Foundations of the Written Rummage:

Crowdsourcing, Transcription and Prospects for the Information Age

Dr. Lang

Senior Paper: Mathematics

MAT-499

April 26, 2011

By Joshua Collins Rio-Ross

Chapter 1: Background and Reasoning of the Written Rummage Project

The developing "information age" is continually unraveling into new ways of discovering, presenting and sharing information. Most new academic material is digitally formatted upon its development: our current academic culture has evolved to a point of digital word processing programs being normative for writing, duplicating and sharing material. However, there remains an ocean of material from times prior to the "information age" that has yet to be converted to digital form. Much of this material can be found in library collections—whether academic, public or private—and thus remains available only to a limited number of locals or willing-and-able sojourners. As time fares onward, those who own collections of handwritten documents are increasingly wanting make the content thereof available to the general public. In fact, the public is continually coming to expect those holding yet unshared information to make that information accessible to them by virtue of its existence as information. "Because it is, it must be able to be known," is the zeitgeist.

Modes of converting this information are not yet "down to a science." While optical character recognition (OCR) devices are useful, they cannot descry all the blemishes and stylistic curls, swirls, and serifs of human handwriting. Some programs and online services have been developed to mitigate this problem, the most prominent of which are Captcha and ReCaptcha. However, even with these tools, the need for human skill still exists, and transcription is often more costly than collection owners are willing to pay.

Dr. Andrew Lang and I undertook developing a way to transcribe handwritten documents as a service to anyone wanting to transcribe a collection of handwritten or

otherwise OCR-incompatible manuscripts into digital, searchable format. We decided to name this project "Written Rummage."

Crowdsourcing: A Brief Introduction to Crowdsourcing and Other Projects

In the 2007 Hollywood release *Live Free or Die Hard*, a haggard New York hacker, Matthew Farrell, finds his life jeopardized after he thinks he earned fifty thousand dollars. He was the successful hacker in a national competition wherein computer users were asked to submit code cracking the security system of a given company, supposedly in an effort to help secure that company's security. Little did they know that they were participating in a terrorist plot to breach FBI's database. If nothing else, this plot demonstrates the power of what has come to be known as "crowdsourcing." Albeit legally, Dr. Lang and I decided we wanted to tap into the utility of crowdsourcing for the sake of providing a service to libraries and other owners of original, handwritten manuscripts.

Crowdsourcing is the process by which a company outsources a function to an undefined network of people rather than hiring one or several professionals to accomplish the same function. The company chooses the "winning" method of performing the function, compensates the successful user the predetermined (generally cheap) reward, and keeps the rights to the work and method. This process can occur by users cooperating, operating individually, or many individuals completing individual tasks that amount to a cohesive whole. It provides the company a solution to its function without having to pay higher wages, while the users benefit with a quick job and compensation, which is especially profitable for users in foreign countries where they have a greater value in the American dollar than their monetary system. Furthermore, the company has

a broader pool of amateur and potentially expert talent to select from and pays strictly when it is satisfied with the product.

Crowdsourcing by nature entails a lack of accountability. In not hiring a professional or working with a specific enclave of people, the company is less able to hold the employed accountable to smooth progress or finishing by a certain deadline. There are weaker forms of accountability, such as predetermined dates for completion, just as there are predetermined expectations for quality, but there is not the level of commitment that comes standard with professional employment with contracts. The worker is viable to either begin a task and not complete it or not attain to the standard of the company. Thus there is no assurance that anyone who undertakes the crowdsourced task will produce the quality of work the company prefers. Because of this risk, the time—and money—taken to inspect the quality and accuracy of the work, especially on large-scale projects, could potentially result in crowdsourcing being less profitable than hiring another more controlled business model. We will attempt to circumvent this discrepancy by not only crowdsourcing the original manuscript but also the proofreading process. If the submitted files match in size, the transcription will be accepted; if not, then we will crowdsource again until the transcription and proofread text match.

Description of Our Project

The Written Rummage project is therefore set to research crowdsourcing and the various ways it has been implemented online. We will then use Mechanical Turk to construct a model for a non-profit business that will transcribe documents for customers. Our sample collection will be a diary from the Frederick Douglass Papers owned by the Library of Congress (titled “Frederick Douglass Diary (Tour of Europe and Africa) on

the website). Once the manuscripts are transcribed, the collected data will then be used to determine the viability of the non-profit business after having analyzed the successes, failures, problems and possible modifications of the project.

For clarity, we will use the anonymous, online workforce of Mechanical Turk to transcribe documents and, in the second order, proofread those transcriptions for accuracy. Proofreading could take more than one round for a given document; nonetheless, for this experiment only one round will likely be performed on each of the transcribed documents. Compensation to workers will be greater for proofreading than original transcription since the proofreaders will read both the original manuscript and the transcription. We will begin compensation for the first transcription at around $0.02-0.03, and around $0.05 for the proofreading. In an effort to ensure that the proofreaders actually do work, we will interpolate mistakes in the text and tell the proofreaders that mistakes already exist and that they are to be fixed if compensation is to be given.

Variables of interest include average overall cost per manuscript once transcribed to desired level, as well as how cost affects both quality and expediency of a given manuscript's transcription. With these, we will then have a better idea of what sort of pricing and time frame we would be able to offer collection owners should they request our service.

Chapter 2: On Crowsourcing; or, A New Way to Surf a Crowd

Crowdsourcing is the process by which a company outsources a function to an undefined network of people rather than hiring one or several professionals to accomplish that function. Usually through the use of the internet, the company attempts to gather multiple contributors to a given project in order to enhance the quality for their pay or in order to acquire as much information as possible for a particular purpose. Crowdsourcing is therefore especially useful when a great amount of diversity is sought for a project, as well as when a particular sort of task needs to be repeated numerous times, such as when taking polls or gathering data.

