SVM, 4) decision tree (CART), and 5) random forest (Sebastiani, 2002). Next, I ran stratified 10-fold cross validations for all five techniques for each of the historian's four topics.

The results show that all techniques have high accuracy (0.75-0.95) but very low recall (0-0.3) due to the highly imbalanced numbers of class samples. For example, there were only 10 out of 189 documents related to "American Hypocrisy" for which accuracy is 0.91-0.95, but recall is 0 across all techniques. This means all related documents were missed for that topic.

To deal with the class imbalance issue, I applied three common techniques: adjusting class weights, random over-sampling, and random under-sampling. Under-sampling showed the best improvement, with accuracy 0.55-0.8 and recall 0.2-0.7. For example, with under-sampling, SVM had highest recall (0.7) and 0.55 accuracy for "American Hypocrisy". Although this was a substantial improvement in recall, it may still not be practical, because there were very few related examples for this topic, and 30% of the related documents were mistakenly excluded by the automated technique.

Although future advancements may make automated techniques more powerful, the above results show RAP may offer a compelling alternative in our demonstrated context of history.

4.6 Broader Implications

4.6.1 Quality connections for historical scholarship

Our results provide a baseline crowd accuracy in the history domain. The Read-Agree-Predict (RAP) algorithm allows non-expert transient crowds to find relevant primary sources in a digital archive as effectively as expert historians and as a byproduct, reveals and prioritizes crowds' confusions. I demonstrated the effectiveness of this approach with an authentic historical dataset and two studies with different historians, topics of interest, and documents. RAP also offers clear advantages over majority vote. Our empirical results and simulations show that RAP consistently outperforms majority vote, and larger crowd sizes increase RAP's accuracy to be on par with experts.

The ability to produce quality connections may give more confidence to historians in trusting data collected via crowdsourcing and in adopting this new crowdsourcing model for their research and classes. By doing so, historians can be more focused on exploring interesting research questions and using related sources to support their arguments. Anecdotally, our historian experts Historian A and Historian B were excited to see how crowds could help with their research. Asked about his interest in RAP-enabled crowdsourced support for archival research, Historian A was enthusiastic:

"Definitely! Yeah, I mean that'll be very useful. It's often kind of difficult to do, especially with a topic like nationalism, because it would be hard to just do a keyword search, because nobody was using the word 'nationalism' in the 1860s. So, to some extent, you just have to read, you know, everything in them [digital archives] and kind of hope something useful comes along."

4.6.2 Opportunities for history education

Historical primary sources are important sources for both scholarly research and education in history domain [129, 130] and teaching students to "think like a historian" is one of the main goals in history education [73, 94, 144]. Within this context, the new crowdsourcing model is particularly useful and may create a win-win situation for both teacher historians and students. One the one hand, this model helps historians do research by organizing related primary sources into their research topics and teach by identifying and prioritizing students' confusions. On the other hand, students obtain opportunities to participate in authentic historical research and to practice historical thinking and knowledge with primary sources and receive feedback accordingly. As prior research shows, comparing students' and domain experts' output of the same task is an effective way to identify students' confusions [21, 22, 101].

By adopting the new crowdsourcing model and RAP in classroom settings, historians can easily organize unprocessed primary sources and collect prioritized confusions that may be pervasive among students via RAP-enabled systems, and direct their time and expertise to the ones with higher potential impact. Demystified materials may help motivate and engage students, as research shows that people are often interested in surprising materials that deny their existing assumptions [35, 51]. While other research from non-historical domains shows that it is possible for the crowd to learn through few microtasks in a short amount of time (e.g., $<$ 30 minutes in total) [56, 88, 156], I did not see that in our results of reading comprehension techniques. The wide adoption of long-term apprenticeship in historical research domain may provide an explanation why we have different results [87]. This new crowdsourcing model also provides opportunities for a longer-term learning process.

4.7 Chapter Summary

With digitized historical and scholarly materials made available online, it is often difficult for researchers to find documents of interest because the topics and themes they are investigating are specialized and abstract. In this chapter, I investigated the possibility of a new crowdsourcing model to connect digitized primary sources to high-level topics, and to reveal and prioritize crowd confusions. In our preliminary study, focusing on the effect of different semantic tasks on comprehension, I found promising results supporting the new crowdsourcing model. I also found that a robust pattern emerged enabling highly accurate predictions of document relatedness based on crowd performance. Based on these results, I developed Read-Agree-Predict (RAP), a crowdsourcing approach which allows crowds to evaluate relevance of primary sources to an abstract theme with high accuracy. As a useful byproduct, RAP also reveals situations of crowd confusion that suggest opportunities for learning interventions. I successfully validated RAP with a new historian and dataset of primary sources. While this research used paid crowd workers, it has implications for applications in classroom settings.

5

Zooniverse: Scaling up Complex Crowdsourced Transcriptions

"Yes, through going through these you hear the real story from the soldiers point of view. This has been one of the most incredible projects that I have ever done on zooniverse, this is always going to stay with me"

– A Zooniverse crowd contributor of the American Soldier project

5.1 Motivation and Research Question

While Incite's class-sourcing was effective in the classroom as with the Mapping the Fourth project and the American Soldier project, this model may not scale up well with a large amount of unprocessed documents that have many variants. For a second case study of Incite's class-sourcing mode, historian Ed Gitre and Bradley Nichols used it for a WWII project, "The American Soldier", and deployed it in multiple history courses at Virginia Tech, generating over 3k transcripts; 3k tags; and 30k connections. Based on this success, Gitre wanted to scale up from a small collection of 3k historical documents to a full set of collections of 65k+ documents, beyond what even multiple large classes could support. This full set of collections contains a wide range of input types that should be transcribed including

multiple choices, textboxes, a combination of both, different input sources (handwritings from different people) and exceptions (e.g., marginalia).

This full set of collections brought a few new challenges. First, we needed a way to greatly increase human power to deal with the amount of work and time constraint. Second, we needed well-designed crowdsourcing workflows to handle the highly-variable documents and to harness the extra human power. Finally, we also needed a good way to aggregate results from the crowd. To overcome these challenges, I ask the following research question.

Research question How can we design a crowdsourcing system to scale up the analysis of complex historical documents?

5.2 Challenges

To answer the research questions, there are three challenges.

5.2.1 Challenge 1: source of human power

The experience with Mapping the Fourth project showed that it required much effort to attract online volunteers from scratch including building a well-designed crowdsourcing, and reaching the right communities. Many successful crowdsourcing transcription projects were built and hosted by reputable GLAM's such as What's on the menu from New York Public Library [15], Virginia Memory from Library of Virginia [7], Digital Volunteers from Smithonian institution and Citizen Archivists from National Archives [8]. These institutions all have designated departments and personnel to build specialized crowdsourcing platforms for their collections.

5.2.2 Challenge 2: Design for analyzing complex historical documents

Although Zooniverse was likely to be able to provide enough human power, it remained unclear how the complex historical documents could be analyzed without supervision. Unlike other projects [cite some projects] that generally had low variants of documents types such as handwritten collections of past writers, the collections (44 rolls) of the American Soldier project were surveys filled by individual solders after World War II and encoded by field experts. Therefore, the collections contained many documents variants. These variants included questions of multiple choice, free-text response, multiple answer, filling in the blank, and different combinations of the above. Each document may contain very different individual handwritings and include different codings made by field experts. In addition to these highly-variable questions, the variants were also likely to contain extra marginalia and modification to the original questions.

A successful crowdsourcing project highly depends on its design but there were few clear rules or guidelines to help guide the design process. While there were many existing transcription projects [cite projects], they mostly focused on some specific collections of documents instead of a full set of highly-variable document types. Although there were transcription guidelines from the transcription projects, there were not clear guidelines on how to design a whole system to crowdsource complex document sets.

5.2.3 Challenge 3: Aggregation of crowdsourced transcriptions

This third challenge is that once we collect multiple transcriptions for one document, how should we use these transcriptions? As prior work points out, there is no "average" across these transcriptions [103].

5.3 Design Process and Final Design

To answer the research question, I used research through design [157] and went through an iterative design process to create a crowdsourcing system that can scale up the analysis of complex historical documents. During this process, I worked closely with the subject matter expert, historian Dr. Ed Gitre, other field experts of the American Soldier, students of a couple history classes and users of Zooniverse, a popular citizen science platform.

This section includes all steps involved in the design process. I included standard steps required or suggested by Zooniverse for completeness and focused on the core design – workflows and tutorials.

5.3.1 Data preparation

This step contains 4 sub-steps: 1) data acquisition, 2) data cleaning, 3) data transformation and 4) data upload.

- Data acquisition: I first worked with Ed Gitre and the National Archives to acquire the 65k+ digitized primary sources.

- Data cleaning: I cleaned duplicates based on file names and sizes.

- Data transformation: I converted the originals into a format Zooniverse would accept

- Data upload: I then uploaded the converted files to Zooniverse.

5.3.2 Prior work survey

It is important to learn from past projects. I surveyed several related projects including Virginia Memory [7], Smithsonian's transcription project [8], Operation War Diary [10], African American Civil War Soldiers [2], Measuring the Anzacs [106], What's on the menu? [15] and Transcribe Bentham [132].

Prepare project information

In this step, I collected project objective, listed team members and research goals.

5.3.3 Reference materials

In this step, I compiled different transcription guidelines gathering and adapting those from related projects such as Virginia Memory, Mapping the Fourth of July, Smithsonian transcription projects. I also compiled and organized domain specific references such as military ranks and abbreviation tables. These reference materials are FAQ's and Field Guide on Zooniverse.

5.3.4 Iterative design of workflows and tutorial

Identifying document variants

At first, it was not clear how many document variants in the collections (44 rolls). After looking at the documents, I found some rolls were in a similar format but most of the rolls were in different formats. I tried using random sampling and later, Dr. Gitre identified there was meta information about the format in each roll. Unfortunately, the meta information

might be missing or insufficient in some rare cases. Therefore, I went through the 44 rolls to identify different document variants and questions types by using both meta information and random sampling and finally identified 29 distinct formats as shown in Table 5.1.

Designing crowdsourcing workflows and tutorials

The challenge in the step is the trade-off between generality and specificity. On the one hand, we want all workflows to be similar. This would greatly lower the design effort and the learning effort for the crowd to do the work. On the other hand, we also want the workflows to provide enough context pertinent to the document to better guide the crowd, ensuring quality results.

My initial workflow design was to see how much effort had to be invested so the task was just asking the crowd to transcribe the hand-written portion of the survey. While this design worked for some very simple documents, it did not work for most of the documents because of the high variability of document types. Some documents did not contain handwritten responses and some might contain handwritten responses in different parts of a document. This design left too much gray area for the user to make decisions.

Based on Dr. Gitre and students from his class, this generally frustrated the users and incurred highly variable results that are hard to aggregate later on.

My second attempt was to design a workflow customized to a specific document with specific question sets. While the instructions and tasks were very specific, leaving little room for interpretation, it took much effort to specify all the details of the document and what was worse was that the design was not applicable to most of the other documents because it was too specific to that particular document. Based on the number of documents and question types, it was infeasible, requiring too much effort to design one workflow for each of them.

MC, # (years old on my last birthday), MC, MC, MC, MC
MC, #, MC, MC, VB
MC, MC, MC
MC, MC, MC, MC, MC, VB
MC, MC, MC, MC, MC, MC, VB
MC, MC, MC, VB
MC, MC, MC, VB (a set of guiding questions)
MC, MC, MC, VB, #, VB
MC, MC, MC, VB, VB
MC, MC, MC, VB, VB (a set of guiding questions)
MC, MC, VB
MC, MC, VB, MC, VB
MC, MC, VB, VB
MC, MC, VB (state name), VB (day-monnth-year), VB (state name), VB
MC, MC+VB, VB
MC+VB, MC+VB, MC, VB
MC+VB, MC+VB, MC+VB, VB
MC, VB
MC, VB, VB
MC, VB, VB (a set of guiding questions)
MC+VB, MC, VB
MC+VB, MC+VB, VB
MC+VB, VB
MC+VB+VB, VB
VB
VB (a set of guiding questions)
VB, MC, MC, MC+VB, VB
VB, MC+VB, MC, VB
VB, VB

Table 5.1: Document variants (MC: multiple-choice, VB: verbatim, #: blank filling with a number)

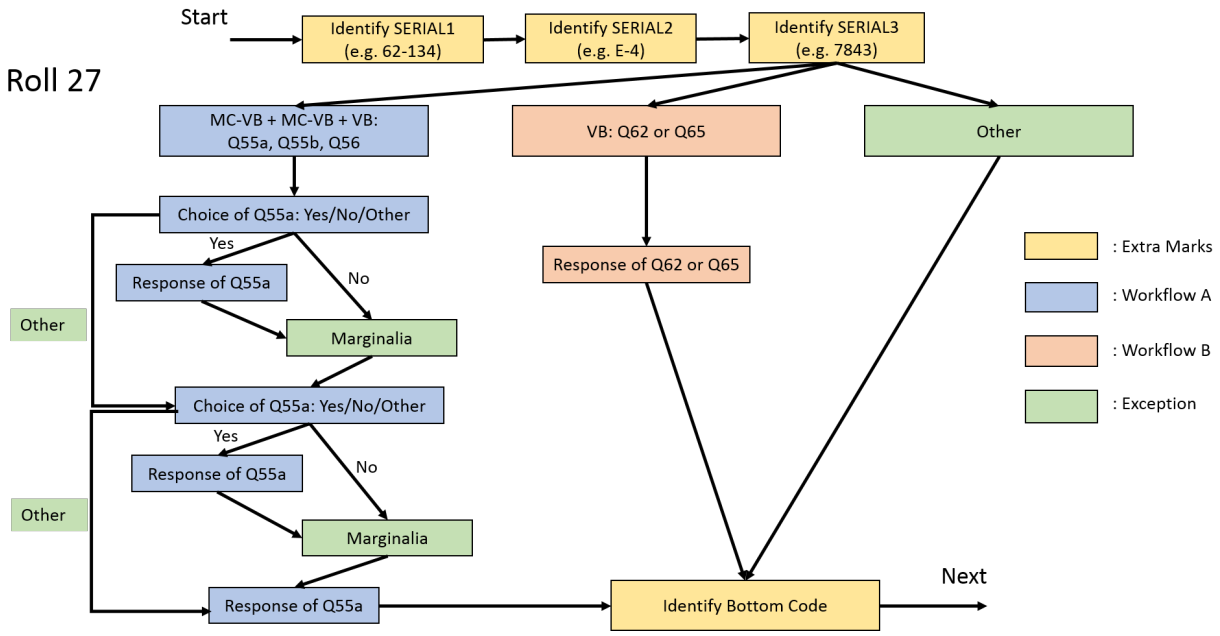


Figure 5.1: Workflows for roll 27

Based on the results of the two attempts and experience of classifying documents, I determined that designing based on document variants would offer a good balance between generality and specificity. Each document variant shares similar question types so that the same instructions and tasks can be reused with no or little modification. The total number of document variants is less than the total number of combinations of document types and question types.

To evaluate the new design, I first chose three of the document variants based on complexity (from low to high) to represent the whole set of collections and then designed workflows and tutorials for them. The simplest workflow contained only one task, that is, transcribe response for some question such as Workflow B in Figure 5.1. A complex workflow may contain several branches and steps like Workflow A in Figure 5.1.

When designing the tutorials, I identified a similar trade-off. One the one hand, we would like to minimize required effort for designers, making one tutorial for all users. On the other

hand, we would like to provide as much detail as possible to give the user full context and step-by-step instructions. Since at this stage, there were only three workflows, I created a tutorial for each of the workflow, to investigate how to most effectively design the tutorials.

With the three fully implemented workflows and associated tutorials, I conducted a user study with the project team and a class of Virginia Tech students to evaluate the design before sending it off for review to be a featured Zooniverse project. This evaluation is described in more detail in a following section as First Evaluation.

Revising the tutorial based on evaluation results

The results of the evaluation (58 responses) showed that the workflow design was effective and gave insights on tutorial design. While being specific and providing much detail were appreciated, many users might not have the patience to read all the context before starting the tasks. I then revised the tutorial design. First, I created one general tutorial that focused on the very basics such as most commonly-used system features and common steps across all the workflows as shown in 5.4. I then moved all the details to the help function within each task and field guide. With this revision, the user can quickly grasp the general idea of the workflow and tasks at the beginning with the general tutorial. When the user needs detailed instructions or more context, the user can access the help function on demand to get information specific to the task.

Improving UI design based on public beta

After these revisions, I submitted the project to Zooniverse for consideration as a featured project. This required a public beta phase where thousands of Zooniverse users tested the site and provided feedback. This feedback was generally positive. I then made some

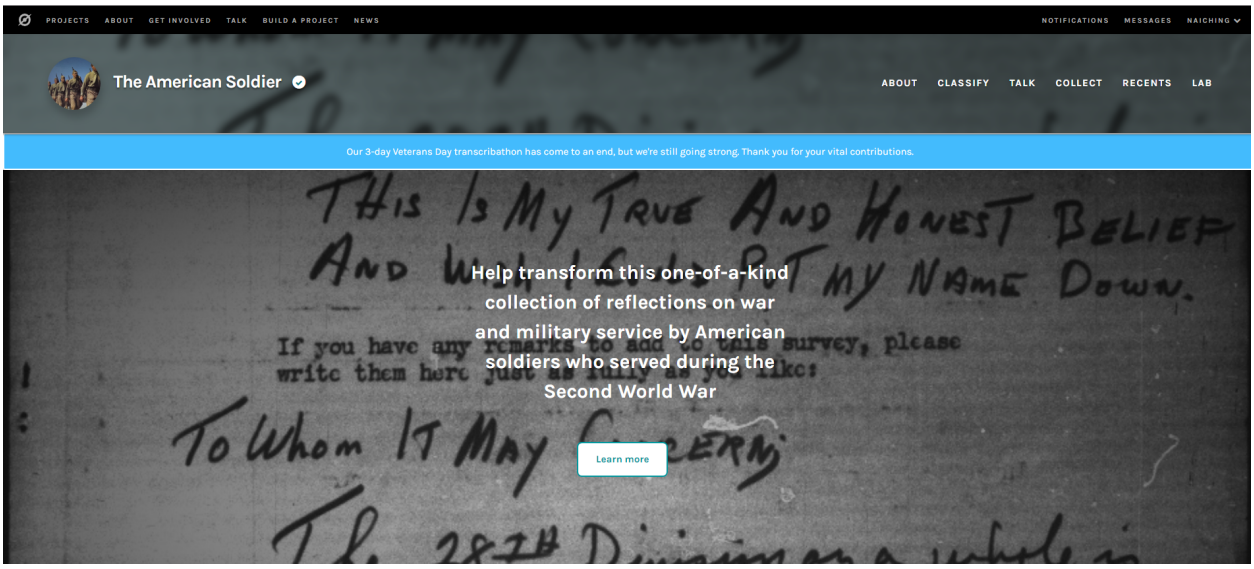


Figure 5.2: Homepage of the American Soldier Project

minor changes based on the beta users' feedback. A following section describes the Second Evaluation in detail. A large portion of the feedback was about Zooniverse interface.

After the revision based on the beta users' feedback, the project was reviewed and approved by Zooniverse staff. Meanwhile, I expanded the design and, finally, created 44 workflows. The project was then promoted as a featured Zooniverse project.

5.3.5 Final design

This section presents different elements of the American Soldier with screenshots. The hompage of the project is shown in Figure 5.2. A regular task view is shown in Figure 5.3. A regular task view with the tutorial is in Figure 5.4. Figure 5.5 is a task view with in-step help information. Finally, Figure 5.6 shows the about page of the project.

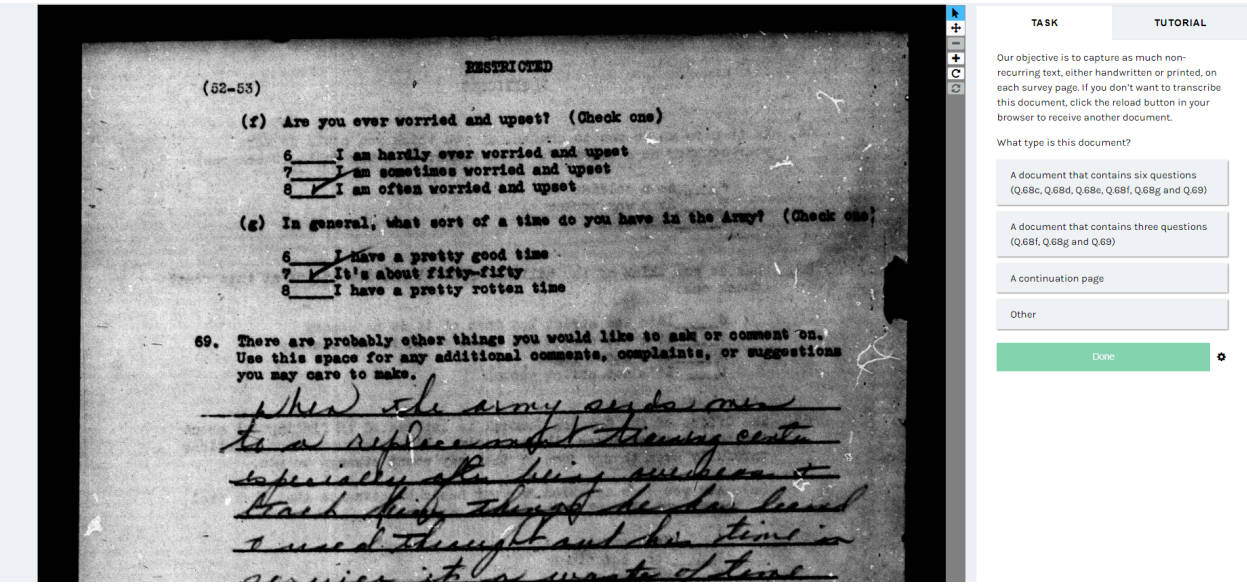


Figure 5.3: A screenshot of task view

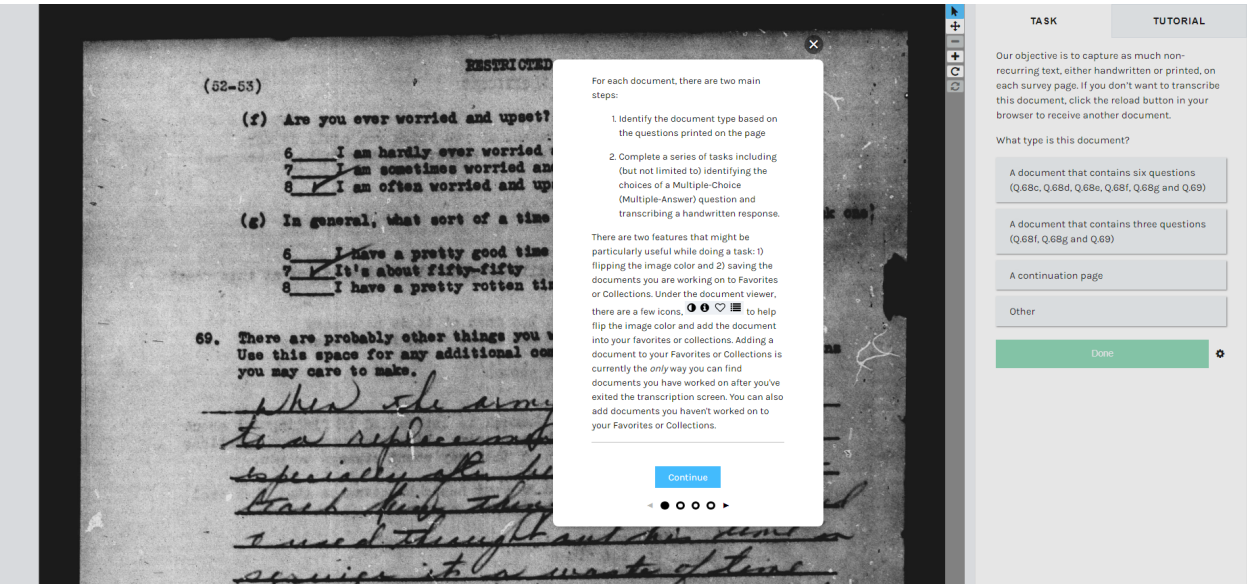


Figure 5.4: A screenshot of task view with tutorial

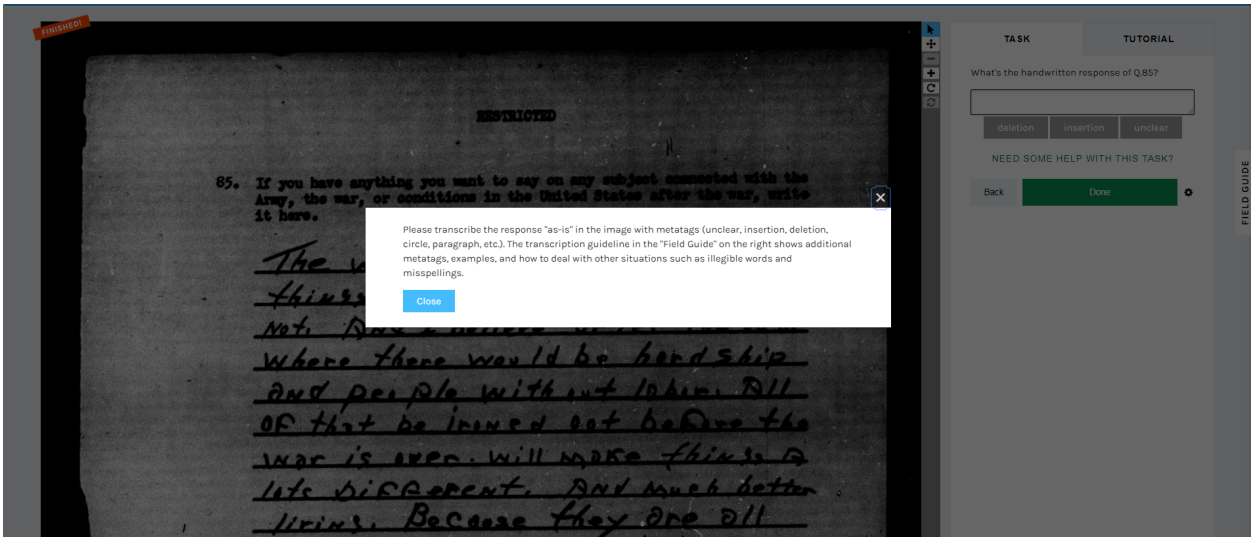


Figure 5.5: A screenshot of task view with in-step help information

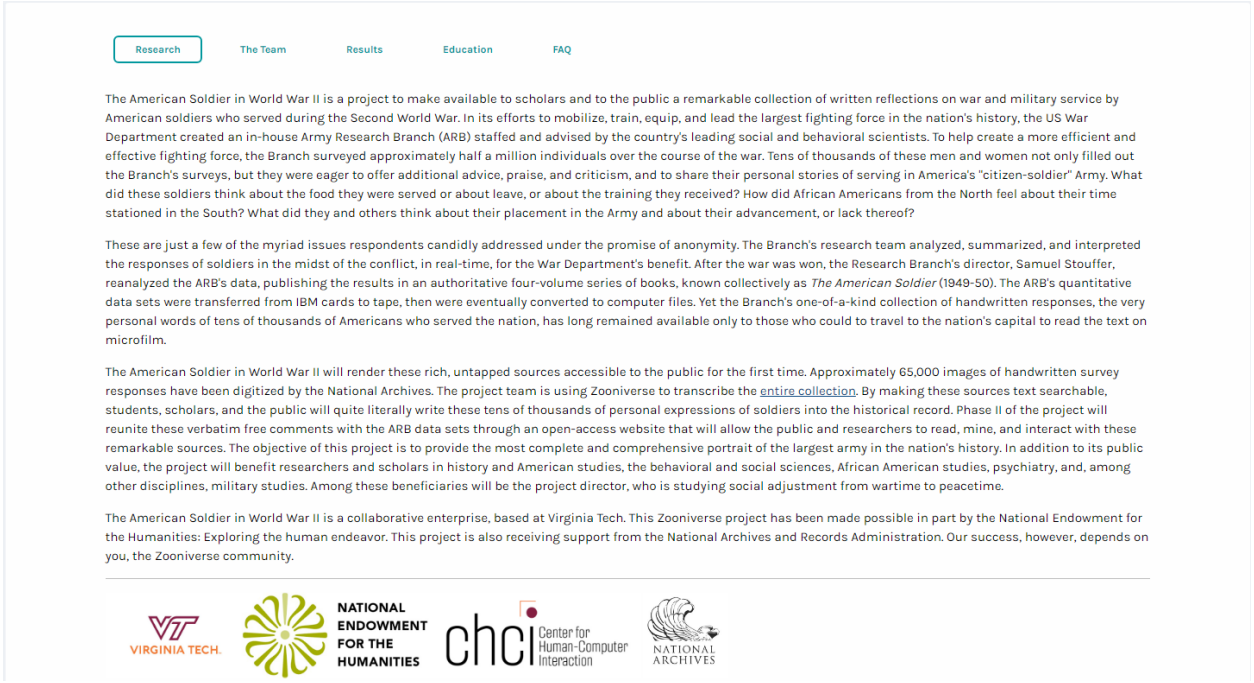


Figure 5.6: A screenshot of about research page

5.4 First Evaluation of Design

I conducted a survey to get feedback on reference materials (field guide), tutorials, workflows, comparison with other similar projects they experienced if any and any other comments.

5.4.1 Participants

Participants were three experts from history and digital humanities of the project and 55 students from a college-level history class who had some transcription experience through the class.

5.4.2 Process (workflow)

The results showed that the design of the workflow was very clear, intuitive, not confusing and fun. Some representative responses are listed as follows.

- (intuitive, not confusing) “this process seems very straight-forward, easy to understand and easy to complete”
- (fun) “I like it and it is fun!”
- (fun) “The process is very enjoyable, it feels rewarding and nothing gives any trouble.”
- (intuitive, not confusing) “The process is neat. The instructions are pretty straight forward and I feel that anyone can pick this up and work on it.”
- (intuitive, not confusing) “It is pretty straightforward after the tutorial”

5.4.3 Tutorial

The results showed that the tutorials were helpful and clear but may sometimes be too lengthy. These tutorials were useful and helpful mostly because they gave the participant an overview of the task and what to expect and do in the following tasks. A sample of the responses is listed as follows.

- (clear, possibly lengthy) "I found it helpful and the description was clear. However, there should definitely be an option to skip the tutorial."
- (clear, helpful) "The tutorial and description is very clear and concise. It is simple and straightforward which makes these tasks simpler."
- (helpful) "I thought the tutorial was helpful in the fact that it described what we had to do as well as make sure we knew which part of the image to transcribe."
- (helpful) "The tutorial is helpful. It breaks down the steps needed in order to transcribe a document. The description is also helpful because it describes the task."
- (clear) "The tutorial was pretty clear as I started to transcribe almost immediately. However, the field guide tips should have been placed in with the tutorial. Some people may not have realized that there was a specific way to deal with words that they had a hard time reading. While the tutorial mentions the field guide, they should bring up all these points in the tutorial so that there is more clarity."

5.4.4 Field guide

The results showed that the field guide was very helpful in details and clear. The field guide was useful in that it gave the transcriber most, if not all, of the details the participant had

to know in order to complete the tasks.

- (helpful) “field guide was very helpful in knowing how exactly how to transcribe the documents”
- (helpful) “The field guide is extremely helpful because I wasn’t sure about a few words and spent so long trying to figure it out, but then saw that the field guide told you what to do with those and it really helped.”
- (helpful) “The field guide is a good basic step by step tool that you need to start transcribing.”
- (clear, helpful) “This is the first site where I have seen this type of guide. I really like how clear the statements are, this leads to less confusion in the long run.”

But due to the design of Zooniverse’s user interface, it may sometimes be hard to locate or notice.

- “I could not find the field guide, so I would recommend making it easier to locate.”
- “I thought the field guide was clear but I did not notice the tab at first”
- “The field guide was helpful because I came across some words I was unsure about. With this being said, I wish the field guide was not located on the side of the page as it took a minute or two to find. It would have been better to include the field guide tips in the tutorial as well.”

5.4.5 Potential improvements

The results suggested that there were some improvements to be made although the design was generally liked and appreciated. These improvements were summarized as follows.

- Tutorials might be too lengthy that the participant might want to skip
- Field guide might be hard to find for some participants
- There were some typos and errors in the description although they did not affect reading much.

5.4.6 Revision

According to the feedback from the survey, I did the following changes to the design. I created a general tutorial so that it could be shorter and focused. I then moved the details that were more specific to the tasks to the "help" feature within each task and field guide if they applied broadly. In this new design, the user can start working as soon as possible with a short tutorial but still can receive extra help and information about each task. This shortened a long 7-step tutorial to a 4-step tutorial. I also added description in the tutorial to remind the user where to locate the field guide. Then I cleaned the language again to make sure it was clear and correct.

5.5 Second Evaluation of Design

This was a standard survey designed by Zooniverse for all the projects going through public beta testing.

5.5.1 Participants

There were 28 beta users from Zooniverse for this survey but 3 of them were not able to use the site and thus excluded from the evaluation.

5.5.2 Task difficulty

Seven out of 25 users found the task easy (very or moderately), while 18 out of 25 found it hard (somewhat to very). These results seem to reflect the general complexity of the primary sources that necessitated our iterative design process. While there are some short and easy documents, most are hard documents.

5.5.3 Usefulness of help text

Most participants (24 out of 25) agreed the help text is useful. This suggests that the help information (tutorial, in-step help, Field Guide) provides needed information and works as intended.

5.5.4 What additional information or capability would you find helpful or interesting?

Suggestions focused on providing more context of the documents and layout of the user interface. This shows that participants are not only interested in contributing via crowdsourced tasks, but also interested in knowing the background of these documents. Providing more contextual information may make a project more engaging.

5.5.5 How did you work through the classification interface?

If the document is too hard, the user may refresh until getting one the user can handle. A lot of zooming in and out involved and magnifying glasses were used when needed. The user generally just followed the workflow and answered the prompts.

5.5.6 Did you find the additional information on other pages useful?

The majority thinks so (17/25). 3 out of 25 did not read and the rest said no. This tells us that the majority actually uses additional information although a small group of people might not be as interested in. This also suggests that it worth the design effort in providing the information which may be a good way of communicating with the participants.

5.5.7 Is project appropriate for the Zooniverse?

The responses unanimously agreed this project was suitable for the Zooniverse.

5.5.8 If we decide to launch this project publicly, do you think you will take part?

About half of the responses were positive about future participation when the project becomes official. Two of them even mentioned about bringing their friends. Only five of the participants specifically said no for future participation. The results show that design of a crowdsourcing project is not the only factor that decides if people want to participant in the project.

5.5.9 Other comments?

The responses also mentioned the project was interesting and even related to their family or themselves as WWII veterans. The responses also recognized the importance and potential values once the project is completed. However, they also mentioned the tasks (handwritings)

were not easy and brought up some bugs (e.g., typos and UI issues) and suggestions to improve the process.

5.5.10 Revision

At this phase, the feedback was mostly about the utility of the design. Some of the feedback was about the Zooniverse's user interface which I did not have much control but only forwarding the feedback to the Zooniverse team. Based on the feedback, I added some more description of the features of the Zooniverse's user interface to help better work on the documents such as recording what's been done and flipping image colors.

5.6 Design Guidelines

Through the iterative design process, I created the American Soldier project on Zooniverse and summarized a set of design guidelines for future projects' reference. These guidelines point out potential trade-offs during the design process, suggest sweet spots of these trade-offs and provide the reasoning process so that future designers can identify different sweet spots if their goals are different.

5.6.1 Design effort vs. specificity of instructions

In this guideline, the goal is to find the balance between design effort and specificity of instructions. On the one hand, we want to put least effort to create working workflows. On the other hand, specific instructions help the user closely match the workflow and the content to complete the task.

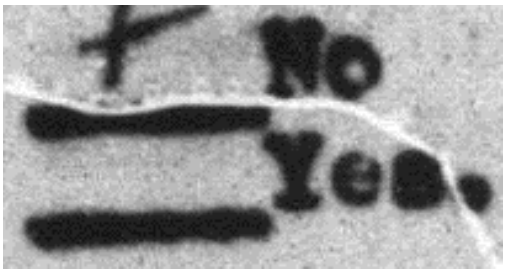
Documents may have different formats or contents, e.g., free-text response, multiple-choice questions, and/or combinations of them. It will be too cumbersome and unrealistic to make a workflow per one document.

A workflow may not work well for two reasons. First, if the workflow is too simple, there may be many situations where the user needs to make decisions. For example, in one of our designs, we asked users to "transcribe the handwritten portion in the document" in our task. While this task worked when there was only one free-text question, it did not work as well when there were different questions with potential handwritten responses. The user needed to decide which handwritten response to transcribe and/or how to put different handwritten responses together. In addition, there might be marginalia. Second, if the workflow is complex enough to give appropriate instructions to different documents, the workflow itself may be too lengthy, complex and error-prone. Changing one step may affect many other steps in the workflow

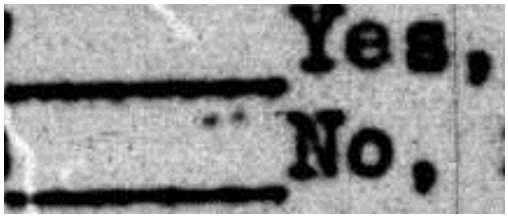
Therefore, having a right balance is important. Our suggestion is to classify the documents based on their similarity and design workflows accordingly so that there are a few less complex workflows that work for all documents. In the American Soldier project, I reduced the number of workflows from as high as 65k+ (or 44 rolls) down to 29 variants by first identifying and classifying these 29 document variants.

5.6.2 Consistency across workflows vs. consistency between a workflow and a document

The goal is to reduce potential user confusion and error. The steps of a workflow should ideally reflect the natural (reading) steps that a person would follow in the documents, that is, follow the natural layout (left to right and top to bottom) including the order of options



(a) No/Yes



(b) Yes/No

Figure 5.7: Different order of yes and no options

of multiple-choice questions. For example, multiple-choice questions may use Yes/No (Yes comes before No option) and No/Yes (No comes before Yes option) in their options as shown in Figure 5.7a and 5.7b.

I first tried to keep the order of all Yes/No options consistent as Yes/No. But based on our evaluation, it was better to follow the order in the document instead of trying to keep it consistent. That is, the consistency between the document and workflow is more important than the consistency between workflows. This approach provides a few benefits, including a close match between the workflow and the document. Also, due the the transient nature of online participation, a user may not complete documents across workflows, so consistency across workflows may be less important.

5.6.3 Granularity of information vs. design effort and user attention

In this guideline, the goal is to reduce design effort and direct user attention to the right place. On the one hand, fewer things to record may expedite the user's transcription process, but may also lose some valuable information during transcription. On the other hand, requiring users to record more things may require more time, leading to greater errors or fa-

tigue, but may also preserve more information during transcription. Although we may want to capture everything from an image, it is probably not feasible to do so. Instead, we want to direct human effort to the parts that are most important, so that the required effort is minimized and the user collects what is most needed. Designers should discuss with domain experts about what should be preserve such as line breaks? paragraphs? misspelling? corrected spelling? uncertainty? printed/handwritten portion? abbreviations? post-processing marks?