This may or may not be an explicit "business transaction," though online services such as Mechanical Turk have contributed toward developing a paid internet workforce. Other crowdsourcing projects, such as Galaxy Zoo and the Open Dinosaur Project, have utilized a free workforce gained through common interest. In either capacity, crowdsourcing has become a platform for innovation and discovery because of the size of the workforce possible.

Examples of Academic Crowdsourcing

Galaxy Zoo

Galaxy Zoo and the Open Dinosaur Project are two of the most salient examples of how crowdsourcing is used to collect data *en masse*. Galaxy Zoo opened its galleries of telescope photos of galaxies to the general public in order to have any willing contributors provide simple classifying information about each galaxy pictured. Some of the information requested were the color, shape and, if applicable, spin direction of the galaxies. In the first round of submissions, Galaxy Zoo received help from over 150,000

users. To verify the information, Galaxy Zoo would send each galaxy picture once to two different users, comparing the first and second submissions of information. If the latter matched the former, then the submissions were accepted; if not, then the image was resubmitted until a continuity of information was achieved. Surprisingly, there arose little frustration of user integrity.

The data collected proved to be extremely useful to the inquiring astronomer minds behind the project. So much so, in fact, that they opted to use the same methodology (of crowdsourcing) to acquire additional, albeit more particular, data about the galaxies. The data acquired in the first set of submissions corrected some weak withstanding theories while at once confirming, through this dense inductive medium, some other commonly held theories about the flow and movement of the universe. One such instance was the confirmation that galaxies in close proximities have the same movement, indicating a continuity of movement from a given point or at least the sharing of some impetus. The next mass data collection is organized to hone in on some other key, pressing questions the astronomers can only answer with a gross amount of empirical data. For further information, visit www.galaxyzoo.com.

Open Dinosaur Project

Another academic crowdsourcing project is The Open Dinosaur Project (ODP), founded by a crew of paleontologists who wanted to make a project that both scientists and the public could openly contribute to "developing a comprehensive database of dinosaur limb bone measurements, to investigate questions of dinosaur function and evolution." No prior education, in paleontology or otherwise, is required for participation. Rather, through a series of checks and balances within the data collection

process, the ODP is able to hold submissions accountable.

The ODP process is fairly straightforward, yet, being as though they are nearing the conclusion of their research and have already begun analysis and paper writing, the process is evidently effective. The first "Investigator," after being supplied necessary materials through Gmail or otherwise, then takes measurements of a given "Specimen" and records those measurements in millimeters on a data sheet. Once the data sheet is complete, the Investigator emails it to the Project Lead along with the necessary bibliographic references. From here the data is added to the "verification list," a collection of data entries that are then sent on to be confirmed by different Investigators than the first. If they submit similar data to that of the first Investigator, then the data is accepted and added to the database.

As aforementioned, the ODP is now analyzing data and preparing various papers reporting the data they have collected. Part of this process entails the drudgery of verifying all bibliographic references of those whom have contributed to the data collection and paper writing. (http://opendino.wordpress.com)

Open Notebook Science Solubility Challenge

While the above projects require little or no prior knowledge or experience of the person completing the crowdsourcing task, others are more specified to a given field. The Open Notebook Science Solubility Challenge "calls upon people with access to materials and equipment to measure the solubility of compounds," with aldehydes, amines and carboxylic acids as a priority. The project is also a competition that gives awards to its most substantial and noteworthy contributors.

Keeping with the theme of crowdsourcing, the project is called "Open Notebook"

because it is open to anyone with the resources and knowledge to contribute to the database. The credibility of the participants' work is held accountable by the format of submission: raw data is not all that one submits, but also the methodology they use in acquiring that data. Thus the Open Notebook community is one of scientists holding one another accountable to each other's process and findings. Sometimes this simply means scientists posting the data from or description of failed experiments, called "dark data." Their website explains:

> Understanding exactly how an experiment was performed is essential to the efficient progress of science. There are no absolute facts in the scientific literature; every measurement reported is only meaningful within the full context of how it was generated. The purpose of a laboratory notebook is to report as much of this context as is reasonable. But to find trends data must be abstracted to a level where they can be manipulated in tables and charts. This is not a problem as long as one can drill down from each data point in a chart to the full context found in the laboratory notebook.

Eventually, the "ONSC" developers intend to publish a paper—or a collection of papers—in a peer-reviewed journal, listing all contributors as co-authors. The project fits into the context of a larger research expedition to "synthesize new anti-malarial agents using the Ugi reaction." (http://onschallenge.wikispaces.com/)

Games with a Purpose

Some crowdsourcing instances use the crowd while the crowd is unaware it is being sourced. For example, there are several games that people play for fun that have an added crowdsourcing component, such as Foldit and The Spectral Game.

Foldit

Foldit is a game developed from a conglomeration of the University of Washington's Computer Science & Engineering and Biochemistry departments. Users are provided a number of tools to solve the sophisticated puzzle of the proper or most useful structure of a given protein. As users return solutions to the various protein puzzles, they contribute to chemical research being done to synthesize antidotes to HIV, Cancer, Alzheimer's and otherwise. The point system is in place to spur users on to competitively provide the most quality proteins based upon a complicated algorithm having to do with proper protein structure, some of which is explained on the game's website (http://fold.it/portal/). The game designers are hoping that in relying upon human puzzle/problem solving intuition to make proteins, they can later teach computers to make proteins in imitating the processes humans use. For now, until such comprehensive data is obtained, crowdsourcing is providing a wider, competitive base to do research that would otherwise be too cumbersome and/or time-consuming for any one team of scientists to accomplish.