In addition, the designers and experts should discuss how these fine-grained details should be recorded. Figure 5.8 shows an example of post-processing marks. The crowdsourcing platform may provide meta tags for this purpose. In our case, Zooniverse provides three meta tags (unclear, delete, insert), which we expanded by following similar logic.

5.6.4 Specificity vs. flexibility of data collection

The goal is to find the balance between collecting specific information while allowing exceptions. Researchers greatly fear missing out on capturing some interesting detail because it was unexpected and their workflow did not capture it. Crowds do not necessarily have the expertise to recognize something interesting that experts would. This is one of the risks and frustrations of delegating work to the crowd.

While classifying documents into types of variants, I discussed with the subject expert and summarized types of potential responses so that I knew what kinds of exceptions there might be. The exceptions we encountered include:

- Marginalia: The respondent may provide extra comments/feedback to questions that do not ask for them. For example, the user may add reasons of choice to a Yes/No as shown in Figure 5.9.

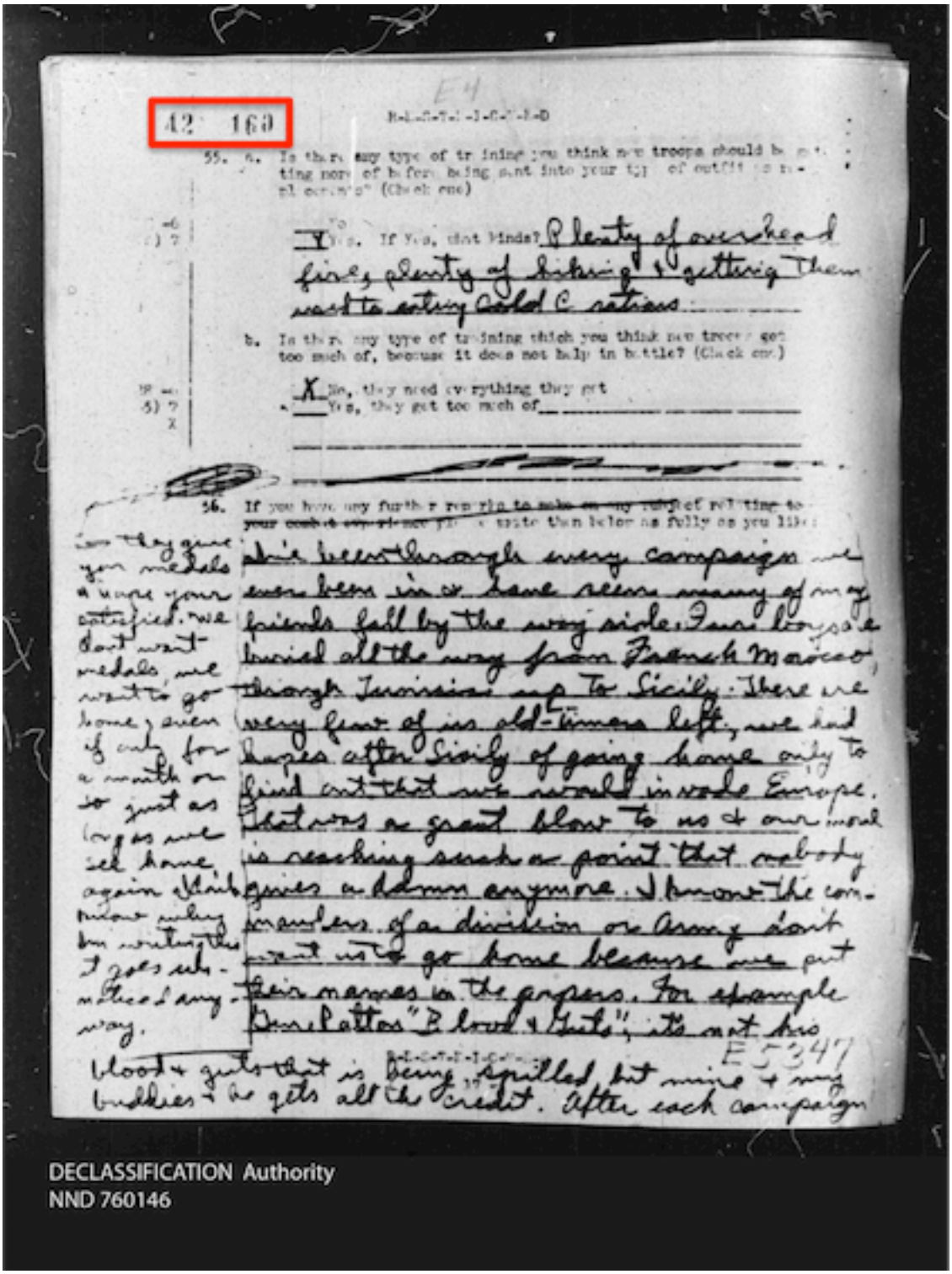


Figure 5.8: Example of post-processing marks

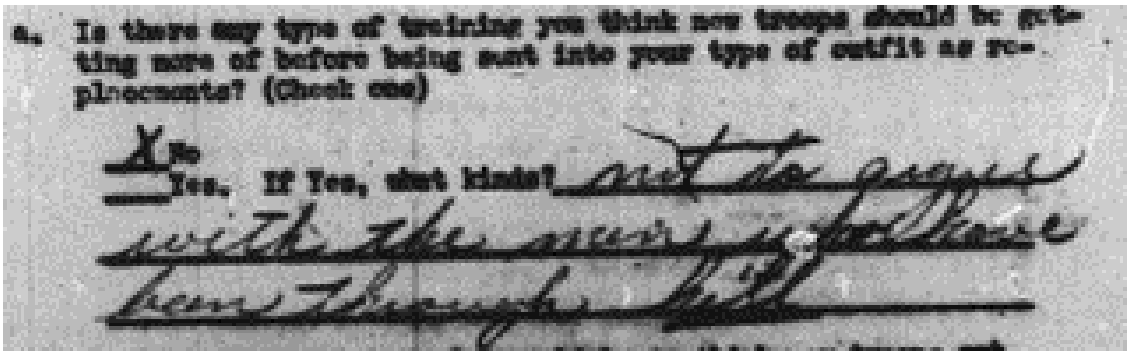


Figure 5.9: Example of exception

- Extra post-processing remarks: Some documents were coded by experts and these codes appear on the document as shown in Figure 5.8.

- Modified questions and responses: The respondent changed the original survey question and then answered the changed question accordingly

5.6.5 Concise tutorial vs. detailed in-step help information

The goal is to both reduce design effort and potential user confusion and error. This is also a trade-off between receiving guidance and working on tasks. Before the workflow, it is valuable to provide a brief tutorial and during the workflow, it is helpful detailed in-step help information.

The tutorial should be concise and giving the user an overall idea of the workflow. While the user may want to read the tutorial to better understand the following tasks, the user may not want to spend much time on reading a lengthy tutorial. They may skip all or part of the tutorial. This will completely defeat the purpose of having a tutorial and waste much design effort for a lengthy tutorial.

Designers should provide detailed in-step help information preferably with visual examples

highlighting the area of interest. This in-step help information is generally need-based. So the user uses when the user has questions or gets stuck with the task, the in-step help information can guide them through the step. The user is also willing to pay more attention to read the help information in order to complete the step.

It may be hard to cover everything in tutorial and the in-step help information. Therefore, in the tutorial and in-step help information, we include a short reference pointing the user to the complete reference (e.g., Field Guide in Zooniverse).

5.6.6 Detailed task guideline vs. contextual information

It is valuable to provide an ultimate place that contains all the related information about the tasks and projects. In addition to task guidelines, it is also important to include related contextual information such as field jargon, abbreviations and specialized terms used in the field of interest because they may appear in the documents. The user is generally interested in knowing more about the background of the project and related information. In Zooniverse, Field Guide is meant for this purpose as shown in Figure.

5.6.7 Robustness vs. redundant effort

The goal is to determine how many transcriptions we should get per each document. On the one hand, we would like to collect many transcriptions for one document so that the transcriptions would cover all words correctly. On the other hand, we would like to collect as few transcriptions per document as possible so that we can finish more documents. Based on our results, one transcription may be too risky as we can see Source 4 in Table 5.2. Three transcriptions seem to work fine. As for generating a good transcription out of a few transcriptions for the same document, there are usually two ways. One is finding one that

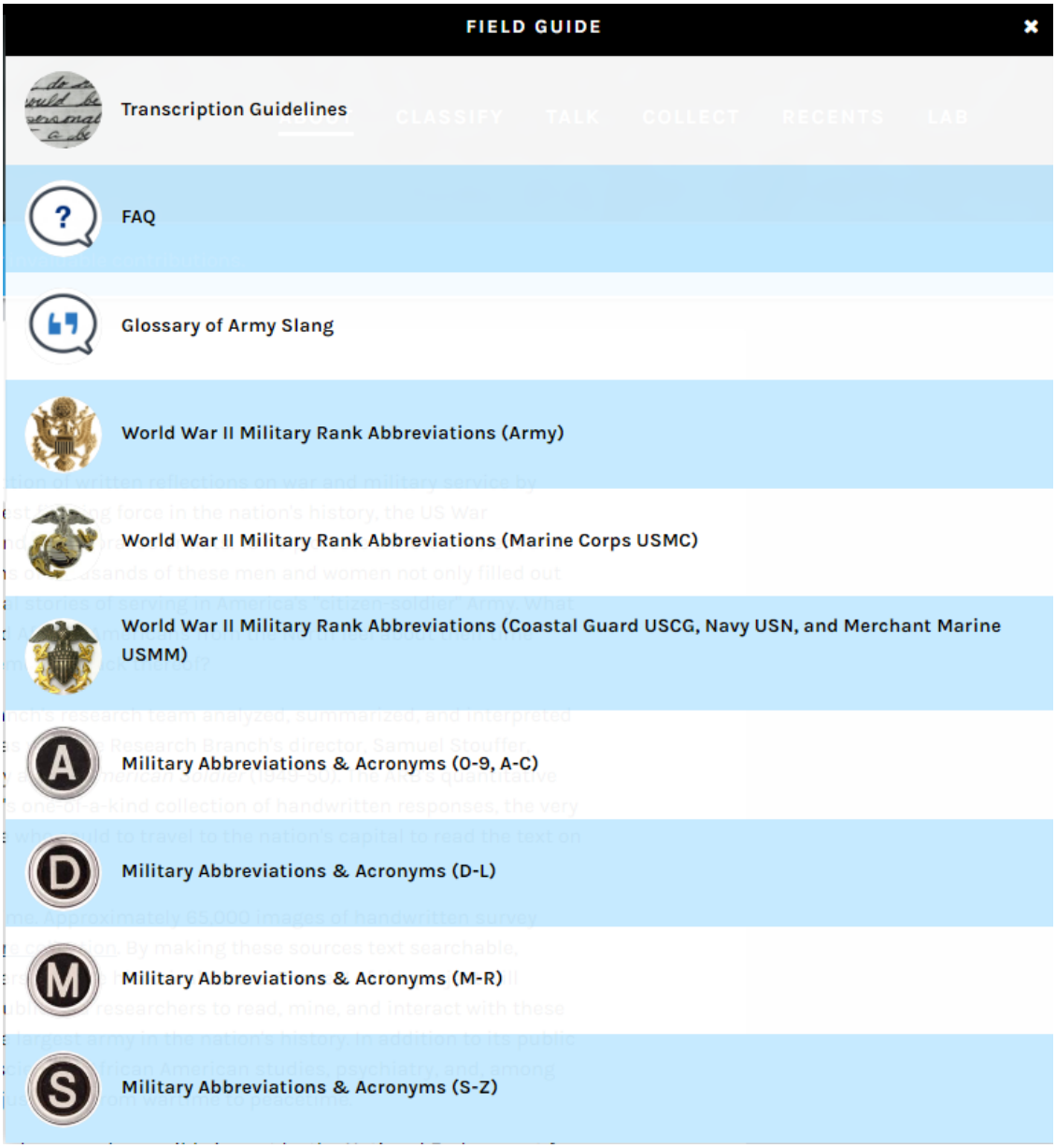


Figure 5.10: Field guide

Source	Transcription
1	I think every soldier should go back through the states before going to the Pacific and [unclear][/unclear] or short furlough.
2 (Pick best)	I think every soldier should go back through to the states before going to the Pacific and have a short furlough
3	I think every soldier should go back through the States before going to the Pacific and have a short furlough.
4	[unclear][/unclear]
Merged (ROVER [60])	I think every soldier should go back through the states before going the the Pacific and [unclear][/unclear] a short furlough.

Table 5.2: One aggregation example

works best [103] and the other is merging words from multiple transcriptions [60]. With three transcriptions per document, both are reasonable methods according to our results. The benefit of pick the best is that the content is generally more consistent because the sentence is from the same author and the drawback is that it does not allow the use of other sources to correct potential mistakes. The benefit of merging is that it can potentially select the correct parts from different sources and the drawback is that the sentence might not be as consistent and may actually merge bad parts into the final results if the merging algorithm is not perfect. One of the examples is shown in Table 5.2. In this example, the merging version merged [unclear][/unclear] into the final results because there was a tie between this choice and 'have' using majority vote. In generally, the merging method is more flexible but its quality is also more sensitive to the choice of merging algorithm.

5.7 General Recommendations for Crowd Research Platforms

Although Zooniverse is a mature platform that has supported many projects, I have gathered feedback from collaborators and users while working on the American Soldier project

that could improve the experience for crowd workers and project creators on Zooniverse and similar crowd research platforms in the future. I have distilled this feedback into the recommendations below.

5.7.1 User interface

Field guide

The current way to locate Field Guide is through a small side bar icon on the very right of the task view shown in Figure 5.5. While Field Guide contains much detailed, useful information, many users reported that they were not even aware of its existence. Two suggestions include 1) make the Field Guide icon more salient and 2) introduce Field Guide in a dedicated tutorial, especially for new users.

Skipping task

Several users also reported that they did not know how to skip a document if the document was too hard for them to transcribe. Consequently, they might end up leaving the task. The current way to do this is to reload the page via the browser's refresh or reload button. Our suggestion is that there could be a dedicated "Skip" button for the user to pass a task in which they are not interested. For example, there might be a "Skip" button beside buttons like "Back" and "Done".

Document viewer

Some users reported that the document view could have been wider for them to view the document more easily; that is, using most of the screen real estate without leaving too much

margin. This is especially useful for those who work with smaller screens. (Update: This recommendation has been accepted by Zooniverse staff and is now the default view.)

5.7.2 Workflow design

Flexible steps

While it is possible to duplicate a workflow in Zooniverse, duplicating *steps* of a workflow is not supported. For projects with complex documents, multiple workflows are expected; e.g., we had 29 unique workflows for the American Soldier project. There are usually some overlapped steps across workflows so it would be a valuable time saver if it were possible to copy some steps of a workflow to another workflow.

Review-based workflows

While gathering multiple submissions for one document is a deliberate decision made by Zooniverse staff, this may not work for all projects, especially when research shows that review-based workflows can improve quality (e.g., [29, 68]) and sidestep aggregation challenges. Our suggestions here are to 1) disclose the rationale behind this deliberate decision, and 2) provide a review-based workflow alternative if the review-based version is more appropriate for some projects.

5.7.3 Aggregation

Retirement count

Since collecting multiple redundant submissions is the current predominant quality control mechanism on Zooniverse, it is important to help the project creator understand how to set an appropriate number of submissions (i.e., retirement count) for each document. This is especially important for creators who do not have related crowdsourcing experience. Based on the American Soldier project, three works well but the actual number may vary depending on the types of documents. The suggestion is that there should be a recommended default number of submission per document along with a piloting process to help the project creator adjust the number accordingly to types of documents. Another suggestion is to let the project creator know how retirement count might affect quality and efficiency. In addition, the recommended default number and discovering process can be shared or obtained via other similar projects and that should make the recommended default number more accurate.

Aggregation mechanisms

Currently, the project creator needs some way to effectively aggregate multiple submissions to produce a final result, but no guidance or tools are provided. We recommend that Zooniverse shares our previously discussed guidelines on this topic (see Section 5.6.7) so that the project creator can be aware available resources. In addition, if possible, Zooniverse might want to implement and integrate some of the methods as baseline tools for project creators because many project creators may not have strong technical backgrounds.

5.7.4 Document and task tracking

Many crowd workers expressed their desired to track their work progress and history, such as documents they transcribed and viewed. This is especially useful when the user observes some patterns across documents over an extended work period that could enrich the data analysis. Users might only vaguely remember a clue or data point they encountered earlier, so it would be very useful for them to go back work history and locate the documents of interest. Some users also wanted to go back to documents they had only partially finished. Currently, the user can only manually add documents to "Favorites" in order to get back to the documents. We suggest that the site automatically track a user's interaction history so that the user can go back to previous documents more easily. The interaction could at least include documents viewed or transcribed (fully or partially) by the user.

5.8 Chapter Summary

In this chapter, I reported the design and implementation process of a crowdsourcing project, the American Soldier project, on Zooniverse. While there are many transcription projects, there were not clear design guidelines about how to design effective crowdsourcing projects to deal with complex documents with many variants. There were also tools to help create crowdsourcing workflows but those tools generally require technical and/or design expertise. Many of them were also design or crowdsourcing markets such as Amazon Mechanical Turk.

This project crowdsources transcriptions of complex documents using volunteer-based platform. Through the design, implementation and evaluation of this project, I discovered a set of design guidelines that would help the design of future crowdsourcing projects dealing with complex document variants without requiring technical and/or design expertise.

These guideline first point out several common trade-offs during the design process and then provide corresponding sweet spots of these trade-offs along with rationale and examples. Future crowdsourcing projects can either take advantage of these suggested sweet spots of the trade-offs if the projects are similar to the American Soldier project or use the rationale to discover their own sweet spots if the projects are very different from the American Soldier project.

6

CrowdSCIM: Scaling up Learning

This chapter is basd on work published at CSCW 2018 [139].

6.1 Motivation and Research Questions

The American Soldier project on Zooniverse demonstrated how analysis of complex historical documents could be scaled up via crowdsourcing, but at the cost of the learning benefits of Incite's class-sourcing approach. I wondered if the SCIM-C framework underpinning Incite could be adapted outside the classroom to a crowdsourcing context, enabling scaled-up productivity without sacrificing learning. I then asked the following research questions:

RQ1: How can we scale up the class-sourcing model while still supporting learning?

RQ2: What are the trade-offs between learning, quality, and efficiency when scaling up?

6.2 Method

I conducted 4 pilot studies and a large-scale experiment to address these questions. The four pilot studies led to the design of CrowdSCIM, a workflow to help crowdworkers learn historical thinking skills while performing micro-tasks supporting historical research. Crowd-SCIM consists of three micro-tasks: Summary-tone, Tag and Connect, corresponding to the

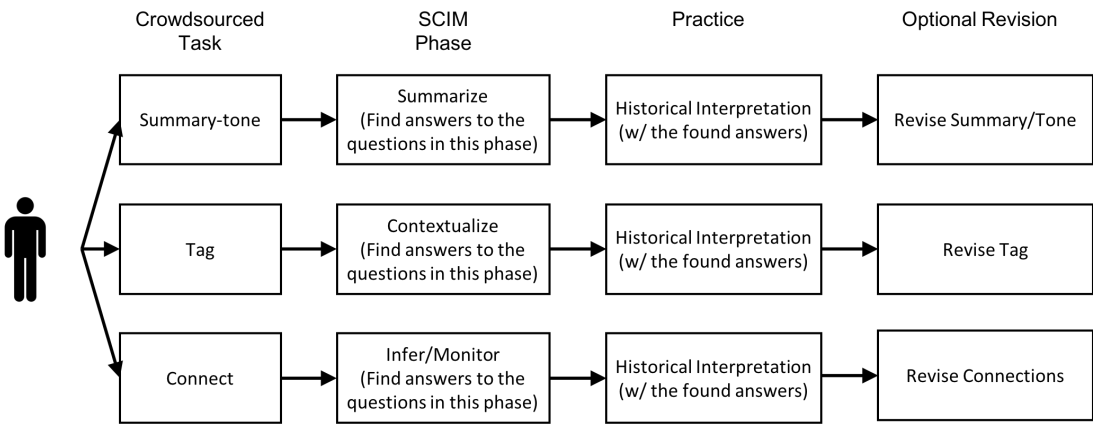


Figure 6.1: The CrowdSCIM workflow

Summarize, Contextualize, and Infer and Monitor phases in SCIM-C, respectively. With this decomposition, each crowdsourced task can be performed individually and each of the phase of SCIM can be learned separately. The workflow is shown in Figure 6.1.

To evaluate the effectiveness of CrowdSCIM, I conducted an experiment comparing CrowdSCIM to three other conditions (Baseline, RvD [156] and Shepherd [56]) in terms of learning, quality, and efficiency. Depending on the task, quality is measured by grades from trained graders or by similarity to gold standard by a field expert. Learning is measured by the difference between grades of a pre-test and a post-test about historical thinking. The experimental design is shown in Figure 6.2.

6.3 CrowdSCIM

6.3.1 Pilot Studies

My goal was to design a crowdsourcing technique to help crowd workers learn historical thinking while working on tasks that may contribute to historical research. To achieve this goal, the crowdsourcing technique had to support learning gains without impeding work

quality. I began with a workflow resembling Incite and, through the series of pilot studies reported below, made iterative changes to arrive at the current CrowdSCIM workflow.

Incite outside of classrooms (Pilot Study 1)

My first step was to see how well Incite could support in learning historical thinking outside of the classroom and without an instructor's intervention. I first customized Incite for the lab study and tested Incite's workflow on the Amazon Mechanical Turk (AMT) paid crowdsourcing platform. I focused on Summary-tone, Tag and Connect tasks. I first asked each participant to complete a pre-test that involved demonstrating historical thinking skills by writing an interpretation of a historical document. The participant then completed the three crowdsourced tasks in sequence, and finally completed a post-test. The tests required writing a historical interpretation for the given primary source. I measured learning by comparing the pre- and post-test scores using rubrics from prior work [70]. Pilot tests with seven participants showed no learning gains from the scores or from their verbatim feedback. In other words, simply giving Incite to crowd workers did not promote learning.

SCIM intervention (pilot study 2)

To try to increase learning, I reviewed the social science education literature and identified the SCIM-C framework [71] as a promising candidate to be adapted for crowd workers. I modified our workflow to add reflective questions from the SCIM-C framework that prompted participants to think about the meaning of the historical document from different perspectives. To minimize the gap between the crowdsourced tasks and SCIM questions, I matched the tasks and questions based on similarity in collaboration with a history professor, Historian A. Specifically, I matched Summary-tone with Summarize because both require a good

summary of the original text. I matched Tag with Contextualize because both ask users to identify entities such as location and time. And I matched Connect with Infer and Monitor because it requires a solid understanding and inference to see if high-level topics are relevant to a given historical document. I tested this revised task design with nine participants from AMT using the same procedure as before. The results again showed no learning effects, suggesting the unmodified SCIM framework is not (directly) applicable to the crowdsourcing context. I observed that although there was no significant learning between tests, the participant's answers often included valuable content that the participant should have included in the post-test. The results seemed to suggest that the participant was able to find required information, but just did not know how to synthesize it in the post-test. Worker feedback also suggested that the task seemed too big for a micro-tasking context. I concluded that it might be too much to ask a crowd worker to complete tasks and learn all four phases of SCIM in one shot and apply all of them in the post-test.

Micro-task design with in-task practice (pilot study 3)

To reduce the workload, improve focus, and increase flexibility, I revised the crowdsourcing workflow to decompose the process into one task at a time. I began with the Tag task because it showed the lowest learning scores for Contextualizing in Pilot Study 2. Our revised procedure again began with the pre-test. The worker made a first attempt at the crowdsourcing task (Tag, in this case). Then, the worker answered a series of reflection questions from the corresponding SCIM phase (Contextualize) and practices writing an interpretation. Next, the worker had the option to revise their Tag task, hopefully incorporating the historical thinking skills from the scaffold. Finally, the worker completed the post-test. I tested this design with six participants from AMT. The results showed a significant learning effect corresponding to Contextualize phase in SCIM with a large effect size (>1.0) for both in-task

and post-test interpretations. The learning effect was slightly higher in in-task practice than in the post-test, an expected result when the scaffold is "faded." Based on these promising results, I tested this workflow with the Summary-tone and Connect tasks, each with five or six participants, and observed similar patterns.

Final workflow

Our iterative process led to the final design of CrowdSCIM, aworkflowto help crowdworkers learn historical thinking skills while performing micro-tasks supporting historical research. CrowdSCIM consists of three micro-tasks: Summary-tone, Tag and Connect, corresponding to the Summarize, Contextualize, and Infer and Monitor phases in SCIM-C, respectively. With this decomposition, each crowdsourced task can be performed individually and each of the phase of SCIM can be learned separately. For the Summary-tone task, the user writes a summary of the document and rates the intensity of each tone from a list. The user then answers four questions from the Summarize phase and writes a historical interpretation containing the answers. Finally, the user can choose whether to revise the summary and tone ratings. For the Tag task, the user tags named entities with categories (e.g., politician, school) for a given primary source. The user then answers the four questions from the Contextualize phase. The user then writes a historical interpretation containing the answers. Finally, the user can revise the tags, if desired. For the Connect task, the user rates relevance of the given historical primary source to each high-level theme in a list. The user then answers four questions from the Infer and Monitor phases. (To balance the workload with the other tasks, I selected two distinctive questions from each phase.) The user then writes a historical interpretation containing the answers. Finally, the user can decide whether to revise the theme ratings. Thus, the generalized CrowdSCIM workflow contains four steps (see Figure 6.1). First, the worker completes an unmodified production microtask (e.g., summarizing,

tagging, or connecting). Second, the worker completes a SCIM phase prompting him or her to reflect on the task just completed by answering a set of questions. Third, the worker writes a historical interpretation synthesizing his or her answers to the questions. Finally, the worker has the option to apply his or her newly sharpened historical thinking skills to the initial production microtask by revising his or her work. Because steps 2–4 in the CrowdSCIM workflow occur after the unmodified production task and do not require content knowledge, CrowdSCIM is relatively straightforward to implement as an enhancement of an existing crowdsourced history workflow. This "add-on" design offers several benefits for requesters. First, it does not require modifying the interface of the initial production task, reducing requester effort and risk. Second, it guarantees at least equal work quality (vs. not using CrowdSCIM) because the intervention is applied after the initial production task. As prior work suggests (e.g., [88]), a primarily learning-oriented workflow may lower work quality, discouraging adoption by requesters. Third, it can be easily turned on and off based on requester needs (see Section 6.6.3 for a discussion of trade-offs).

6.4 Evaluation

To evaluate the effectiveness of CrowdSCIM, I conducted an experiment comparing CrowdSCIM to three other conditions in terms of learning, quality, and efficiency.

6.4.1 Apparatus and procedure

The experiment was conducted entirely online. After completing an online IRB-approved consent form, each unique participant was randomly assigned to one of the four crowdsourcing workflows: CrowdSCIM, RvD [156], Shepherd [56], and a baseline similar to Incite. The

participant was also randomly assigned three different historical documents — one for pre-test, one for the task, and one for post-test — from a pool of five documents. The participant then used the web interface I developed to complete a three-stage work process. First, the participant completed a pre-test on writing a historical interpretation for a historical document. Second, the participant completed the randomly assigned task (Summary-tone, Tag, or Connect). Three of the conditions involved a three-step process: the initial task, an intervention, and an optional revision to the initial task. The CrowdSCIM intervention involved answering four questions derived from SCIM-C. The RvD intervention involved reviewing existing work from other participants. The Shepherd intervention involved self-assessing the participant's own work. The Baseline condition required completing only the task itself and had no intervention. Unlike the Baseline, other three conditions provided an option for the participant to revise the work after the intervention. Third, the participant completed a post-test on writing another historical interpretation for a different historical document. After the work process, the participant completed a post-task survey for demographic information and feedback.

6.4.2 Participants

I recruited novice crowd workers from Amazon Mechanical Turk (AMT). I restricted to US-only workers to increase the likelihood of English language fluency, with a 95% HIT (human intelligence task) minimum acceptance rate and 50 or more completed HITs. I recruited 360 workers and randomly assigned 30 to each of the three crowdsourced tasks of each of the four conditions (30 participants $\times$ 3 crowdsourced tasks $\times$ 4 techniques = 360). Each worker was unique and assigned to only one HIT to ensure that the required expertise was learned within that HIT. Thus, there were 30 unique workers per each crowdsourced task per crowdsourcing technique. I paid participants at least minimum wage ($7.25/hour) based

on average task times in pilots.

6.4.3 Materials

To ensure the validity of our test materials, I used the same historical documents and grading rubric used in previous evaluations of SCIM-C [70]. The SCIM-C materials were selected, constructed, and tested by domain experts including a historian, a teacher educator, an educational psychologist, and a high school social studies teacher. The documents also cover a variety of eras and topics in American history, including the American Civil War, the American Revolution, the Great Depression, and Women's Rights. The random assignment and wide coverage of these documents helped eliminate the possibility that the potential effect was caused by some specific topic or document. Two of the sources are shown on the left panel in Figures 6.3 and 6.4. Some of the task outputs, such as tone ratings, tags, and connections, can be graded automatically if gold standard data is provided. Other task outputs, including historical interpretations and summaries, requires manual grading (see Appendices A and B for the rubrics).

6.4.4 Experimental design

I conducted a between-subjects GRCB (Generalized Randomized Complete Block) design with one treatment factor (crowdsourcing technique), one block factor (crowdsourced tasks), and three dependent variables (learning, quality, and efficiency). The overall experimental design is shown in Figure 6.2 where the process of the Summary-tone task is bolded.

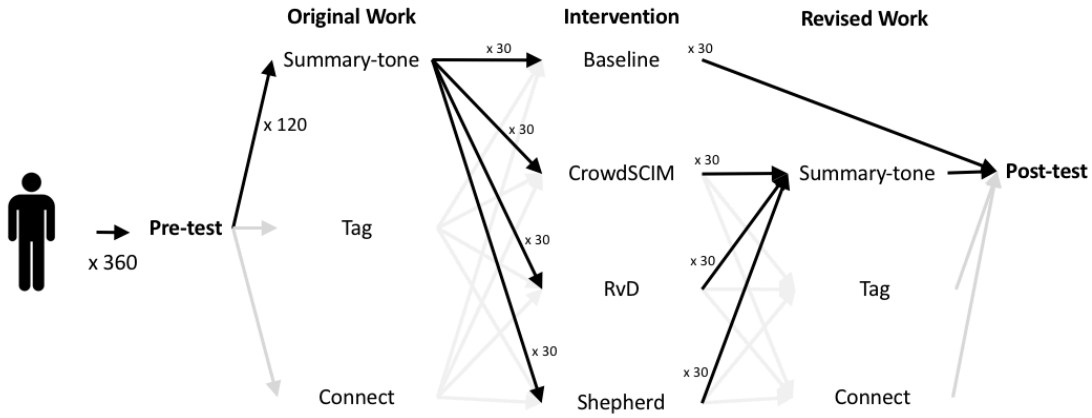


Figure 6.2: Experimental design with the process of the Summary-tone task highlighted

Crowdsourced tasks (block factor)

The crowdsourced tasks Summary-tone, Tag, and Connect were adapted from Incite. The Summary-tone task was to write a maximum three-sentence summary about the given historical document, and then, on a four-point Likert scale, rate how intensely each tone was expressed in the document from a given list of tones (e.g., informative, optimistic). The Tag task was to label named entities with correct categorical information (person, location, and organization). The Connect task was to rate how relevant each theme (e.g., Racial Equality) was to the given document on a 4-point Likert scale.

Crowdsourcing techniques (treatment factor)

The independent variable, crowdsourcing technique, had four levels: CrowdSCIM, RvD, Shepherd, and Baseline. While Baseline did not contain any intervention nor revision steps, the other three levels contained an intervention and an optional revision step. The revision step is the same across all the three techniques, so the only difference is the intervention. The Baseline did not contain any intervention nor revision steps. In addition to the pre-

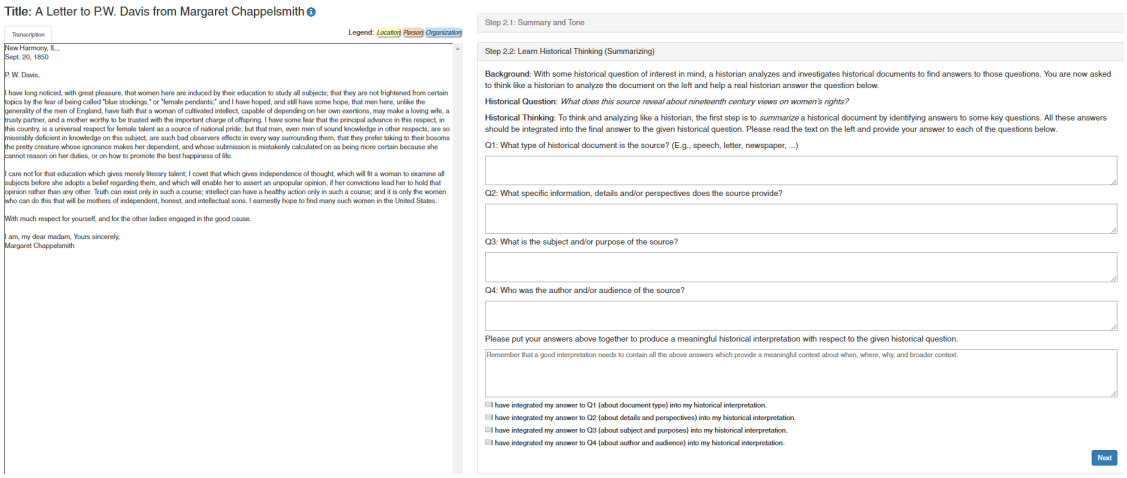


Figure 6.3: A screenshot of the CrowdSCIM intervention for the summary-tone task

and post-tests, the participant was asked to complete the assigned crowdsourced task. The CrowdSCIM intervention is described above. Participant answered four SCIM questions depending on the assigned task. Figure 6.3 includes the four questions of the Summarize phase corresponding to the Summary-tone task (see Appendix C for a complete list of these questions). After the intervention, the participant had a chance to revise their response to the crowdsourced task if so desired. The RvD and Shepherd mimicked the design from the original studies using the same rubric as the graders. After completing the crowdsourced task, both interventions asked the participant to assess the quality of work based on a given rubric. After the intervention, the participant also had a chance to revise their response to the crowdsourced task if so desired. The major difference was that RvD asked the participant to assess another participant's work, while Shepherd asked the participant to self-assess his or her own work. The RvD intervention is demonstrated in Figure 6.4. In the Shepherd intervention, I replaced "the worker" and "the worker's" with "I" and "my", as indicated in the original Shepherd study.

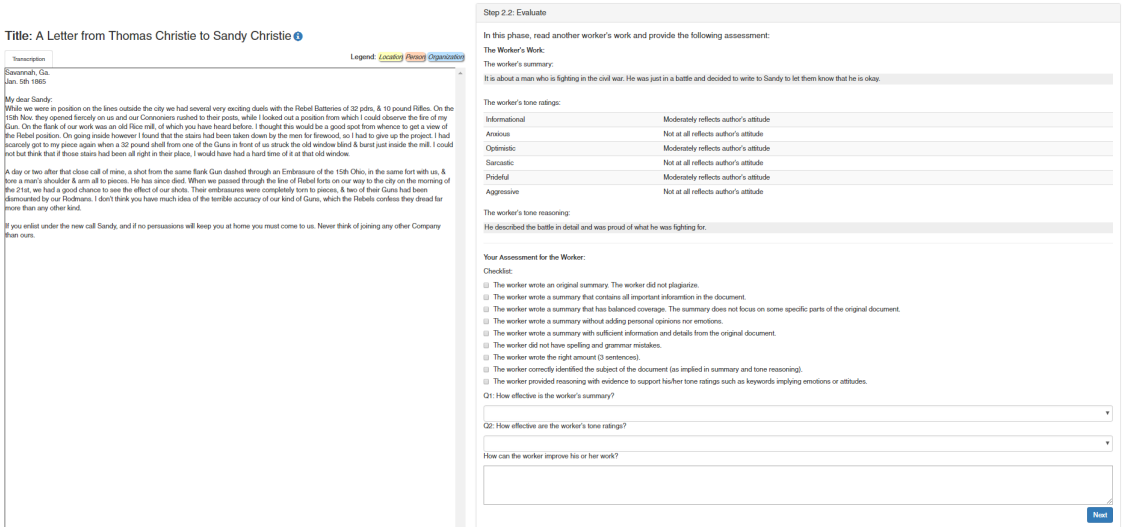


Figure 6.4: A screenshot of the RvD intervention for the summary-tone task

Dependent variables

To measure learning, I followed the same procedure used in previous SCIM-C studies (e.g., [70]) to compare the difference between the participant's score of the historical interpretation in the post-test and the pre-test. The interpretations were graded by two graders who were trained with the same materials used in previous SCIM-C studies and blind to the crowdsourcing techniques and crowdsourced tasks. The same grading rubric (see Appendix A for details) from prior SCIM-C studies was also used for grading. Interrater reliability was determined by comparing the graders' responses (binary yes/no) to the 12 scoring rubric questions across 60 interpretations from the pilot studies and calculating Cohen's κ. Cohen's κ ranges from 0.0 (agreement is no better than chance) to 1.0 (perfect agreement), and is appropriate for measuring interrater reliability for categorical data. The graders had a Kappa score of 0.89, indicating high reliability. To measure the quality of the summary, I used the score of the summary. The summaries were graded by the same two graders who were blind to the crowdsourcing techniques. They used the rubric developed from previous work and guidelines gathered from school writing centers, and approved by a history professor,

Historian B (see Appendix B for details). Interrater reliability was determined by comparing the graders' responses (yes or no) to the rubric questions across all unrevised summaries. The graders had a κ score of 0.83, indicating high reliability. Quality was divided into three categories: low (0–3), medium (4–6), and high (7–10) based on a 10-point scale. A high quality summary contains no or minor issues that do not affect reading. A medium quality summary misses some important information, detail or context. A low quality summary misses much of important information, detail and/or context. These categories were nominal labels to help make sense of the scores, but I used raw scores for all of the following data analyses. To measure the quality of the tones, I compared each worker's response with a gold standard response provided prior to the study by a history professor, Historian A. Specifically, I measured the Cohen's weighted κ between the crowd's response and the gold standard response. To measure the quality of the tags, I compared each worker's response with gold standard response also provided prior to the study by Historian A. Specifically, I measured the precision and recall of the tags created by the crowd. To measure the quality of connections (i.e., theme ratings), similarly, I measured the Cohen's weighted κ between the crowd's response and the gold standard response provided by Historian A. To measure how the work quality is affected by the crowdsourcing technique, I calculated the difference between revised and original work as quality change, if applicable (there was no revision for the Baseline condition). I also measured the crowd's efficiency in analyzing documents in terms of time and attempts as attrition. Time describes how long it takes for a task to be completed and is an indicator of how much effort the task requires. Attempts describes how many workers accept and return a HIT before it is completed and is an indicator of the perceived task difficulty.