The Spectral Game

The Spectral Game is a web-based game where players try to match molecules to various forms of interactive spectra including 1D/2D NMR, Mass Spectrometry and Infrared spectra. Each correct selection earns the player one point and play continues until the player supplies an incorrect answer. The game is usually played using a web browser interface, although a version has been developed in the virtual 3D environment of Second Life. Spectra uploaded as Open Data to ChemSpider in JCAMP-DX format are used for the problem sets together with structures extracted from the website. The spectra

are displayed using JSpecView and ChemDoodle Web Components, Open Source spectra viewing platforms which afford zooming and integration. As people play the game, they have an option to flag and comments on spectra. The flags and comments are then reviewed by the ChemSpider team to curate their spectral data.

Many other instances of crowdsources are extant online and are listed on Wikipedia—itself another case of crowdsourcing's increasing breadth and influence.

Crowdsourcing Digitization Efforts

Google Books

By now, most of the world has happened upon Google Books. Anytime someone searches Google for a book title or even a phrase found in a book, Google automatically runs the title or phrase through Google Books to find a match and presents that match near the top of the results. Users can then usually view snippets of whatever work resulted with matches to the phrase. These snippets tend to go for a series of several pages before the rest of the work is blocked to the user. Along the screen's side one will also find an arsenal of links to websites where one can purchase the work the user is viewing.

Google Books has not been without controversy. Providing snippets of thousands of books means also uploading thousands of books into a database, then making those available to users with Google as the resource. The Author's Guild sued Google in 2005, saying that their mass-digitizing project committed copyright infringement. After settling (for $125 million) Google undertook restructuring its project in a way that appeases authors and publishers, while still allowing them to follow through with

digitizing. One can visit http://books.google.com/googlebooks/agreement/ to see the details of this new structure.

Google Books has released The Library Project, allowing all participating libraries to include their books in Book Search. For Library Books still in copyright, Google will offer small snippets of the book, its basic information and where it can be purchased. According to the Google Books website, The Library Project has also partnered with over 20,000 publishers and authors to make their books discoverable on Google. The goal of the project is best explained by the site:

> The Library Project's aim is simple: make it easier for people to find relevant books – specifically, books they wouldn't find any other way such as those that are out of print – while carefully respecting authors' and publishers' copyrights. Our ultimate goal is to work with publishers and libraries to create a comprehensive, searchable, virtual card catalog of all books in all languages that helps users discover new books and publishers discover new readers.

Since Google is establishing a coalition of libraries, of sorts, it is also offering library-type services to its users. Though the agreement has not been fully approved, Google intends to offer Google users the opportunity to purchase full online access to millions of books by logging onto their Book Search account and accessing their "electronic bookshelf." Libraries and universities will likewise be able to purchase institutional subscriptions for their respective members and/or students. Finally, users will always be directed to where they can purchase a book should they desire to do so.

Access to Google Books will be dependent on which of three types of books it is determined to be: in-copyright, in-print; in-copyright, out-of-print; and out-of-

copyright. Authors of the books still under copyright will be able to determine how much information is actually shared on Google Books when their book is found through the search engine. Out-of-print books can yet again be sold, should the author so prefer, giving an incentive for authors of such books to load them into Google's database. Out-of-copyright books are made available for reading, download and printing.

Google has therefore made great strides at legally digitizing millions of books to make them accessible to the public. They intend to make these searchable and thus useful for a variety of purposes, not the least of which is research. However, no evidence has yet been forthcoming about Google's plans to do any work transcribing or digitizing handwritten manuscripts that are not yet searchable or OCR-compatible.

Historical Documents—Ships Logs

Another smaller imaging and digitizing endeavor is known as the Corral Project. According to their website, http://www.corral.org.uk/Home, the Corral Project's "principal objective is to image ship's logbooks particular historic and scientific value and to digitise the meteorological observations in those logbooks" [sic]. The logbooks are supplied by the UK National Archives at Kew, Surrey. While the sheer historical significance of the logbooks is enough for some to take interest in the project, the core reason for the digitization is to provide meteorologists with a supply to peek into weather patterns before the twentieth century. This information can then be used to try to determine cycles in the earth's meteorology, thus either to confirm or debunk existing theories about weather patterns.

As aforementioned, the logbooks are supplied by the UK National Archives. From their website, they provide images of pages from the logbooks that users can then

use to transcribe the data contained in those images into Excel spreadsheets and/or text documents. Once the data is verified, the information is sent on for analysis. While meteorological study was not so precise then as it is now—were there as many known variables—most logbooks track barometric pressure and temperature at sea. This information is useful in that it provides that "peek backwards" for patterning and also in that it provides data from earth's oceanic surface rather than just land.

The original Corral Project ran about a year, between October 2008 and September 2009. Since then, larger projects—and likely more grant money—with more logbooks are being sought. The developers consider the project a success and all the more necessary for climatological study amidst rampant discussion about global warming and its various manifestations. Their full paper can be seen on their web site: http://www.corral.org.uk/Home/project-documents.

HathiTrust

HathiTrust is steadily becoming one of the prominent names among digital libraries. In partnering with a number of consortia and individual academic institutions—an exhaustive list of which can be found at hathitrust.com/community—they have begun building a "community of research libraries committed to the long-term curation and availability of the cultural record."