6.5 Results

6.5.1 Learning: Only CrowdSCIM improves learning

The mean of the pre-test scores was 1.4 (sd=1.2) out of a maximum 12 points. The learning (score change between pre-test and post-test) of each task for each of the three crowdsourcing techniques is shown in Table 6.1 and Figure 6.5. For the Baseline technique, there is almost no learning; average learning scores for the Summary-tone, Tag and Connect tasks are 0.07, 0.17 and 0.13, respectively. For the CrowdSCIM technique, average learning scores for the Summary-tone, Tag, and Connect tasks are 1.9, 3.3, and 1.43, respectively. For the RvD technique, average learning scores for the Summary-tone, Tag, and Connect tasks are 0.60, 0.47, and 0.10, respectively. For the Shepherd technique, average learning scores for the Summary-tone, Tag, and Connect tasks are 0.27, 0.43, and 0.80, respectively. A two-way ANOVA showed a significant main effect of crowdsourcing technique ($F(3, 354)=26$, $p<0.01$), insignificant main effect of crowdsourced tasks ($F(2, 354)=2.5$, $p=0.08$), and a significant interaction effect ($F(6, 348)=8.5$, $p=0.01$) on learning. Since there was a significant interaction effect and I were interested in how these crowdsourcing techniques affect the quality based on the 5 measures, I ran a one-way ANOVA's for each of the tasks. To control the overall Type I error level (α_E) as 0.05, I used Bonferroni's adjustment for each ANOVA, whose Type I error level (α_I) became 0.017. For the Summary-tone task, a one-way ANOVA showed a significant effect of crowdsourcing technique ($F(3, 116)=7.5$, $p<0.01$). Post-hoc Tukey tests indicated that learning of CrowdSCIM was significantly higher than other three techniques (all $p<0.01$) and no significant difference among other three techniques. For the Tag task, a one-way ANOVA showed a significant effect of crowdsourcing technique ($F(3, 116)=19$, $p<0.01$). Post-hoc Tukey tests indicated that learning of CrowdSCIM was significantly higher than other three techniques (all $p<0.01$) with no significant differences among

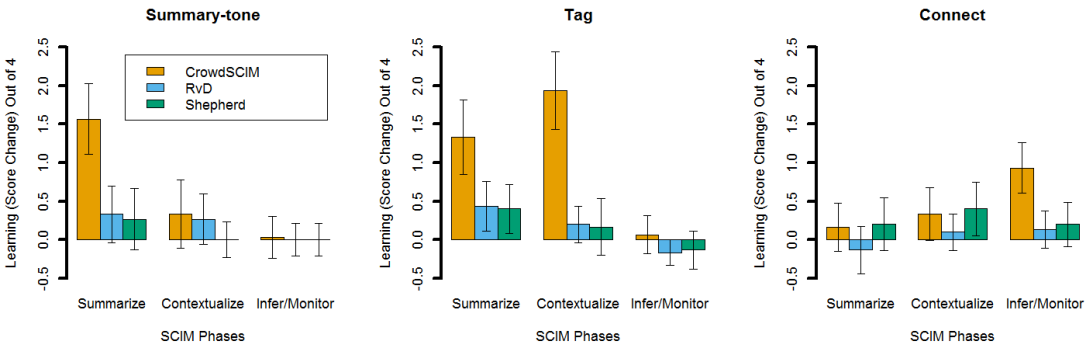


Figure 6.5: Individual phase learning across the crowdsourcing techniques

Task and Technique	Baseline		CrowdSCIM		RvD		Shepherd	
	mean	sd	mean	sd	mean	sd	mean	sd
Summary-tone	0.07	1.2	**1.9***	2.3	0.60	1.7	0.30	1.4
Tag	0.17	1.6	**3.3***	2.5	0.50	1.5	0.43	1.7
Connect	0.13	1.7	**1.4***	1.7	0.10	1.2	0.80	1.8

Table 6.1: Learning (score change) of different tasks across all crowdsourcing techniques (*: $p < 0.05$)

other three techniques. For the Connect task, a one-way ANOVA showed a significant effect of crowdsourcing technique ($F(3, 116)=4.5$, $p<0.01$). Post-hoc Tukey tests indicated that learning of CrowdSCIM was significantly higher than Baseline and RvD (both $p<0.01$) with no significant differences among other three techniques. To better understand what abilities workers learned, I divided overall learning into SCIM phases: Summarize, Contextualize, and Infer/Monitor, as shown in Figure 6.5. For CrowdSCIM, we can see that CrowdSCIM almost always creates learning gains, especially when task and phase are aligned, and never hurts learning (i.e., post-test worse than pre-test). In addition, CrowdSCIM in the Tag task also helped workers learn the Summarize ability. In contrast, RvD and Shepherd show much smaller learning gains and can actually hurt learning in some cases, e.g., Infer/Monitor in the Tag task for both RvD and Shepherd.

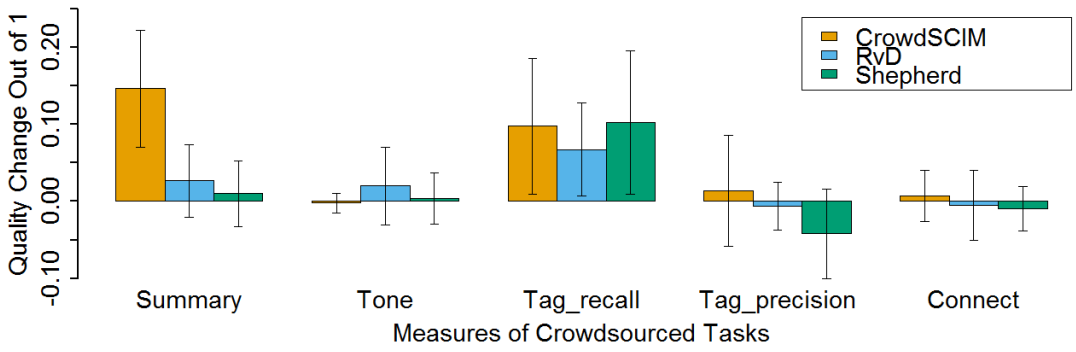


Figure 6.6: Quality change of the crowdsourcing techniques for each crowdsourced task

6.5.2 Quality: Only CrowdSCIM improves summary quality

Since there were five quality measures in the three tasks (summary, tone, recall of tag, precision of tag, and connect), I first normalized all the measures to a 0–1 scale and then conducted a two- way ANOVA. The two-way ANOVA showed a significant main effect of crowdsourcing technique ($F(3, 580)=3.9$, $p=0.01$), a significant effect of crowdsourcing task (the five measures) ($F(4, 580)=7.0$, $p<0.01$), and a significant interaction effect ($F(12, 580)=1.9$, $p=0.03$). Since there was a significant interaction effect and I were interested in how these crowdsourcing techniques affect the quality based on the five measures, I ran a one-way ANOVA for each of the measures. To control the overall Type I error level (α_E) as 0.05, I again used Bonferroni's adjustment for each ANOVA, whose Type I error level (α_I) became 0.01. The overall quality change of each crowdsourcing technique for each crowdsourced task is shown in Figure 6.6 and Table 6.2.

Summary: Similar original quality but CrowdSCIM brings quality change

The mean summary quality of the original work was 4.0 (sd=2.5) out of maximum 10 points. This mean corresponds to the lowest score in the medium quality category that still contains

Task	Summary		Tone		Tag (rec.)		Tag (pre.)		Connect	
Technique	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
CrowdSCIM	**0.15***	0.21	0.00	0.03	0.10	0.25	0.01	0.20	0.01	0.09
RvD	0.03	0.13	0.02	0.14	0.07	0.17	-0.01	0.09	0.00	0.13
Shepherd	0.01	0.12	0.00	0.19	0.10	0.26	-0.04	0.16	-0.01	0.08

Table 6.2: Quality change of the crowdsourcing techniques for each crowdsourced task (*: $p < 0.05$; Out of maximum 1.0)

some important information, detail, and context. The mean summary quality change was 1.5 for CrowdSCIM (sd=2.1), 0.3 for RvD (sd=1.3), and 0.1 for Shepherd (sd=1.2). A one-way ANOVA showed a significant main effect of crowdsourcing technique ($F(3, 116)=7.3$, $p<0.01$) on summary quality change. Post-hoc Tukey tests showed the summary quality change of CrowdSCIM was significantly higher than other techniques (all $p<0.01$) and no significant difference among other techniques.

Tone: Similar original quality without quality change

The mean tone rating quality of the original work was 0.54 (sd=0.30) out of maximum 1. The Cohen's κ of 0.54 is generally considered "moderate agreement" with the gold standard [86]. The mean quality changes in tone rating were 0.00 for CrowdSCIM (sd=0.03), 0.02 for RvD (sd=0.14), and 0.00 for Shepherd (sd=0.09). A one-way ANOVA showed no significant main effect ($F(3, 116)=0.41$, $p=0.75$) on quality change in tone rating.

Tag: Similar original quality without quality change

The mean recall of the original work was 0.59 (sd=0.28) out of maximum 1. The mean precision of the original work were 0.61 and 0.28 (out of maximum 1). The mean quality changes for recall were 0.10 for CrowdSCIM (sd=0.25), 0.07 for RvD (sd 0.17), and 0.10 for Shepherd (sd=0.26).Aone-wayANOVAshowed no significant main effect ($F(3, 116)=1.7$,

p=0.17) on quality changes for recall. The mean quality changes for precision were 0.01 for CrowdSCIM (sd=0.20), -0.01 for RvD (sd=0.09), and -0.04 for Shepherd (sd=0.16). A one-way ANOVA showed no significant main effect (F(3, 116)=0.91, p=0.44) of crowdsourcing technique on quality change for precision.

Connection: Similar original quality without quality change

The mean connection quality (theme rating) of the original work was 0.65 (sd=0.26) out of maximum 1. The κ value 0.65 is generally considered "substantial agreement" with the gold standard [86]. The mean quality changes for connection were 0.01 for CrowdSCIM (sd=0.09), 0.00 for RvD (sd=0.13), and -0.01 for Shepherd (sd=0.08). A one-way ANOVA showed no significant main effect (F(3, 116)=0.20, p=0.90) on quality change for connection.

6.5.3 Efficiency: Different efficiency but similar attrition

Time: Baseline requires the least time while CrowdSCIM needs the most

Except that the Baseline did not include intervention nor revision, all techniques contained pre-test, task, intervention, revision, and post-test. The time spent on each stage for each technique is shown in Table 6.3. Since I were interested in how these crowdsourcing techniques affect the efficiency of the tasks, I ran a two-way ANOVA for each of the task-related activities (task, intervention and revision). To control the overall Type I error level (α_E) as 0.05, I again used Bonferroni's adjustment for each ANOVA whose Type I error (α_I) became 0.017.

The mean time required to complete the pre-test was 6.2 minutes (sd=4.5). Overall, it took 6.4 minutes (sd=4.8) to complete a task. For each task, it took 7.3 minutes (sd=5.1) to

Stage	Pre-test		Task		Intervention		Revision		Post-test		Total	
Technique	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
Baseline	6.4	4.4	6.0	3.9	N/A		N/A		6.8	7.5	19	12
CrowdSCIM	5.8	3.5	7.1	6.6	**10***	6.9	1.8	3.1	7.5	7.3	33	18
RvD	6.6	5.9	6.0	3.6	2.8	1.8	1.5	1.7	6.6	5.4	24	13
Shepherd	6.0	3.8	6.2	4.6	1.9	1.4	1.9	2.0	6.3	4.3	23	10

Table 6.3: Time spent at different work stages across different crowdsourcing techniques (minutes; *: $p < 0.05$)

complete the Summary-tone task, 6.3 minutes (sd=5.6) to complete the Tag task, and 5.4 minutes (sd=3.3) to complete the Connect task. A two-way ANOVA showed a significant main effect of crowdsourced task ($F(2, 348)=4.8$, $p=0.01$) on time spent on the task. Post-hoc Tukey tests showed it took significantly more time to finish the Summary-tone than the Connect task ($p=0.01$). In general, it took 5.1 minutes (sd=5.7) to complete any of the three learning interventions. Across the three techniques, it took, on average, 10 minutes (sd=6.9) to complete the CrowdSCIM intervention, 2.8 minutes (sd=1.8) to complete the RvD intervention and 1.9 minutes (sd=1.4) to complete the Shepherd intervention. A two-way ANOVA showed a significant main effect of crowdsourcing technique ($F(2, 261)=120$, $p<0.01$) on time spent on the intervention. Post-hoc Tukey tests showed it took significantly more time to finish the CrowdSCIM intervention than the other two interventions for each of the crowdsourced tasks (all $p<0.01$). On average, it took 1.7 minutes (sd=2.4) to complete the revision of a crowdsourcing technique. For each of the three techniques, it took 1.8 minutes (sd=3.1) for CrowdSCIM, 1.5 minutes (sd=1.7) for RvD and 1.9 minutes (sd=2.0) for Shepherd. Two-way ANOVA showed a significant main ef- fect of crowdsourced task ($F(2, 261)=4.6$, $p=0.01$) on time spent on revision. Post-hoc Tukey tests showed it took significantly more time to finish the Summary-tone revision than the Connect revision ($p=0.01$). The mean time required to complete the post-test was 6.8 minutes (sd=6.3).

Attrition (via attempts): Extra work takes extra time but similar attrition rate

The mean attempts required to complete a task were 3.2 (sd=2.5). A two-way ANOVA showed no significant main effect of crowdsourced tasks nor techniques on attempts before a task is completed.

6.6 Discussion

6.6.1 Learning: CrowdSCIM supports learning while other techniques do not

The results of the pre-test scores (only 1.4 out of 12, on average) showed that crowd workers generally lacked sufficient historical thinking skills to write a strong historical interpretation for the given primary source. The learning results for the Baseline condition further suggested that an instructor's intervention may be necessary for Incite users to learn historical thinking. Example historical interpretations are shown in Appendix D. The CrowdSCIM results showed significant learning gains when the crowdsourced task and the SCIM phase were aligned (e.g., Summary-tone with Summarize, Tag with Contextualize, and Connect with Infer/Monitor). These results indicate that the iterative design process I employed through pilot studies was both effective and necessary. Both the first pilot study and the Baseline learning results from our evaluation showed that merely doing the crowdsourced tasks did not help with learning. Further, the second and third pilot studies showed that simply applying SCIM-C from the classroom to crowdsourced settings would not work, either. Only by decomposing the SCIM-C technique into micro-tasks could the learning gains be realized. Finally, the results of com- parison with RvD and Shepherd support the intuition that a domain expertise-related intervention is more effective than task-related interventions in achieving learning gains in

historical thinking skills. Looking at the most-learned ability for each task, CrowdSCIM improved worker learning by 1-2 points (out of maximum 4) for the corresponding phase in SCIM. That amount of learning is comparable to previous SCIM-C studies [70] in which the learning gains were 1.26 for Summarize, 1.48 for Contextualize, and 0.61 for both Infer and Monitor combined after three 2.5-hour tutorials across three instructional episodes. Although our study recruited novice workers from AMT and participants of prior SCIM-C studies were school students, the learning gains were comparable across these two very different participant pools, settings, and time frames. CrowdSCIM in the Tag task also helped workers learn the Summarize ability. This suggests that there is a strong correlation between the two phases and abilities. In contrast, the results of RvD and Shepherd showed insignificant learning, and these techniques could even hurt learning in some cases.

6.6.2 Quality: Crowd's work quality is moderate and CrowdSCIM improves summary

The quality results for summarization tasks showed that the participants in general were able to generate summaries of middling quality (4 out of 10). CrowdSCIM was able to improve the average summary quality to 5.5 (see Appendix D for an example). RvD and Shepherd did not improve the summary quality, in contrast to previous work [56, 156]. Unfamiliarity with historical primary sources might make it difficult for these workers to write a very good summary. The results of tone rating showed there was a moderate baseline agreement between the crowd and the expert historian. For the Connect task, the baseline crowd results were even better, showing substantial agreement with the expert. Because the baseline performance for these tasks is already reasonably good, improvement may not be necessary for some use cases. The recall results of the Tag task showed the baseline crowd was able to tag about 60% of the expert's tags. Due to the large number of responses

and tags to analyze, our calculations only recognized exact matches; i.e., minor differences such as “Va.” and “Va” were considered different. Therefore, this number should be seen as a lower bound because the recall could be higher with alias handling techniques. For historians, recall is generally more important than precision because they are accustomed to false positives, whereas relevant primary sources are rare and missing one is costly. Aside from the Summary task, none of the three crowdsourcing techniques improved quality results for any of the other tasks: tone, tag, or connect. Two possible explanations for CrowdSCIM's lack of effect are that the SCIM phases are too abstract for workers to immediately transfer to micro-tasks, or that there is a ceiling effect caused by solid initial performance. Notably, none of the learning techniques hurt work quality, either. 6.3 Efficiency: Extra work takes extra time but attrition is similar As expected, Baseline is the most efficient technique for the crowdsourced task, followed by Shepherd and RvD, and finally CrowdSCIM. CrowdSCIM's learning intervention took significantly longer (10 min) than RvD (3 min) or Shepherd (2 min). However, the revision step took the same amount of time for all three techniques (about 3 minutes). Further, across all SCIM phases and task types, the total task times were suitable for crowd work (10 minutes or less). I primarily included a revision step in all conditions to help quantify the advantage of each learning technique, but future work could explore omitting the initial task completion in CrowdSCIM to improve efficiency. Although the intervention and revision time of CrowdSCIM was longer, the attrition rate of CrowdSCIM was similar other techniques. This seemed to suggest that CrowdSCIM intervention provided some extra attraction to keep the attrition rate as the same level as others.

6.6.3 Trade-offs

No one technique works best for all situations Based on the results, there is no one technique among the four I evaluated that can serve all purposes, but depending on the requester's goals, some approaches work significantly better than others. In general, if the task design is learning-oriented, CrowdSCIM is the clear winner, because workers show the highest learning gains while producing work of similar or better quality, although this approach is slower than the others. If the design is quality-oriented, CrowdSCIM should be used for summary tasks, and Baseline for the other tasks, since all approaches perform similarly, but Baseline is fastest. If the design is efficiency-oriented, Baseline is the fastest, but workers will not learn anything, and summary quality will be degraded. Further, the "add-on" design of CrowdSCIM makes it easy to switch between Baseline and CrowdSCIM if the requester's needs change frequently.

6.7 Chapter Summary

As crowdsourcing markets become more popular and pervasive, researchers and practitioners have begun to seriously consider what future crowd work should ideally look like, suggesting some potential trade-offs such as learning and productivity. One the one hand, I would like to help workers learn and develop new skills. On the other hand, learning and skill development do not come without cost (e.g., immediate productivity or money). In this study, I investigated potential trade- offs between learning domain expertise and productivity for historical research, a domain that has seen little attention from crowdsourcing researchers. I adapted a technique from educational research to create a crowdsourcing workflow, CrowdSCIM, that allows novice crowd workers to learn historical thinking skills while completing useful historical research tasks. Results from our experiments showed that CrowdSCIM was

effective at helping workers learn domain expertise while producing work of equal or higher quality compared to baseline and prior work conditions. I also use CrowdSCIM as an example to discuss broader implications for future crowd work in terms of training and education and implications for history education and research.

7

Conclusion and Future Work

In this chapter, I conclude my dissertation and discuss future work. I begin by returning to my research questions and how each of the studies in my dissertation addresses them. I then state the major contributions offered by this research. Finally, I discuss some opportunities for future work in crowdsourcing history and beyond.