Part of HathiTrust's objective is to develop a research database with enough space to support a community of libraries' materials while still maintaining which materials still properly belong to each respective library. To make this a possibility, they have developed both a "page turner mechanism" to be able to read, download and interact

with texts and images from HathiTrust, as well as a branding mechanism that watermarks all resources a user is interacting with. Beyond simply making resources available to individual researchers, HathiTrust strongly emphasizes libraries sharing resources with one another. Also, in an effort to make HathiTrust's materials available, they have partnered with World Cat and are trying pilot releases to ensure usability. Over time, they are also hoping to support materials other than books and journals.

Whatever HathiTrust provides, they are intent upon the partners' resources being secure and their user's resources being accessible. Once the library is fully launched, users will be able to execute full-text, cross-repository searching to find materials. (www.hathitrust.org/mission_goals)

HathiTrust is not free. Because their service requires security, organization, governance, maintenance and otherwise, they have to seek support from their partners as well as academic grants. For now, they are developing a new cost model that takes into account partners with material to contribute, and supporting partners without digital content to contribute. The site says the new cost model:

> will distribute the costs of sustaining the repository in a way that more accurately reflects the benefits each partner receives from the preserved collections. The new model will also allow institutions to join HathiTrust that wish to participate in the curation and management of the repository in return for specialized services, but do not necessarily have digital content to contribute. (http://www.hathitrust.org/cost)

Specific formulas used to determine cost can be found on the website.

Through 2011, HathiTrust had an estimated 9.5 million volumes shared among 60

partners. As more universities and major libraries grow weary of maintaining or unable to maintain their digital libraries and certain sections of their physical libraries, they will likely seek the governance of an organization like HathiTrust. Digitization is therefore increasingly demanded and transcription efforts are soon to follow.

Mechanical Turk

The System

Mechanical Turk is an online crowdsourcing resource developed by Amazon(.com). Rather than having precisely one crowdsourcing cause, Mechanical Turk was designed to provide a framework for companies or independent users—called "Requesters"—to utilize a crowdsourcing workforce—called "Workers." Obviously the needs of various Requesters differ, meaning that MTurk needed to provide a variety of templates for Requesters to use to meet their needs and present their requested task to Workers in such a way that it is clear and able to be completed. Because MTurk is business-based—if Requesters are pleased with the results of one or more Workers, those Workers are compensated through MTurk acting as a financial intermediary—Requesters specify in their task templates how much compensation a worker can expect if their work is accepted by the Requester. The Worker is permitted to choose which tasks they want to attempt; the Requester is able to choose which Workers have provided adequate work and thus who will be compensated. Sometimes Requesters have only one task to be completed; at other times, Requesters may want the same task completed numerous times, such as The Sheep Market project.

The Sheep Market Project

The Sheep Market was an experiment initiated by Aaron Koblin. Upon hearing

about the release of MTurk's crowdsourcing technology, Koblin thought to explore the philosophical, aesthetic, and socio-economic implications of a company establishing such a workforce. After a few minor visual projects as sort of "test runs" of MTurk, Koblin decided to initiate a major project: The Sheep Market. The Sheep Market called for workers to "draw a sheep facing left" and offered two cent ($0.02) compensation to each submitted image that fulfilled that criteria. (It is worth noting that hundreds of images were rejected because they for one of a limited number of reasons did not meet these criteria). On TheSheepMarket.com, the user can scroll over any one of the first 10,000 accepted submissions—compiled in a sort of mural—to view them; one can also click on that image to view its sheepish drawing process. Koblin eventually released the mural as an art exhibit in several major museums around the world. One can view his published analysis of the project on his website, www.aaronkoblin.com/work/thesheepmarket.

Unlike The Sheep Market, Written Rummage had the same sort of task to be done for a variety of different instances. To be more specific, while the task of handwritten-to-digital transcription is common to all of Written Rummage's requests, the instance of what particular manuscript to be transcribed varies with each request. Using MTurk in with this kind of framework is not unprecedented.

Casting Words

One of the most prominent companies making use of MTurk is Casting Words. Casting Words, like Written Rummage, is an online transcription agency. However, where Written Rummage concerns the transcription of the written word, Casting Words has broken ground in using MTurk's crowdsourcing resources to transcribe audio recordings. Casting Words' clientele submits an audio recording of variant circumstance

and quality, after which Casting Words submits that recording to the MTurk workers for a designated compensation, usually based upon length and audio quality. According to the Casting Words website, the company employs multiple workers to review the same piece of audio to ensure quality control. Once the transcription is determined to be of appropriate quality, it is submitted back to the client, either in plain text, HTML, or MS Word formats, depending upon preference and/or need.

Outro

The contexts and uses of crowdsourcing technology are only beginning to disseminate, diversify, and flourish. It is the hope of Written Rummage to expand the use of this technology to serve educational research. To this, we will expatiate hereafter.

Chapter 3: Finding and Selecting a Sample Collection

Introduction

Before proceeding with developing a model for transcription, it was first necessary to select a sample collection of manuscripts to transcribe. Key to this process was finding a collection that would adequately represent the sort of manuscripts we are likely to be working with when contracted by libraries and other collection owners. We also had to choose a collection that is large enough to provide an adequate pool of data to analyze and use for business projections, while also being small enough that it can viably be transcribed before deadlines of this project demand it come to a close. Tacked onto these, if a collection were chosen that brings with it an institution or collector that could be a potential partner, that would only expedite the latter marketing process.

Holy Spirit Research Center

We began locally. Oral Roberts University's library contains the Holy Spirit Research Center (HSRR), one of the largest collections of Pentecostal and Charismatic documents in the country. Surely, we thought, they would want to digitize their collection, thereby making it more accessible to in-house researchers, distance-learning students and scholars in need of those resources with no sensible way of getting to them or even knowing they exist.

Initially we asked the collection's keeper about what handwritten documents they held that they might be interested in transcribing. He explained that most documents held in the collection and of note in the Pentecostal movement were typically printed and published—numerous Pentecostal journals chronicling testimonies of miracles and

conversions, prayer requests and praise reports were in circulation starting around the Azusa Street Revival in the first decade of the 20th century. Since print was the standard mode of communication in the movement, handwritten letters and treatises are more scarcely found in collections of the HSRR's sort.

An immediate concern arose: is our project germane to the HSRR's needs given that almost all their materials are already in print form? The collection's keeper further explained that the library is actively preparing to digitize their library using new "old-and-fragile-and-barely-bound-manuscript-compatible" Snapter photo technology. With this device, libraries can place their books inside the machine, where the book can lie comfortably between two cameras that will take pictures of both open pages, gingerly turn the page, and rinse and repeat until the whole work is digitized and converted into a PDF. If we could at all help, our work would begin once some of the collection's manuscripts were digitized in this way.

Nonetheless, the library seemed more interested in having someone make their texts searchable than how that process was undertaken, and nonetheless did not have a definite timeframe establishing when they would own Snapter's technology and begin digitizing. Given the OCR's typical use of identifying the familiar sorts of fonts one would find in old Pentecostal journals, our method of transcription would not be the optimal route HSRR would take to make their resources searchable. We opted to seek out partners with a need more particular to the gap we are hoping to fill.

UCLA's Catalogue of Digitized Medieval Manuscripts

During the course of our research Dr. Lang discovered a nascent catalogue being developed by UCLA (UCLA Catalogue of Digitalized Medieval Manuscripts) that digitizes collected medieval—and sometimes older—manuscripts, stores them into their database and posts them online to be freely viewed and even read. Because these documents predate the Printing Press, all of them are handwritten. They are collected from around the world and from a variety of medieval cultures, meaning they are written in a variety of languages, including Latin, Old English, Middle English, German, and French, all of which conveniently share the same alphabet. Digitization of the documents is performed by photographing pages against a black background and uploading them to be viewed as either books or individual specimens, as the case may be. UCLA likely uses some process similar to the one HSRR described, though this can be neither confirmed nor denied from the website.

One could thereby—inconveniently—read through the whole manuscript by clicking on each image individually and zooming in to make it viewable. The pictures have high enough pixels that they do not distort when enhanced; however, the process of reading images of penmanship through a series of pictures is less than desirable. It seems that this mode of viewing is present despite UCLA's overarching intentions with the catalogue. Though they certainly aim to exhibit the documents, the notion of online galleries has yet to catch vogue. UCLA prides itself upon being avant guard in nearly all of their developments, and this catalogue seems little exception. Consequently, we hoped they would take interest in the technology we are developing for the transcription of handwritten manuscripts via crowdsourcing.

One distinct limitation of crowdsourcing is not knowing what the pool of workers knows. UCLA's collection has a distinct multi-cultural appeal in representing various Germanic and Romantic languages through various portions of their history. However, that breadth of representation also means transcription would require an equivalent breadth of representation in workers. Mechanical Turk, our crowdsourcing resource, has allows us to make tasks available to over 500,000 workers, representing more than 190 countries. Though one can expect an eclectic skill set within such a market, we have little guarantee that sufficient users competent in the given language of a collection would be available to transcribe. Further, if they were to describe, my deficiency in these languages would disable me from verifying manuscript quality. An English collection proved preferable.

Library of Congress' The Jefferson Papers

After discovering that at least for immediate purposes UCLA's collection proved undesirable, we decided to browse the internet for other digital collections of handwritten manuscripts. The best candidate is an online collection presented by the Library of Congress called "The Thomas Jefferson Papers." It is purportedly the largest collection of Thomas Jefferson's handwritten manuscripts in the world (over 27,000), including letters and speeches. These have several advantages. Being as though letters and speeches are written to be read, they are predominantly legible. Also, though the speech is dated in some respects, the letters and speeches are written in English, which would provide transcribers with context to aid in transcription. Furthermore, the collection has a cultural relevance—always a major plus in substantiating a project.

The Thomas Jefferson Papers were a major step in finding our full collection. Eventually the collection did not show the sort of consistency desired for a pilot run. Plenty of documents in the collection were written letters and speeches, but many were also assorted receipts, daily plans, acquaintances' addresses and indecipherable jots, notes and phrases. Should the Written Rummage project fully develop, we would not shy away from collections baring such marks; however, for the pilot run a more consistent series of manuscripts with a paragraph format is preferred.

Library of Congress' The Frederick Douglass Papers

Fortunately historical collections can keep significant historical figures like Thomas Jefferson and his American successors as close neighbors. A series of links led us to The Frederick Douglass Papers, a similar collection to the Thomas Jefferson Papers described above, but having in it a series more fitting for our project than any collection yet encountered. Within The Frederick Douglass Papers there is a series called "Diary, 1886-94" (aka: "Frederick Douglass Diary (Tour of Europe and Africa)") bearing the description, "A single diary kept by Douglass during his 1886-87 tour of Europe and Africa, with notes added in later years" (http://memory.loc.gov/ammem/doughtml/doughome.html).

While a few of these documents have similar idiosyncrasies to the Thomas Jefferson Papers, these are hardly significant enough to impede using the collection for Written Rummage. The diary is 72 images, each with their own URL and their own link to a higher quality, "zoomable" image. Given Frederick Douglass' historical significance and reputed knack for rhetoric, this diary is assumed to be rich with enjoyably and

pertinently transcribable handwritten text.