7.1 Addressing the Research Questions

7.1.1 Incite and class-sourcing

RQ 1a: How can we design a crowdsourcing system to both support historical research and history education?

RQ 1b: How do teachers and students use such a system?

The class-sourcing model and Incite presented in Chapter 3 provided answers to the two research questions. The class-sourcing model conceptualizes a win-win situation where educator historians outsource some of their research work to their students and the students get opportunities to participate in authentic research materials and experience.

The design of Incite, an implementation of the class-sourcing model, integrated advanced crowdsourced tasks beyond transcriptions along with embedded historical thinking phase.

The evaluation of Incite showed that the class-sourcing model worked as expected. Educator historians collected data beyond transcriptions useful for their research and students gained real experience of working with real pieces of history. In addition the the conceptual model, the key design of Incite was to identify the common element of historical research and history education, that is, primary sources. And the next step is to identify what kind of analysis would be useful for historical research and what kind of expertise students are expected to acquire. The third step is to match the two parts and design corresponding crowdsourced tasks.

The idea of class-sourcing is not unique to history domain. In many other domains, researchers are often educators as well such as university professors. The results of this study encourages more implementations of class-sourcing model in other domains where researcher have dual roles. This model provides benefits for both research and education purposes.

With the opensourced Incite, I would also encourage historians and possibly their institutions to class-source their personal digital archives. One of the design considerations of Incite was to make it a plugin to a widely-used content management system, Omeka, in digital humanities. Ideally, the installation of Incite is process of only a few clicks if the archive is hosted with Omeka.

7.1.2 RAP: Scaling up crowdsourced historical connections

RQ 2a: How well does the novice crowd make connections between historical primary sources and high-level research topics?

RQ 2b: How can crowds connect related primary sources to scholarly topics as accurately as historians?

RQ 2c: How can crowdsourcing systems identify opportunities for public history interven-

tions?

The RAP study presented in Chapter 4 provided empirical results about how well novice perform for a higher-level task, that is, connecting primary sources to research topics. While novice crowds can already help with this task, this study show that majority vote could further improve the quality. And with RAP, novice crowds could make connections as accurately as historians. In addition, this study also showed that using novice crowds provided new opportunities of learning interventions for historians to engage in public history. With RAP, we could even prioritize these opportunities.

7.1.3 Zooniverse: Scaling up Complex Crowdsourced Transcriptions

RQ 3: How can we scale up the analysis of complex historical documents while minimizing cost and effort?

The Zooniverse study presented in Chapter 5 disclosed challenges, process and common trade-offs of designing a crowdsourcing project to handle a large dataset with complex document variants. While there were examples and tools available, these resources were not practical for a typical historian without related expertise, funding and effort. This study presented a generalizable and practical solution by providing a set of design guidelines with examples and rationale.

While we try to make crowdsourcing a more and more powerful (research) tool, we might also want to consider work that would lower the learning curve so that more and more people (researchers) can actually benefit from the these powerful crowdsourcing tools and systems. As discussed in Chapter 5, many existing tools to help design crowdsourcing workflows require additional expertise of programming and design, which experts from other domains

may not have.

With the results of this study (thousands of registered users and comments in talk boards), historians are encouraged to take advantage of the design guidelines and crowds on crowdsourcing platform such as Zooniverse to help advance their historical research and public history.

7.1.4 CrowdSCIM: Scaling up learning

RQ 4a: How can we help crowd workers learn domain expertise in history domain (historical thinking)?

RQ 4b: How does the introduction of learning affect work quality and efficiency of crowdsourcing?

The CrowdSCIM study presented in Chapter 6 demonstrated how we could help crowd worker learn historical thinking and the evaluation indicated how learning may improve quality and lower efficiency. The add-on design of CrowdSCIM also proved to be effective in retaining quality. The CrowdSCIM workflow is also generalizable to other similar text analysis domains such as journalism.

This study showed that the negative impact of the introduction of learning into a crowdsourcing can be negligible, especially with the add-on design that separates the original tasks and learning interventions. On the other hand, learning can bring benefits to the crowdsourced results such as increasing the quality and enable the crowd to do more complex tasks.

Additionally, introducing learning may help provide a more beneficial and ethical crowdsourcing environment for crowd workers. Beyond immediate tasks, workers can learn important skills that may affect the rest of their lives.

The results show that CrowdSCIM is an effective tool for learning historical thinking so educator historians might want to consider using it in their classes as a first-pass so that they can focus more on more advanced parts. In addition, the empirical results of the study also suggest that historians be confident in crowdsourcing that crowds can help with at least the first round of analyses.

7.1.5 Connections among the four studies

To address the first challenge to supporting historical research and history education discussed in Chapter 1, my first study class-sourcing and Incite focused on classroom settings as the first step to explore supporting historians' current practices with crowdsourcing techniques. The dynamics and interactions between the instructor and students are important factors of the design. Therefore, Incite provides many features to support this goal such as group management and progress tracking. From the feedback of the participating instructors, we can see Incite provided a few benefits including 1) introducing more complex crowdsourced tasks, 2) more complex documents, and 3) teaching and learning history.

With the success of class-sourcing and Incite, my following studies address other two challenges by focusing on scaling up these three benefits with more general crowds such as paid crowd workers and volunteers when instructors are not available or have limited amount of time to help cultivate these relationships. RAP, introduced in Chapter 4, investigated how we can scale up more complex crowdsourcing analysis beyond transcription using connecting as the target crowdsourced task with paid crowd workers. Zooniverse, presented in Chapter 5, explored how we can scale up more complex documents with volunteers. CrowdSCIM, described in Chapter 6, examined how we can scale up crowds' learning historical thinking while contributing to historical research.

CrowdSCIM is fundamentally a learning add-on for crowdsourcing tasks that helps crowds learning historical thinking skills while performing tasks. It could be added onto connection tasks, as presented in RAP, or transcription related tasks, as presented in Zooniverse study and those workers would also learning historical thinking while performing the same tasks as well or better. Taking RAP and Zooniverse studies for examples, the connection task in RAP is very similar to the one tested in CrowdSCIM study so I expect the add-on process of CrowdSCIM for the connect task can be directly applied to the connection task in RAP to obtain the learning benefit. The Summarize phase of CrowdSCIM can be add-on to the transcription task in Zooniverse since it is similar to the design of the summary task in Incite. I would expect to see learning although the learning might be different due to the slight difference between the two crowdsourced tasks. If we want to ensure the same learning, it is easy to add summary and tone sub-tasks to the crowdsourced task on Zooniverse because the summary and tone sub-tasks are generic and do not need to change any documents or existing workflows.

7.2 Broader Implications

7.2.1 Implications for historical research and history education

Historical documents are critical sources for both scholarly research and learning in the domain of history [129, 130], and teaching students to think like a historian is one of the main goals in history education [73, 94, 144]. From Incite, RAP to Zooniverse, I evaluated how three different types of crowd (students, paid workers, and volunteers) can support historical research. There are trade-offs of using these different crowds depending on the goals.

While I only evaluated CrowdSCIM with paid crowds to support rapid iteration and scaling up, I expect that crowds of traditional history students could also benefit from using CrowdSCIM. SCIM-C is designed to be a sequential process that a student should follow from Summarize to Corroborate, but our experiments suggest that different phases in historical thinking may be learned or improved individually. CrowdSCIM may also offer a useful supplement to the classroom teaching. For example, it may be used as a first pass, allowing the instructor to focus on learning material that CrowdSCIM does not provide. Or it may be used in a targeted way to improve one specific ability of historical thinking that a student or teacher identifies as weaker than the others by working on corresponding tasks with CrowdSCIM. Finally, in settings where the ratio of number of students per expert is high, such as in MOOCs or citizen science projects, CrowdSCIM may provide a scalable way for students to learn historical thinking with minimal intervention from experts.

Quality results of RAP and CrowdSCIM also show that the crowd can already do a reasonable job for most crowdsourced tasks, although there is often room for improvement. For example, the summary captures key information of the primary source, so it can help the historian quickly decide whether a source is worth extra attention. Taking the test documents as examples, the average length of the documents is 253.2 words and the average length of a crowdsourced summary is 49.9 words. This suggests a historian could save about 80% of the reading time while searching for relevant documents. In addition, the tagging results show that crowds have moderate to substantial agreement with a historian in identifying documents that are relevant to the historian's topics of interest.

Scenarios

Through these four studies, I worked with several historians and experts in digital humanities who have different experiences and expectations with crowdsourcing. In this section, I would

like to discuss how they might (want to) take advantages of the results and systems.

Class-sourcing model along with Incite is a good start for those historians who are uncertain with the use of crowdsourcing because it minimizes the the change of existing research and teaching practices. Furthermore, Mapping the Fourth project provides guidelines for how to integrating Incite into classroom use and suggestions for various types of assignments. This model also gives students opportunities to interact with real pieces of history and may reveal students' confusions so this is also helpful for those who are interested in improving teaching and learning experiences to their classrooms. The main drawback of this model is that the scalability may not be very good depending on the size of the class. For example, historian Ed Gitre was interested in using crowdsourcing in his research and class but did not have much direct experience with crowdsourcing projects before. He asked to participate in testing Incite in his classes and was able to incorporate Incite into his curriculum. With four classes, he had about 3,000 documents analyzed by students.

Read-Agree-Predict (RAP) from Chapter 4 can be a useful tool to provide high quality of connections between primary sources and research topics and identify opportunities for public history interventions on demand. This is particular useful for those who are willing to pay and/or need some quick help externally. While this technique is scalable and can be requested on demand, it costs money. Historian Paul Quigley was excited to see how RAP can help him connect primary sources to the topics he uses for research and would be interested in using RAP when appropriate.

Design guidelines presented in Chapter 5 may be valuable for those who have experiences with digital humanities and citizen scientists/archivists and have a lot of documents to be processed. The guidelines with crowdsourcing platforms such as Zooniverse can provide high scalability for processing a large amount of documents. This approach is also a good source for those who look for engaging the public and public history intervention. A practical

example is also from historian Ed Gitre. With a large amount of documents, class-sourcing and Incite might not scale wel so volunteer-based Zooniverse became an ideal source of human power. By weighing the trade-offs presented in the design guidelines, I was able to build the American Soldier project on Zooniverse.

CrowSCIM from Chapter 6 can be used with Incite with its flexible add-on design. Historians might want to use this as a supplemental material for teaching historical thinking. It can be used as a first pass so historians can teach more advanced materials or historians can use it to provide practice opportunities. Afterwards, they can decide whether to keep the learning depending on their needs.

7.2.2 Implications for crowdsourcing research

While Shepherd [56] and RvD [156] have demonstrated significant value in other task domains, the results suggest they are not well-suited for promoting learning or improving output quality in historical research. Why not? While neither Shepherd or RvD was designed for historical research, both were previously evaluated with writing or summarization tasks, so it is perhaps most surprising that summarization was the only task where CrowdSCIM yielded significantly better (vs. similar) work quality. This result suggests that writing about historical primary sources creates unique challenges. I propose two reasons why CrowdSCIM is better-suited for learning (and, in the case of summaries, doing) analysis of historical documents. One is that the reflective questions in step 2 provide scaffolding to help workers engage in deeper thinking and stimulate higher-level cognitive processes. Specifically, the questions provide structure in the form of a "specific cognitive strategy" [70] that reduces the initial complexity of an open-ended process. They also problematize the task by drawing workers' attention to issues they might not normally consider [70]. A second reason may be

that the practice interpretation in step 3 helps workers to synthesize and internalize their new expertise before attempting to transfer it to a new application (i.e., the revision). This mental organization may make the expertise more readily available for the post-test and beyond.

Previous work (e.g., [88, 109]) suggests that learning domain knowledge (factual knowledge) may hinder work quality, but the results show learning domain expertise (such as analytical skills and thinking strategies) may help with some task types, such as writing a summary, without impeding work quality for other types of tasks. While the main focus of CrowdSCIM study is using crowdsourcing to support historical research, the CrowdSCIM workflow may be generalize to other domains focused on sensemaking and analysis of primary source documents. Historians have been described as "detectives searching for evidence among primary sources to a mystery that can never be completely solved" [145], which shares similarities with other investigative domains such as journalism, law enforcement, and political fact-checking. To adapt CrowdSCIM for other domains, I envision a generalized workflow comprising 1) an unmodified initial text analysis task, 2) a scaffolded learning intervention, 3) a practice task, and 4) a revised attempt at the initial task. Given my focus in this dissertation on supporting historical research, CrowdSCIM's learning intervention in step 2 was adapted from SCIM-C's historical thinking prompts. For other domains, however, these prompts could be substituted for alternative domain-specific reflective questions most relevant to the given task. For example, a CrowdSCIM variant for supporting crowdsourced journalism might provide reflective questions derived from ethnographic studies of expert practice for workers to learn to review user-generated content for newsworthy themes [95, 133]. A crowdsourced fact-checking effort could ask questions based on verification principles to help crowds learn to research politically oriented claims and assess the reliability of their sources [84, 119]. In a law enforcement context, novice workers could learn to analyze police reports for patterns of

interest, guided by reflective questions about motive, opportunity, and lack of alibi [63, 64]. My experiences with CrowdSCIM and historical documents also suggest some caveats in adapting this approach for other domains. First, although reflective questions for the target task may already exist, they were likely developed for a different audience, such as students or junior practitioners, and will likely require iterative design to repurpose for novice crowd workers in a micro- tasking context. My pilot studies showed that SCIM-C, while effective in traditional classrooms, required extensive modularization for crowds. Second, I recommend aligning each task with the most relevant subset of reflective questions; with CrowdSCIM, proper alignment made the difference between productivity gains and learning losses. While future work is needed, I anticipate that CrowdSCIM's specific orientation towards historical thinking and working with primary sources, as well as its more general approach to decomposing complex thought processes into just-in-time learning interventions, may be applicable to these and other domains sharing similar processes.

Future of crowd work

The results from CrowdSCIM demonstrate possibilities for a better "future of crowd work", [83]. Instead of doing repetitive, low-payment tasks, crowd workers can learn and develop new skills to steadily handle more complex and creative tasks and improve work quality and payment. When learning can improve the work quality, as in the Summary task, learning may be seen as "training" directly related to the work, and the requester could pay for the training as in a traditional job market. When the learning does not improve the work quality, as in the Tone, Tag, and Connect tasks, learning may be seen as "education" not directly related to the work, and the requester may choose not to pay for it, but rather provide it as free education. At the same time, crowd workers can also decide if they want to do more tasks to get paid or spend the time with skill development.

7.3 Contributions

My dissertation provides a crowdsourcing system to support historical research and history education in a classroom setting, a new crowdsourcing algorithm for relevance judgements and new insights into designs of crowdsourcing workflows and a new crowdsourcing system to support crowd learning while doing crowdsourced tasks.

Specifically, I contributed:

1. a crowdsourcing system, Incite, to support historical research and history education in a classroom setting (Chapter 3);
2. a comparative investigation of effects of two common crowdsourced tasks (tagging and summarizing) in digital archives on quality of crowdsourced connections and a crowdsourcing algorithm to accurately connect historical primary sources to high-level topics/themes historians use for their research (Chapter 4);
3. a set of design guidelines explaining potential trade-offs and suggesting sweet spots to help minimize cost and effort required for designing crowdsourced projects (Chapter 5);
4. a comparative investigation of effects of crowdsourcing workflows on quality of results and historical thinking (Chapter 6).

7.4 Future work

While my dissertation focuses on history domain, many of the ideas and systems should be generalizable to other domains. For example, the class-sourcing model and the workflow design of CrowdSCIM should work well for other domains focusing on text analysis.

A big theme of my dissertation is the learning component in crowdsourcing. Considering how education is emphasized in our society, the introduction of learning to crowdsourcing can create many win-win situations. There are many students looking for opportunities and work to practice and gain real experience. If crowdsourced tasks can provide these learning opportunities, I can have a very large (volunteer) based crowd and these crowds may even want to pay to do these tasks. Scalable crowd learning workflows may also benefit more people and can be used in online systems such as MOOCs to help solve the disproportion between the number of available teachers and participating students. While crowdsourcing more complex tasks is a desirable goal, learning can play a more important role because learning can help more people acquire required skills and expertise to work on more complex tasks.

Bibliography

[1] About the Field. URL http://ncph.org/what-is-public-history/about-the-field/.

[2] African American Civil War Soldiers. URL https://www.zooniverse.org/projects/usct/african-american-civil-war-soldiers/.

[3] Edit Articles | Citizen Archivist. URL https://www.archives.gov/citizen-archivist/edit/.

[4] Mapping the Fourth of July | Civil War Crowdsourcing | Virginia Tech. URL july4.civilwar.vt.edu.

[5] National Endowment for the Humanities (NEH) Funding Levels. URL http://www.humanitiesIndicators.org/content/indicatorDoc.aspx?i=75.

[6] DIYHistory | Transcribe, . URL https://diyhistory.lib.uiowa.edu/.

[7] Making History - Transcribe, . URL http://www.virginiamemory.com/transcribe/.

[8] Smithsonian Digital Volunteers, . URL https://transcription.si.edu/.

[9] Old Weather. URL https://www.oldweather.org/.

[10] Operation War Diary. URL https://www.operationwardiary.org/.

[11] steve.museum research report available: Tagging, Folksonomy and Art Museums | museumsandtheweb.com. URL http://www.museumsandtheweb.com/blog/jtrant/stevemuseum_research_report_available_tagging_fo.html.

[12] Trove crowdsourcing behaviour | National Library of Australia, . URL https://www.nla.gov.au/our-publications/staff-papers/trove-crowdsourcing-behaviour.

[13] Trove, . URL http://trove.nla.gov.au/.

[14] Using Archives: A Guide to Effective Research | Society of American Archivists. URL http://www2.archivists.org/usingarchives.

[15] Whats on the menu? URL http://menus.nypl.org/.

[16] Zooniverse, . URL https://www.zooniverse.org/about/publications.

[17] Zooniverse, . URL https://www.zooniverse.org/.

[18] For the common good: The Library of Congress Flickr pilot project. 2008. URL http://www.loc.gov/rr/print/flickr_report_final.pdf.

[19] THE STATE OF THE HUMANITIES: FUNDING 2014. Technical report, Humanities Indicators, American Academy of Arts & Sciences, 2014. URL http://www.humanitiesindicators.org/binaries/pdf/HI_FundingReport2014.pdf.

[20] Omar Alonso, Daniel E. Rose, and Benjamin Stewart. Crowdsourcing for Relevance Evaluation. *SIGIR Forum*, 42(2):9–15, November 2008. ISSN 0163-5840. doi: 10.1145/1480506.1480508. URL http://doi.acm.org/10.1145/1480506.1480508.

[21] John R. Anderson, C. Franklin Boyle, and Brian J. Reiser. Intelligent tutoring systems. *Science(Washington)*, 228(4698):456–462, 1985. URL http://www.academia.edu/download/31310363/Science_1985_Anderson.pdf.

[22] John R. Anderson, C. Franklin Boyle, Albert T. Corbett, and Matthew W. Lewis. Cognitive modeling and intelligent tutoring. *Artificial intelligence*, 42(1):7–49, 1990. URL http://www.sciencedirect.com/science/article/pii/000437029090093F.

[23] Paul AndrÃ©, Haoqi Zhang, Juho Kim, Lydia Chilton, Steven P. Dow, and Robert C. Miller. Community Clustering: Leveraging an Academic Crowd to Form Coherent Conference Sessions. In *First AAAI Conference on Human Computation and Crowdsourcing*, March 2013. URL http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7512.

[24] Paul AndrÃ©, Aniket Kittur, and Steven P. Dow. Crowd Synthesis: Extracting Categories and Clusters from Complex Data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 989–998, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2540-0. doi: 10.1145/2531602.2531653. URL http://doi.acm.org/10.1145/2531602.2531653.

[25] Paul AndrÃ©, Robert E. Kraut, and Aniket Kittur. Effects of Simultaneous and Sequential Work Structures on Distributed Collaborative Interdependent Tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 139–148, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2557158. URL http://doi.acm.org/10.1145/2556288.2557158.

[26] Sasha Barab and Kurt Squire. Design-based research: Putting a stake in the ground. *The journal of the learning sciences*, 13(1):1–14, 2004.

[27] Keith C. Barton and Linda S. Levstik. *Teaching history for the common good*. Routledge, 2004.

[28] Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(3):351–368, 2012. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpr057. URL https://www.cambridge.org/core/journals/political-analysis/article/

evaluating-online-labor-markets-for-experimental-research-amazoncoms-mechanical-tu
348F95C0FBCF21C3B37D66EB432F3BA5.

[29] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. Soylent: A Word Processor with a Crowd Inside. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 313–322, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0271-5. doi: 10.1145/1866029.1866078. URL http://doi.acm.org/10.1145/1866029.1866078.

[30] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=944919.944937.

[31] Samuel A. Bobrow and Gordon H. Bower. Comprehension and recall of sentences. *Journal of Experimental Psychology*, 80(3, Pt.1):455–461, 1969. ISSN 0022-1015. doi: 10.1037/h0027461.

[32] J. Bohannon. Psychologists grow increasingly dependent on online research subjects. *Science Magazine*, 7, 2016.

[33] Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, September 2004. ISSN 0031-3203. doi: 10.1016/j.patcog.2004.03.009. URL http://www.sciencedirect.com/science/article/pii/S0031320304001074.

[34] Jonathan Bragg, Mausam, and Daniel S. Weld. Crowdsourcing Multi-Label Classification for Taxonomy Creation. In *First AAAI Conference on Human Computation and Crowdsourcing*, November 2013. URL http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7560.

[35] H. W. Brands. Response to Hochschild. *Historically Speaking*, 9(4):6–7, 2008. ISSN 1944-6438. doi: 10.1353/hsp.2008.0063. URL http://muse.jhu.edu/content/crossref/journals/historically_speaking/v009/9.4.brands.html.

[36] Burke H. Bretzing and Raymond W. Kulhavy. Notetaking and depth of processing. *Contemporary Educational Psychology*, 4(2):145–153, April 1979. ISSN 0361-476X. doi: 10.1016/0361-476X(79)90069-9. URL http://www.sciencedirect.com/science/article/pii/0361476X79900699.

[37] Carrie J. Cai, Shamsi T. Iqbal, and Jaime Teevan. Chain Reactions: The Impact of Order on Microtask Chains. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 3143–3154, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858237. URL http://doi.acm.org/10.1145/2858036.2858237.

[38] Joseph Chee Chang, Aniket Kittur, and Nathan Hahn. Alloy: Clustering with Crowds and Computation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 3180–3191, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858411. URL http://doi.acm.org/10.1145/2858036.2858411.

[39] Chen Chen, Xiaojun Meng, Shengdong Zhao, and Morten Fjeld. ReTool: Interactive Microtask and Workflow Design Through Demonstration. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 3551–3556, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3025969. URL http://doi.acm.org/10.1145/3025453.3025969.

[40] Justin Cheng and Michael S. Bernstein. Flock: Hybrid Crowd-Machine Learning Classifiers. In *Proceedings of the 18th ACM Conference on Computer Supported Co-*

operative Work & Social Computing, CSCW '15, pages 600–611, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-2922-4. doi: 10.1145/2675133.2675214. URL http://doi.acm.org/10.1145/2675133.2675214.

[41] E. H. Chi, L. Hong, J. Heiser, and S. K. Card. Scentindex: Conceptually Reorganizing Subject Indexes for Reading. In *2006 IEEE Symposium On Visual Analytics Science And Technology*, pages 159–166, October 2006. doi: 10.1109/VAST.2006.261418.

[42] Ed H. Chi, Lichan Hong, Michelle Gumbrecht, and Stuart K. Card. ScentHighlights: Highlighting Conceptually-related Sentences During Reading. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, IUI '05, pages 272–274, New York, NY, USA, 2005. ACM. ISBN 978-1-58113-894-8. doi: 10.1145/1040830.1040895. URL http://doi.acm.org/10.1145/1040830.1040895.

[43] Ed H. Chi, Michelle Gumbrecht, and Lichan Hong. Visual Foraging of Highlighted Text: An Eye-Tracking Study. In Julie A. Jacko, editor, *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments*, number 4552 in Lecture Notes in Computer Science, pages 589–598. Springer Berlin Heidelberg, July 2007. ISBN 978-3-540-73108-5 978-3-540-73110-8. doi: 10.1007/978-3-540-73110-8_64. URL http://link.springer.com/chapter/10.1007/978-3-540-73110-8_64.

[44] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. Cascade: Crowdsourcing Taxonomy Creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1999–2008, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1899-0. doi: 10.1145/2470654.2466265. URL http://doi.acm.org/10.1145/2470654.2466265.

[45] D. Coetzee, Seongtaek Lim, Armando Fox, Björn Hartmann, and Marti A. Hearst. Structuring Interactions for Large-Scale Synchronous Peer Learning. In *Proceedings*

of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW '15, pages 1139–1152, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-2922-4. doi: 10.1145/2675133.2675251. URL http://doi.acm.org/10.1145/2675133.2675251.

[46] Robin George Collingwood and Willem J. van der Dussen. *The idea of history*. Oxford University Press on Demand, 1993.

[47] Allan Collins, John Seely Brown, and Susan E. Newman. Cognitive apprenticeship: Teaching the craft of reading, writing and mathematics. *Thinking: The Journal of Philosophy for Children*, 8(1):2–10, 1988.

[48] National Research Council and others. *How people learn: Brain, mind, experience, and school: Expanded edition*. National Academies Press, 2000. URL https://books.google.com/books?hl=en&lr=&id=QZb7PnTgSCgC&oi=fnd&pg=PR1&dq=bransford+how+people+learn&ots=FsQVkIesZE&sig=qESNaxmqFmysC8uqFFdNdTvJ2LI.

[49] Fergus I. M. Craik and Robert S. Lockhart. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11(6):671–684, December 1972. ISSN 0022-5371. doi: 10.1016/S0022-5371(72)80001-X. URL http://www.sciencedirect.com/science/article/pii/S002253717280001X.

[50] Fergus I. M. Craik and Endel Tulving. Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, 104(3):268–294, 1975. ISSN 1939-2222 0096-3445. doi: 10.1037/0096-3445.104.3.268.

[51] Murray S. Davis. That's Interesting: Towards a Phenomenology of Sociology and a Sociology of Phenomenology. *Philosophy of the Social Sciences*, 1(4):309–344, December 1971. ISSN 0048-3931. URL http://search.proquest.com/docview/1300114738/citation/4774E4B025D413DPQ/1.

[52] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, September 1990. ISSN 1097-4571. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9. URL http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9/abstract.

[53] Gianluca Demartini, Djellel Eddine Difallah, and Philippe CudrÃ©-Mauroux. ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-scale Entity Linking. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 469–478, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1229-5. doi: 10.1145/2187836.2187900. URL http://doi.acm.org/10.1145/2187836.2187900.

[54] Mira Dontcheva, Robert R. Morris, Joel R. Brandt, and Elizabeth M. Gerber. Combining Crowdsourcing and Learning to Improve Engagement and Performance. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 3379–3388, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2557217. URL http://doi.acm.org/10.1145/2556288.2557217.

[55] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. Toward a Learning Science for Complex Crowdsourcing Tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2623–2634, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858268. URL http://doi.acm.org/10.1145/2858036.2858268.

[56] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. Shepherding the

Crowd Yields Better Work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1013–1022, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1086-4. doi: 10.1145/2145204.2145355. URL http://doi.acm.org/10.1145/2145204.2145355.

[57] Ryan Drapeau, Lydia B. Chilton, Jonathan Bragg, and Daniel S. Weld. MicroTalk: Using Argumentation to Improve Crowdsourcing Accuracy. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2016. URL http://www.cs.washington.edu/ai/pubs/drapeau-hcomp16.pdf.

[58] B. Farrimond, S. Presland, J. Bonar-Law, and F. Pogson. Making History Happen: Spatiotemporal Data Visualization for Historians. In *Second UKSIM European Symposium on Computer Modeling and Simulation, 2008. EMS '08*, pages 424–429, September 2008. doi: 10.1109/EMS.2008.42.

[59] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219885. URL https://doi.org/10.3115/1219840.1219885.

[60] J. G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 347–354, December 1997. doi: 10.1109/ASRU.1997.659110.

[61] Elena L. Glassman, Juho Kim, Andrés Monroy-Hernández, and Meredith Ringel Morris. Mudslide: A Spatially Anchored Census of Student Confusion for On-

line Lecture Videos. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1555–1564, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3145-6. doi: 10.1145/2702123.2702304. URL http://doi.acm.org/10.1145/2702123.2702304.

[62] Elena L. Glassman, Aaron Lin, Carrie J. Cai, and Robert C. Miller. Learnersourcing Personalized Hints. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 1626–1636, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3592-8. doi: 10.1145/2818048.2820011. URL http://doi.acm.org/10.1145/2818048.2820011.

[63] Nitesh Goyal and Susan R. Fussell. Designing for Collaborative Sensemaking: Leveraging Human Cognition For Complex Tasks. *arXiv preprint arXiv:1511.05737*, 2015.

[64] Nitesh Goyal and Susan R. Fussell. Effects of Sensemaking Translucence on Distributed Collaborative Analysis. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 288–302, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3592-8. doi: 10.1145/2818048.2820071. URL http://doi.acm.org/10.1145/2818048.2820071.

[65] Catherine Grady and Matthew Lease. Crowdsourcing Document Relevance Assessment with Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 172–179, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1866696.1866723.

[66] Michael D. Greenberg, Matthew W. Easterday, and Elizabeth M. Gerber. Critiki: A Scaffolded Approach to Gathering Design Feedback from Paid Crowdworkers. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, C&C

'15, pages 235–244, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3598-0. doi: 10.1145/2757226.2757249. URL http://doi.acm.org/10.1145/2757226.2757249.

[67] Stuart Greene. Students as authors in the study of history. In *Teaching and Learning in History*, pages 137–170. Routledge, 1994.

[68] Derek L. Hansen, Patrick J. Schone, Douglas Corey, Matthew Reid, and Jake Gehring. Quality Control Mechanisms for Crowdsourcing: Peer Review, Arbitration, & Expertise at Familysearch Indexing. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 649–660, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1331-5. doi: 10.1145/2441776.2441848. URL http://doi.acm.org/10.1145/2441776.2441848.

[69] Simon Hengchen, Mathias Coeckelbergs, Seth van Hooland, Ruben Verborgh, and Thomas Steiner. Exploring archives with probabilistic models: Topic Modelling for the valorisation of digitised archives of the European Commission. In *First Workshop Â«Computational Archival Science: digital records in the age of big dataÂ», Washington*, volume 8, 2016. URL http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2016/05/1.pdf.

[70] David Hicks and Peter E. Doolittle. Fostering Analysis in Historical Inquiry Through Multimedia Embedded Scaffolding. *Theory & Research in Social Education*, 36(3): 206–232, July 2008. ISSN 0093-3104. doi: 10.1080/00933104.2008.10473373. URL http://dx.doi.org/10.1080/00933104.2008.10473373.

[71] David Hicks, Peter E. Doolittle, and E. Thomas Ewing. The SCIM-C strategy: expert historians, historical inquiry, and multimedia. *Social Education*, 68(3):221–226, April 2004. ISSN 00377724.

[72] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM. ISBN 978-1-58113-096-6. doi: 10.1145/312624.312649. URL http://doi.acm.org/10.1145/312624.312649.

[73] Cynthia Hynd, Jodi Patrick Holschuh, and Betty P. Hubbard. Thinking like a historian: College students' reading of multiple historical documents. *Journal of Literacy Research*, 36(2):141–176, 2004. URL http://jlr.sagepub.com/content/36/2/141.short.

[74] Michael J. Jacobson and Rand J. Spiro. Hypertext Learning Environments, Cognitive Flexibility, and the Transfer of Complex Knowledge: An Empirical Investigation. *Journal of Educational Computing Research*, 12(4):301–333, June 1995. ISSN 0735-6331. doi: 10.2190/4T1B-HBP0-3F7E-J4PN. URL http://journals.sagepub.com/doi/abs/10.2190/4T1B-HBP0-3F7E-J4PN.

[75] Michael J. Jacobson, Chrystalla Maouri, Punyashloke Mishra, and Christopher Kolar. Learning with Hypertext Learning Environments: Theory, Design, and Research. *J. Educ. Multimedia Hypermedia*, 4(4):321–364, December 1995. ISSN 1055-8896. URL http://dl.acm.org/citation.cfm?id=227170.227173.

[76] Charlene Jennett, Laure Kloetzer, Daniel Schneider, Ioanna Iacovides, Anna Cox, Margaret Gold, Brian Fuchs, Alexandra Eveleigh, Kathleen Methieu, Zoya Ajani, and others. Motivations, learning and creativity in online citizen science. *Journal of Science Communication*, 15(3), 2016. URL http://oro.open.ac.uk/47008/.

[77] Gabriella Kazai. In Search of Quality in Crowdsourcing for Search Engine Evaluation. In *Advances in Information Retrieval*, pages 165–176. Springer, Berlin, Heidelberg,

April 2011. doi: 10.1007/978-3-642-20161-5_17. URL http://link.springer.com/chapter/10.1007/978-3-642-20161-5_17.

[78] Juho Kim and others. *Learnersourcing: improving learning with collective learner activity*. PhD thesis, Massachusetts Institute of Technology, 2015. URL http://dspace.mit.edu/handle/1721.1/101464.

[79] Juho Kim, Robert C. Miller, and Krzysztof Z. Gajos. Learnersourcing Subgoal Labeling to Support Learning from How-to Videos. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 685–690, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1952-2. doi: 10.1145/2468356.2468477. URL http://doi.acm.org/10.1145/2468356.2468477.

[80] Walter Kintsch and Teun A. van Dijk. Toward a model of text comprehension and production. *Psychological Review*, 85(5):363–394, 1978. ISSN 1939-1471(Electronic);0033-295X(Print). doi: 10.1037/0033-295X.85.5.363.

[81] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. CrowdForge: Crowdsourcing Complex Work. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 43–52, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0716-1. doi: 10.1145/2047196.2047202. URL http://doi.acm.org/10.1145/2047196.2047202.

[82] Aniket Kittur, Susheel Khamkar, Paul AndrÃ©, and Robert Kraut. CrowdWeaver: Visually Managing Complex Crowd Work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1033–1036, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1086-4. doi: 10.1145/2145204.2145357. URL http://doi.acm.org/10.1145/2145204.2145357.

[83] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1301–1318, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1331-5. doi: 10.1145/2441776.2441923. URL http://doi.acm.org/10.1145/2441776.2441923.

[84] Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. Integrating On-demand Fact-checking with Public Dialogue. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 1188–1199, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2540-0. doi: 10.1145/2531602.2531677. URL http://doi.acm.org/10.1145/2531602.2531677.

[85] Anand P. Kulkarni, Matthew Can, and Bjoern Hartmann. Turkomatic: Automatic Recursive Task and Workflow Design for Mechanical Turk. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 2053–2058, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0268-5. doi: 10.1145/1979742.1979865. URL http://doi.acm.org/10.1145/1979742.1979865.

[86] J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006-341X. doi: 10.2307/2529310. URL http://www.jstor.org/stable/2529310.

[87] Edith Law, Krzysztof Z. Gajos, Andrea Wiggins, Mary L. Gray, and Alex Williams. Crowdsourcing As a Tool for Research: Implications of Uncertainty. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1544–1561, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4335-0. doi: 10.1145/2998181.2998197. URL http://doi.acm.org/10.1145/2998181.2998197.

[88] Doris Jung-Lin Lee, Joanne Lo, Moonhyok Kim, and Eric Paulos. Crowdclass: Designing Classification-Based Citizen Science Learning Modules. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*, September 2016. URL https://aaai.org/ocs/index.php/HCOMP/HCOMP16/paper/view/14027.

[89] Michele Linden and M. C. Wittrock. The Teaching of Reading Comprehension according to the Model of Generative Learning. *Reading Research Quarterly*, 17(1):44–57, 1981. ISSN 0034-0553. doi: 10.2307/747248. URL http://www.jstor.org.ezproxy.lib.vt.edu/stable/747248.

[90] Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. Exploring Iterative and Parallel Human Computation Processes. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 68–76, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0222-7. doi: 10.1145/1837885.1837907. URL http://doi.acm.org/10.1145/1837885.1837907.

[91] Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. TurKit: Human Computation Algorithms on Mechanical Turk. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 57–66, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0271-5. doi: 10.1145/1866029.1866040. URL http://doi.acm.org/10.1145/1866029.1866040.

[92] Shuhua Monica Liu and Jiun-Hung Chen. A multi-label classification based approach for sentiment classification. *Expert Systems with Applications*, 42(3):1083–1093, February 2015. ISSN 0957-4174. doi: 10.1016/j.eswa.2014.08.036. URL http://www.sciencedirect.com/science/article/pii/S0957417414005181.

[93] Kurt Luther, Nathan Hahn, Steven P. Dow, and Aniket Kittur. Crowdlines: Supporting Synthesis of Diverse Information Sources through Crowdsourced Outlines. In *Third*

AAAI Conference on Human Computation and Crowdsourcing, September 2015. URL http://www.aaai.org/ocs/index.php/HCOMP/HCOMP15/paper/view/11603.

[94] Nikki Mandell. Thinking like a Historian: A Framework for Teaching and Learning. *OAH Magazine of History*, 22(2):55–59, April 2008. ISSN 0882-228X,. doi: 10.1093/maghis/22.2.55. URL http://maghis.oxfordjournals.org/content/22/2/55.

[95] Ville J. E. Manninen. Sourcing practices in online journalism: an ethnographic study of the formation of trust in and the use of journalistic sources. *Journal of Media Practice*, 18(2-3):212–228, September 2017. ISSN 1468-2753. doi: 10.1080/14682753.2017.1375252. URL https://doi.org/10.1080/14682753.2017.1375252.

[96] Christina Manzo, Geoff Kaufman, Sukdith Punjasthitkul, and Mary Flanagan. "By the People, For the People": Assessing the Value of Crowdsourced, User-Generated Metadata. 9(1), 2015. URL http://www.digitalhumanities.org/dhq/vol/9/1/000204/000204.html.

[97] Winter Mason and Siddharth Suri. Conducting behavioral research on Amazonâ€™s Mechanical Turk. *Behavior Research Methods*, 44(1):1–23, March 2012. ISSN 1554-3528. doi: 10.3758/s13428-011-0124-6. URL https://doi.org/10.3758/s13428-011-0124-6.

[98] Tyler McDonnell, Matthew Lease, Tamer Elsayad, and Mucahid Kutlu. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2016. URL https://www.ischool.utexas.edu/~ml/papers/mcdonnell-hcomp16.pdf.

[99] Andrea L. McNeill, Peter E. Doolittle, and David Hicks. The effects of training, modality, and redundancy on the development of a historical inquiry strategy in a

multimedia learning environment. *Journal of Interactive Online Learning*, 8(3):255–269, 2009.

[100] William McNeill. Why Study History? (1985), 1985. URL https://www.historians.org/about-aha-and-membership/aha-history-and-archives/archives/why-study-history-(1985).