Chapter 4: The Process: Using Mechanical Turk and Google Docs

Being in the nascent stages of Written Rummage's development, we are not yet so privileged as to have the system automated. Instead, the pilot run is for the purpose of providing a proof of concept and then with that proof of concept finding a viable system to automate. That process' finer details are the concern of this chapter.

As mentioned above, The Frederick Douglass Papers, our chosen sample collection, comes complete with a URL for each of its respective manuscript images. We can then use those URLs to develop tasks in Mechanical Turk for workers to perform.

Before anything we had to develop a Mechanical Turk Requester account, providing necessary information like who we are and what means of payment we intend to use to fund the account out of which Mechanical Turk takes a commission and pays its workers. Further, the name of our business had to be established: "Written Rummage."

Immediately after creating the account, Mechanical Turk educates its Requesters about how to effectively use its services and how to build a clear, useful and workable "HIT," or task to be completed by workers. MTurk provides a tripartite sequence for getting work done: design, publish and manage. Naturally, one must begin with the design of the HIT.

Myriad templates are made available by MTurk to implement or modify. Resources are available for large-scale surveys, open-end questions, picture tagging, content filtering and many other simple tasks. For now, surveys and content filtering do not fall within our field of interest, but long strings of character input do, leaving us with a modified "open-end questions" template.

Requesting initial and proofread transcriptions means creating two different

templates. The former will hereafter be called the "Text Image template" or "TI" for expediency. MTurk's first design page, shown in Appendix A, specifies the overarching descriptions and fundamental information of the HIT. Each HIT is given a title, description and a set of keywords by which it can be found should a "Turker" be keyword searching for tasks. Also, each HIT is given a "time allotment" to determine how long a worker has to complete a task before they forfeit it. One day should be ample time to complete the task with several stops for food, movie watching and internet lollygagging. The "HIT expires in" section decides how long the HIT lasts before being pulled off the market; in this case we chose three days. Since workers either have their work accepted or declined every time they complete an assignment, they build a history called an "approval rate." Requesters can then filter who is allowed to work on their HITs by worker approval rates. 90% is our minimum rate. Finally, as shown, a reward per assignment is established.

The reward per assignment is one of the constant variables affecting the Written Rummage project. Upon first launch, we were offering $0.01 for Text Image HITs. While the submissions' quality was usually commendable for a first draft, the time taken between submissions was, in short, too long—weeks, at times. Thus emerged a primary factor affected by reward per assignment: project completion speed. Though paying a penny per task is pocket-friendly, the amount of time to produce a finished product would be unacceptable. We therefore began increasing the compensation per task—both TI and Proofread—incrementally. Typical compensation now is $0.10-0.15 per TI and $0.10 per Proofread template. At this level, whole batches of 6-8 HITs can be completed before they expire in a day or two. Efficiency at this level is preferable for providing a

legitimate non-profit product to private collectors.

Clicking “next” at the bottom of the screen moves the design process on to the next phase. Here the Requester designs what the worker sees when interacting with a HIT. Thorough instructions and the resources for accomplishing a task are provided using this screen. Once the parameters are provided, the Requester can click to preview and finish the HIT. In Appendix B is an image of a TI preview containing what is asked of “Turkers” undertaking our HITs. The hyperlink “VIEW MANUSCRIPT” above the comment section opens the page that they are to transcribe. The comment section is where workers input the requested data.

Appendix B shows what Turkers see when they begin a TI HIT, but before that HIT is ever made available to Mechanical Turk’s workforce, it must first be “published” by the Requester. Once the template is complete, the Requester can click “publish” at the top of the screen, which opens a page that allows the Requester to upload .CSV (comma-separated values) files, which are used to create the HIT. A .CSV file typically stores tables of data separated either by commas or “return lines” in text files or cells in spreadsheet files. Every time we make a HIT, we must make a .CSV to determine which links will be available to the workers so that they can transcribe. For instance, in the TI template, a simple column of data—usually 6-8 entries, though many more are permitted—will create one HIT for every link entered in that column. Thus the template shown above will be duplicated and “stamped out” as many times as the amount of data entries, with each “VIEW MANUSCRIPT” hyperlink corresponding to a data entry in the .CSV file. The entire collection of these HITs is called a “batch.” A screenshot showing a typical .CSV file for a Proofread template will be shown a little below.

Once the batch is entirely uploaded, Mechanical Turk immediately distributes it to the crowd of workers. Each HIT will have the "shelf life" determined in the "HIT expires in" section of its design template. Either the HITs will be picked up by workers or they will be left to expire, meaning they will have to be republished in another batch. Throughout the process of the batch being completed, the Requester can check for submissions and choose to either accept the work submitted—and thus compensate that worker—or reject it, thereby not compensating him or her. A Requester must provide a reason for rejecting work; usually "HIT not attempted" or "This is not the requested material" is sufficient. Once work is submitted, the Requester has a certain amount of time to accept or reject the assignment, as was determined in the HIT's design template, before the worker is automatically compensated for the work he or she did. When this occurs, we do not receive the submitted work and compensation still happens—punctuality is key. Shown in Appendix C is a screenshot of the "review results" page.

Listed on the "review results" page are the unique ID assigned to each HIT, the Worker ID that also allows the Requester to track which workers consistently turn in acceptable and unacceptable work, the inputs that each HIT used and, finally, the worker's submission under the "Comment" column. After the requested comment has been submitted, the process for storing that submitted work and preparing it for proofreading begins.

Among the myriad innovations Google has provided its users over the years, the most useful for our purposes has been Google Docs. Google Docs allows users to use word processing and spreadsheet resources akin to what Microsoft Office has to offer, while saving those documents "on a cloud" so that users can access them whenever they

log into Google. Three sorts of documents are produced for our project using two formats: the copied-and-pasted TI submission as the first transcription of a given page from the Frederick Douglass Papers, the copied-and-pasted Proofread submission serving as the second and hopefully more accurate transcription, both of which are word processing documents; and the "Transcription Track" spreadsheet, which keeps tabs on which documents have undergone which degree of transcription.

After the requested transcription has been submitted in comment form, it is, as described above, copied and pasted into a Google Doc and saved such that it can be viewed by anyone with the URL. Part of Google Docs' usefulness for our project is contingent upon this assignment of a viewable URL; that URL can in turn be used for other workers to view and proofread, as will be described below. An image depicting a typical Google Doc is provided in Appendix D.

In the upper-left-hand corner of Appendix D, one can see the document's title juxtaposed with its privacy setting. First transcriptions are title in the form "Author_# of manuscript," while proofread transcriptions are titled as "Author_# of manuscript_p." This allows for clearer storage, since Google Docs can be saved in a list that can be sorted alphabetically, as well as a number of other less convenient ways. Privacy settings allow the document's creator to determine its availability to the broader internet audience. Three primary options exists: "Private," which allows the document to only be viewed by the owner and whomever they expressly share it with; "Anyone with the link," which allows anyone with the link or URL to the page to have access; and "Public on the web," which allows the document to be found and viewed by anyone using Google's search engine. Initial TI transcriptions are given the "Anyone with the link"

privacy setting so that workers can access the link from the HIT they are provided with, while proofread transcriptions are kept private. There is no need to make the documents available to the entire Google cosmos.

The manuscript "rough draft" transcription has found its home in Google Docs and awaits further use. Next in the process is developing the Proofread HIT in Mechanical Turk. No great innovation is necessary for the HIT template; rather, we use the same open-end template as the TI. The difference is in providing the workers with two links: "View Original Manuscript" and "View Digital Text." Appendix E is a sample Proofread HIT. Workers are typically compensated about a dime to compare the original manuscript and the first transcription, making corrections where needed. The qualification required is slightly higher at 95% (over the 90% of the TI HIT). Workers are expected to be able to complete the proofreading in a day, again allotting for any number of distractions and meals and breaks that could break up an otherwise fairly quick task.

When publishing a Proofread assignment, the .CSV file that must be uploaded into MTurk takes a slightly different shape. Rather than using one column of links, two columns are uploaded: the column of original manuscript links and the column of Google Doc links corresponding to those original manuscripts. Ultimately, rather than gradually obtaining these links piecemeal, a sort of "Pangea" .CSV file was made from which to draw all the corresponding manuscript and Google Doc links. "Pangea" is then systematically chipped away into smaller, more manageable .CSV files (shown in Appendix F), which are in turn uploaded into MTurk.

The managing process of Proofread HITs is nearly identical to that of the TI

HITs. The submissions are copied and pasted into Google Docs and saved in the fashion already described. Every time a document is saved as either a first transcription or a Proofread file, it is also marked accordingly on the Transcription Track spreadsheet saved in Google Docs. Once every document is adequately proofread, the project is complete. The last screenshot, found in Appendix G, shows the Transcription Track spreadsheet in all of its functionality.

Chapter 5: Data, Data Analysis and Existing Problems

A sufficient amount of work has been done to produce some semblance of a report on the Written Rummage project. Not all transcriptions from the sample Frederick Douglass Diary have been successfully transcribed and proofread; however, the process of transcription has been tried and crafted and reformed enough to present some coherent facts, figures, interpretations, projections and further problems to be solved in order to catalyze this project into a non-profit business.

As of now, 68/72 manuscripts from the Frederick Douglass Diary have had the initial TI transcription completed, and 28 of those have been proofread and resubmitted with actual improvements made. A striking total of $13.418 has been paid to workers for HITs and to Mechanical Turk as commission for its services. Discussion below will provide insight into how these figures, combined with current Written Rummage working rates for TI and Proofread HITs, project complete project prices being significantly cheaper than anything professional transcription agencies offer. First, some factors that have been integral to this study deserve special attention an obligation.

The primary immediate elements of concern relate to the pace at which a project can be completed, how to maintain and ensure quality control of the transcriptions, and how our rates compare to those of other transcription services that collectors might be subjected to in our product's absence.

To be forthright, on a broad scale this project's pace has been abysmally slow. Certainly, this was not without due course: learning curves are inevitably steep and slippery; developing a system for a new service with new technology is no exception. Thus, Written Rummage, beginning in Fall 2009, has yet to fully complete its pilot run to

definitively indicate its product's haste and quality. Start-up was slow for a number of reasons, such as learning to develop adequately usable HITs for users to use and fixing them accordingly, and in some respects even this process is not concretized. More saliently, the issue of compensation has proven to be a decisive factor in the pace at which transcription occurs.

Before turning to HIT compensation entirely, a fuller description of which processes might be in need of review or automation is in order. One inherently anticlimactic fact of the Written Rummage process is that when more submissions flood in, so does more tedium. Receiving more submission of either sort of HIT means having to copy and paste each submission to its own Google Doc, saving and renaming that Google Doc, establishing its privacy settings and marking its status in the Transcription Track spreadsheet. Some quicker, perhaps automated method of converting submissions to Word documents or Google Docs would make the project far more manageable, as hours on end can easily be spent clicking back and forth between MTurk and Google tabs when the work pours forth. Contact with MTurk will soon be made inquiring as to how this can be done through their resources.