[101] Douglas C. Merrill, Brian J. Reiser, Michael Ranney, and J. Gregory Trafton. Effective Tutoring Techniques: A Comparison of Human Tutors and Intelligent Tutoring Systems. *Journal of the Learning Sciences*, 2(3):277–305, July 1992. ISSN 1050-8406. doi: 10.1207/s15327809jls0203_2. URL http://dx.doi.org/10.1207/s15327809jls0203_2.

[102] Piotr Mitros. Learnersourcing of Complex Assessments. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, L@S '15, pages 317–320, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3411-2. doi: 10.1145/2724660.2728683. URL http://doi.acm.org/10.1145/2724660.2728683.

[103] Nathaniel Deines, Melissa Gill, Matthew Lincoln, and Marissa Clifford. Six Lessons Learned from Our First Crowdsourcing Project in the Digital Humanities, February 2018. URL http://blogs.getty.edu/iris/six-lessons-learned-from-our-first-crowdsourcing-project-in-the-digital-humanities

[104] Kristen Nawrotzki, editor. *Writing History in the Digital Age*. University of Michigan Press, 2013. ISBN 978-0-472-07206-4 978-0-472-02991-4. URL http://hdl.handle.net/2027/spo.12230987.0001.001.

[105] Sherrie L. Nist and Mark C. Hogrebe. The Role of Underlining and Annotating in Remembering Textual Information. *Reading Research and Instruction*, 27(1):12–25,

September 1987. ISSN 0886-0246. doi: 10.1080/19388078709557922. URL http://dx.doi.org/10.1080/19388078709557922.

[106] NYPL/Zooniverse. Measuring the ANZACs. URL https://www.measuringtheanzacs.org/#/.

[107] Maureen Ogle. The Perils and Pleasures of Going "Popular"; or My Life as a Loser. *Historically Speaking*, 8(4):29–31, 2007. ISSN 1944-6438. doi: 10.1353/hsp.2007.0027. URL http://muse.jhu.edu/content/crossref/journals/historically_speaking/v008/8.4.ogle.html.

[108] Johan Oomen and Lora Aroyo. Crowdsourcing in the Cultural Heritage Domain: Opportunities and Challenges. In *Proceedings of the 5th International Conference on Communities and Technologies*, C&T '11, pages 138–149, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0824-3. doi: 10.1145/2103354.2103373. URL http://doi.acm.org/10.1145/2103354.2103373.

[109] Vineet Pandey, Amnon Amir, Justine Debelius, Embriette R. Hyde, Tomasz Kosciolek, Rob Knight, and Scott Klemmer. Gut Instinct: Creating Scientific Theories with Online Learners. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 6825–6836, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3025769. URL http://doi.acm.org/10.1145/3025453.3025769.

[110] Sarah E. Peterson. The cognitive functions of underlining as a study technique. *Reading Research and Instruction*, 31(2):49–56, December 1991. ISSN 0886-0246. doi: 10.1080/19388079209558078. URL http://dx.doi.org/10.1080/19388079209558078.

[111] Paul Quigley, Kurt Luther, David Hicks, Daniel Newcomb, and Nai-Ching Wang. New directions for inquiry: Citizen student archivists crowdsourcing the past. In

96th Annual Conference of the National Council for the Social Studies (NCSS 2016), Washington, D.C., USA, 2016.

[112] Mia Ridge. From Tagging to Theorizing: Deepening Engagement with Cultural Heritage through Crowdsourcing. *Curator: The Museum Journal*, 56(4):435–450, October 2013. ISSN 00113069. doi: 10.1111/cura.12046. URL http://doi.wiley.com/10.1111/cura.12046.

[113] Mia Ridge. *Crowdsourcing our cultural heritage*. Digital research in the arts and humanities; Digital research in the arts and humanities. Ashgate, Farnham, Surrey, England, 2014. ISBN 978-1-4724-1022-1.

[114] Jennifer Rutner and Roger Schonfeld. Supporting the Changing Research Practices of Historians. Technical report, Ithaka S+R, New York, December 2012. URL http://sr.ithaka.org/?p=22532.

[115] Gerard Salton. *Introduction to Modern Information Retrieval*. Mcgraw-Hill College, New York, September 1983. ISBN 978-0-07-054484-0.

[116] John W. Saye and Thomas Brush. Scaffolding critical reasoning about history and social issues in multimedia-supported learning environments. *Educational Technology Research and Development*, 50(3):77–96, September 2002. ISSN 1042-1629, 1556-6501. doi: 10.1007/BF02505026. URL https://link.springer.com/article/10.1007/BF02505026.

[117] Thomas Schnell and Daniel Rocchio. A Comparison of Underlying Strategies for Improving Reading Comprehension and Retention. *Reading Horizons*, 18(2), January 1978. URL http://scholarworks.wmich.edu/reading_horizons/vol18/iss2/4.

[118] Fabrizio Sebastiani. Machine Learning in Automated Text Categorization. *ACM Comput. Surv.*, 34(1):1–47, March 2002. ISSN 0360-0300. doi: 10.1145/505282.505283. URL http://doi.acm.org/10.1145/505282.505283.

[119] Ivor Shapiro, Colette Brin, Isabelle Bédard-Brûlé, and Kasia Mychajlowycz. Verification as a Strategic Ritual. *Journalism Practice*, 7(6):657–673, December 2013. ISSN 1751-2786. doi: 10.1080/17512786.2013.765638. URL https://doi.org/10.1080/17512786.2013.765638.

[120] Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 614–622, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401965. URL http://doi.acm.org/10.1145/1401890.1401965.

[121] S. Andrew Sheppard, Andrea Wiggins, and Loren Terveen. Capturing Quality: Retaining Provenance for Curated Volunteer Monitoring Data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 1234–1245, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2540-0. doi: 10.1145/2531602.2531689. URL http://doi.acm.org/10.1145/2531602.2531689.

[122] Jakub Šimko, Marián Šimko, Mária Bieliková, Jakub Ševcech, and Roman Burger. Classsourcing: Crowd-Based Validation of Question-Answer Learning Objects. In *International Conference on Computational Collective Intelligence*, pages 62–71. Springer, 2013. URL http://link.springer.com/chapter/10.1007/978-3-642-40495-5_7.

[123] K. L. Smart and J. L. Bruning. An examination of the practical importance of the von Restorff effect. In *annual meeting of the American Psychological Association, Montreal, Canada*, 1973.

[124] R. Smith. An Overview of the Tesseract OCR Engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633, September 2007. doi: 10.1109/ICDAR.2007.4376991.

[125] Ray Smith, Daria Antonova, and Dar-Shyang Lee. Adapting the Tesseract Open Source OCR Engine for Multilingual OCR. In *Proceedings of the International Workshop on Multilingual OCR*, MOCR '09, pages 1:1–1:8, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-698-4. doi: 10.1145/1577802.1577804. URL http://doi.acm.org/10.1145/1577802.1577804.

[126] Raymond W. Smith. Hybrid Page Layout Analysis via Tab-Stop Detection. In *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, ICDAR '09, pages 241–245, Washington, DC, USA, 2009. IEEE Computer Society. ISBN 978-0-7695-3725-2. doi: 10.1109/ICDAR.2009.257. URL http://dx.doi.org/10.1109/ICDAR.2009.257.

[127] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and Fastâ€"but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1613715.1613751.

[128] Peter Stearns. Why Study History? (1998), 1998. URL https://www.

historians.org/about-aha-and-membership/aha-history-and-archives/archives/why-study-history-(1998).

[129] Peter N. Stearns, Peter C. Seixas, and Sam Wineburg. *Knowing, teaching, and learning history: National and international perspectives*. NYU Press, 2000.

[130] Bill Tally and Lauren B. Goldenberg. Fostering historical thinking with digitized primary sources. *Journal of Research on Technology in Education*, 38(1):1–21, 2005. URL http://www.tandfonline.com/doi/abs/10.1080/15391523.2005.10782447.

[131] Simon Tanner, Trevor MuÃ±oz, and Pich Hemy Ros. Measuring mass text digitization quality and usefulness. *D-lib Magazine*, 15(7/8):1082–9873, 2009.

[132] University College London-Gower Street-London-WC1E 6BT Tel: +4420 7679 2000. UCL Transcribe Bentham. URL http://blogs.ucl.ac.uk/transcribe-bentham/.

[133] Peter Tolmie, Rob Procter, David William Randall, Mark Rouncefield, Christian Burger, Geraldine Wong Sak Hoi, Arkaitz Zubiaga, and Maria Liakata. Supporting the Use of User Generated Content in Journalistic Practice. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 3632–3644, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3025892. URL http://doi.acm.org/10.1145/3025453.3025892.

[134] J. Trant. Tagging, Folksonomy and Art Museums: Early Experiments and Ongoing Research. *Journal of Digital Information*, 10(1), January 2009. ISSN 1368-7506. URL https://journals.tdl.org/jodi/index.php/jodi/article/view/270.

[135] Rajasekar Venkatesan, Meng Joo Er, Mihika Dave, Mahardhika Pratama, and Shiqian Wu. A novel online multi-label classifier for high-speed streaming data applications.

Evolving Systems, pages 1–13, 2016. URL http://link.springer.com/article/10.1007/s12530-016-9162-8.

[136] Nai-Ching Wang. Crowdnection: Connecting High-level Concepts with Historical Documents via Crowdsourcing. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, pages 146–151, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4082-3. doi: 10.1145/2851581.2890377. URL http://doi.acm.org/10.1145/2851581.2890377.

[137] Nai-Ching Wang, David Hicks, Paul Quigley, and Kurt Luther. A crowdsourced approach to evaluating the relevance of digitized primary sources for historians. *Human Computation*, to appear.

[138] Nai-Ching Wang, David Cline, David Hicks, Kurt Luther, Kelly McPherson, Craig Perrier, and Paul Quigley. The design, development and implementation of funded transdisciplinary digital history projects: Illustrative cases of k-16 collaboration in action. In *132nd Annual Meeting of the American Historical Association (AHA 2018)*, Washington, D.C., USA, 2018.

[139] Nai-Ching Wang, David Hicks, and Kurt Luther. Exploring Trade-Offs Between Learning and Productivity in Crowdsourced History. *Proc. ACM Hum.-Comput. Interact.*, 2 (CSCW):178:1–178:24, November 2018. ISSN 2573-0142. doi: 10.1145/3274447. URL http://doi.acm.org/10.1145/3274447.

[140] Sarah Weir, Juho Kim, Krzysztof Z. Gajos, and Robert C. Miller. Learnersourcing Subgoal Labels for How-to Videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 405–416, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-2922-4. doi: 10.1145/2675133.2675219. URL http://doi.acm.org/10.1145/2675133.2675219.

[141] Andrea Wiggins and Yurong He. Community-based Data Validation Practices in Citizen Science. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 1548–1559, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3592-8. doi: 10.1145/2818048.2820063. URL http://doi.acm.org/10.1145/2818048.2820063.

[142] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z. Gajos, Walter S. Lasecki, and Neil Heffernan. AXIS: Generating Explanations at Scale with Learnersourcing and Machine Learning. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, L@S '16, pages 379–388, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3726-7. doi: 10.1145/2876034.2876042. URL http://doi.acm.org/10.1145/2876034.2876042.

[143] Sam Wineburg. Historical thinking and other unnatural acts. *The Phi Delta Kappan*, 80(7):488–499, 1999. URL http://www.jstor.org/stable/20439490.

[144] Sam Wineburg. Thinking like a historian. *Teaching with primary sources quarterly*, 3(1):2–4, 2010. URL https://www.weteachnyc.org/media/filer_public/24/4a/244ab1eb-4540-4ca1-a60e-5ef96326b365/research_history.pdf.

[145] Sam Wineburg. Thinking Like a Historian. *Library of Congress*, 2010. URL http://www.loc.gov/teachers/tps/quarterly/historical_thinking/article.html.

[146] Samuel S. Wineburg. On the Reading of Historical Texts: Notes on the Breach Between School and Academy. *American Educational Research Journal*, 28(3):495–519, September 1991. ISSN 0002-8312, 1935-1011. doi: 10.3102/00028312028003495. URL http://aer.sagepub.com/content/28/3/495.

[147] Samuel S. Wineburg and Suzanne M. Wilson. Subject matter knowledge in the teaching of history. *Advances in research on teaching*, 2:305–347, 1991.

[148] M. C. Wittrock. Learning as a generative process. *Educational Psychologist*, 11(2): 87–95, November 1974. ISSN 0046-1520. doi: 10.1080/00461527409529129. URL http://dx.doi.org/10.1080/00461527409529129.

[149] M. C. Wittrock and Kathryn Alesandrini. Generation of Summaries and Analogies and Analytic and Holistic Abilities. *American Educational Research Journal*, 27(3):489–502, September 1990. ISSN 0002-8312, 1935-1011. doi: 10.3102/00028312027003489. URL http://aer.sagepub.com/content/27/3/489.

[150] Merlin C. Wittrock. Generative Processes of Comprehension. *Educational Psychologist*, 24(4):345–376, September 1989. ISSN 0046-1520. doi: 10.1207/s15326985ep2404_2. URL http://dx.doi.org/10.1207/s15326985ep2404_2.

[151] Anbang Xu, Huaming Rao, Steven P. Dow, and Brian P. Bailey. A Classroom Study of Using Crowd Feedback in the Iterative Design Process. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1637–1648, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-2922-4. doi: 10.1145/2675133.2675140. URL http://doi.acm.org/10.1145/2675133.2675140.

[152] Lixiu Yu, Aniket Kittur, and Robert E. Kraut. Distributed Analogical Idea Generation: Inventing with Crowds. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 1245–1254, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2557371. URL http://doi.acm.org/10.1145/2556288.2557371.

[153] Carole L. Yue, Benjamin C. Storm, Nate Kornell, and Elizabeth Ligon Bjork. Highlighting and Its Relation to Distributed Study and Studentsâ€™ Metacognitive Beliefs. *Educational Psychology Review*, 27(1):69–78, July 2014. ISSN 1040-726X, 1573-

336X. doi: 10.1007/s10648-014-9277-z. URL http://link.springer.com/article/10.1007/s10648-014-9277-z.

[154] Min-Ling Zhang and Zhi-Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, July 2007. ISSN 0031-3203. doi: 10.1016/j.patcog.2006.12.019. URL http://www.sciencedirect.com/science/article/pii/S0031320307000027.

[155] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2014. URL http://ieeexplore.ieee.org/abstract/document/6471714/.

[156] Haiyi Zhu, Steven P. Dow, Robert E. Kraut, and Aniket Kittur. Reviewing Versus Doing: Learning and Performance in Crowd Assessment. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 1445–1455, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2540-0. doi: 10.1145/2531602.2531718. URL http://doi.acm.org/10.1145/2531602.2531718.

[157] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. Research Through Design As a Method for Interaction Design Research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 493–502, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-593-9. doi: 10.1145/1240624.1240704. URL http://doi.acm.org/10.1145/1240624.1240704.

Appendices

Appendix A

SCIM Scoring Rubric (Based on [70])

Summarizing (1 point each) (1) Does the response indicate the subject of the source? (2) Does the response indicate the audience for the source? (3) Does the response indicate the author of the source? (4) Does the response include specific details from the source? Contextualizing (1 point each) (1) Does the response indicate when the source was produced? (2) Does the response indicate where the source was produced? (3) Does the response indicate why the source was produced? (4) Does the response indicate the immediate or broader context? Inferring/Monitoring (1 point each) (1) Does the response include explicit and/or implicit inferences? (2) Does the response include inferences based on omissions? (3) Does the response indicate the need for information beyond the source? (4) Does the response evaluate the usefulness or significance of the source?

Appendix B

Summary Scoring Rubric

1. The worker wrote an original summary. The worker did not plagiarize.
2. The worker wrote a summary that contains all important information in the document.
3. The worker wrote a summary that has balanced coverage. The summary does not focus on some specific parts of the original document.
4. The worker wrote a summary without adding personal opinions nor emotions.
5. The worker wrote a summary with sufficient information and details from the original document.
6. The worker did not have spelling and grammar mistakes.
7. The worker wrote the right amount (3 sentences).

Appendix C

SCIM Questions Used In CrowdSCIM (Based on [70])

Summary-tone (Summarize)

1. What type of historical document is the source?
2. What specific information, details and/or perspectives does the source provide?
3. What is the subject and/or purpose of the source?
4. Who was the author and/or audience of the source?

Tag (Contextualize)

1. When and where was the source produced?
2. Why was the source produced?
3. What was happening within the immediate and broader context at the time the source was produced?
4. What summarizing information can place the source in time and place?

Connect (Infer/Monitor)

1. What interpretations, inferences, perspectives or points of view may be drawn from or indicated by the source?

2. What inferences may be drawn from absences or omissions in the source?

3. What additional evidence beyond the source is necessary to answer the historical question? useful or significant is the source for its intended purpose in answering the historical question?

Appendix D

Sample Participant Responses

D.1 Sample historical interpretations of CrowdSCIM across pre-test and three crowdsourced tasks

D.1.1 Pre-test

"The farming communities are having a real tough time. Most of the crops have died out over the last season, due to weather related catastrophes. Merchants have began cutting off credit, so farmers are really going to start feeling the pinch. A family member also seems to have gotten a good job as a personal chauffeur."

— P287 with 1 point for describing detail in Summarize

D.1.2 Summary-tone

"Estella writes to her sister about life on the farm in 1911. She relates details about the tough weather and pest conditions that led to a worrisome set of circumstances for all the farmers. The situation was dire enough that merchants were cutting off credit to farm families. She was also hopeful, though, that the current rains might change the conditions in their favor so that they might have a good crop of wheat. She also writes of her daughter who is headed

for a job in NYC and another mutual acquaintance who was changing his job from chauffeur to pilot."

— P310 with 3 points in Summarize and 1 point for time in Contextualize

D.1.3 Tag

"This letter from Estella Stigebower to her sister Ella Roesch, written on August 3rd, 1911, indicates conditions on the plains were very harsh during the early 20th century. Writing from Marion, Nebraska, Stigebower responds to her sister's previous letter by indicating she is sorry to hear the crops were a total failure. Her letter also demonstrates that crop failure was experienced by many people in the plains, largely as a result of storms and plagues of grasshoppers and army worms. Crops had failed to such an extent that merchants were forced to cut off credit to farmers. Finally, Stigebower offers a glimpse of how other family members outside of the plains have been fairing, indicating they have been able to get jobs and not mentioning any specific hardship they have had."

— P332 with full points in Summarize and Contextualize

D.1.4 Connect

"Conditions of life in farming communities during this time period were difficult. Crops were failing not only for this family, but for many families- suggesting that many families were unable to support themselves properly. The lack of help from the town (cutting the credit system) also suggests that the wealthier families had no desire to help the other citizens. Although this point of view is missing, it seems that like today, the wealthier businesses only wanted to help themselves and not those who needed assistance. If the article included

information about other businesses, it would add to the perspective of ”the other side”.”

— P359 with 1 point for details in Summarize and 2 points for inference and monitoring in Infer/Monitor

D.2 An example of an improved summary using Crowd-SCIM

D.2.1 Original Summary

“Estella has written a letter to her sister and family, catching them up on what’s going on in her life. She’s concerned about her family’s failed crops and has seen similar issues in her area”

— P312 with score 4/10

D.2.2 Revised Summary

“In the Midwest, pre-WWI, Estella has written a letter to her sister and family, catching them up on what’s going on in her life. She’s concerned about her family’s failed crops and has seen similar issues in her area where corn fields were destroyed by grasshoppers and hail storms. She also talks about other people, including Lillie and Gus, who was working as a chauffeur for $125/month and hoping to run an airship soon.”

— P312 revised work with score 9/10 for extra context (location and time), details (grasshoppers and hail storms), and coverage (relatives)