Building .CSV files to upload into HIT batches is also another major time expenditure. This process could likely be automated much more easily than the submission transfer process discussed above. Constructing an initial "Pangea" .CSV has greatly expedited the process, since every subsequent HIT is a copy-and-paste snippet from the first grander collection of manuscript-submission corresponding entries. However, a method could likely be made reading the URL addresses of the submitted digital PDF manuscript images either into a series of smaller .CSV files or at least

one “Pangea” file from which smaller ones could manually be made. Similarly, an automated program loading the collection of Google Doc URLs to a spreadsheet could also expedite the process for building .CSVs for the Proofread HITs.

Among the major problems for Written Rummage is developing a reliable means of ensuring product quality without making the transcription and proofreading processes superfluous. The simple way of framing this problem is: “How do you know when a manuscript—and ultimately a whole collection—is complete to industry standards?” As it stands, Written Rummage can receive and proofread transcriptions; nonetheless, measuring accuracy is another undertaking. All measurement of progress thus far has been relative to the goal of transcribing all documents in the collection and proofreading those submitted transcriptions once. In order for quality control to be more effective, a different tack in the proofreading process must be taken, as this process has proven unreliable.

The last chapter described the Proofread HIT’s two-link format. While this format at first glance seems simple enough of a commission for a worker, having the first submission available to them also invites a lackadaisical approach. Two workers already have consistently submitted copied-and-pasted versions of the first submissions as proofread submissions. One could easily deduce the lack of effort from MTurk’s helpful figure, “Average Time per Assignment,” which for these submissions was 12 seconds. When the HIT submissions seem shady, more work left to the Requester, since in this case we had to compare and contrast the “proofread” submission and the first transcription to determine if there were any modifications made. We are therefore proofreading the work of the proofreaders, making the compensation for the time and

effort of a worker a frivolous expenditure. Measures must be taken preventing this sort of internal plagiarism and the resultant wasted time in quality control.

One possible option could be an adaptation from the Open Dinosaur Project's verification process. After a first submission is received, the same photo and tools are then released again to the crowd to be measured by someone else whom is unaware that the first submission has even taken place. With the two collections of data are then compared to each other. If they match, the data is received; if they do not, the photos are released again until matching data is received. While this cannot be perfectly mimicked in Written Rummage because of how dense with data each transcription is, a similar multiple-publication-of-HIT system could be built so that multiple Turkers work on the same HITs and thereby validate one another's work.

So then, does that not repeat the problem of having to read multiple transcriptions for similarities and differences? And, even so, in the case that none of the transcriptions perfectly match, which, if any, is chosen and how much money should be spent refining the documents? These are the continuing problems of the project. A means could likely be developed to automatically detect and identify the differences in long strings of characters, at least solving the first problem. As to the second, a criterion could likely be developed based upon the percent similarity that the documents share. Should two documents share a predetermined sufficiently high degree of similarity, perhaps a rule of arbitrary choice could be employed. Though no definitive solution has yet been formulated on this point. With that, a discussion of

Early trials of TI transcriptions were experiments testing how low rates could be while still having projects accepted and submitted. For the first month or so, therefore,

Ti HITs were paying between $0.01-$0.03 per page and Proofread HITs were paying between $0.03-0.05 per page. (In our beginnings, Proofread HITs were thought to entail more effort by virtue of having two documents to compare to one another; however, upon further review, the need for this payment gap proved nonexistent). While these rates did yield transcriptions and proofread copies, they unfortunately did not do so hastily. Twelve transcribed and proofread documents were done dirt cheap, but until two or three months had passed. Though other logistics were addressed during this time, the project's more material progress was not so apparent. Another issue exacerbating the first was the length of the HIT shelf-lives. HIT expiration periods force Requesters to act swiftly in adjusting payment rates and other HIT complications. Early on these periods usually spread between 5-7 days, whereas they are now 2-3 days maximum.

Since Fall 2009, compensation per page has significantly increased. TI HITs peaked at $0.15 per page, while the Proofread HITs have been stayed at $0.10; however, the Text Image request will be reduced to $0.10, since no marked difference in pace or quality of submission has been detected between the two rates. These rates, in contrast to more frugal ones mentioned above, tend to yield a full batch of HITs—typically 6-8 manuscript assignments in size—completed, or at worst 5-6. This pace far more viable for transcribing the documents that might be submitted by private collectors, especially those whom expect our service to adhere to deadlines. One-week deadlines still seem too pressing; one-month deadlines, depending upon the collection size, seem more viable.

When compared to traditional per-page transcription services, Written Rummage is still a shining star. For the 68 initial transcriptions and 28 proofread transcriptions that have been garnered, a total of $13.418 has been paid to workers and to Mechanical Turk

for its commission. Projections to finish the project in the manner it has been started approximate to a total of $18.70. This is reducible to $0.26 per page for transcription and proofreading. Should modifications be made to the proofreading process such that up to $0.40 were required per page, transcription for collections of similar size would be approximately $31.68. To contextualize these numbers, some transcription rates of professional agencies are provided in Appendix H.

A quick glance at Appendix H shows that the cheapest rates among the transcription services researched are Sage Word Service's "Up to 150 pages: $2.00/page; 150 and above: $1.5/page" prices. Rates range from $1.50/page to $8.00/page for The Transcription Place's "Standard Rush (1 – 2 business days)" service. Given our sample collection, the total transcription cost could therefore cost anywhere between $108.00-576.00. Our projected cost is less than a third of the lowest price and is an eighteenth the cost of the premium service.

The sheer difference in cost is adequate impetus to continue working through Written Rummage's quality control complications. Taking tips from other major crowdsourcing projects is likely the best tack to immediately take. An increasing demand for digitally searchable research information makes the cost-effective approach ideal for collections either feeling the pressure to meet that demand or wanting to make the full extent of their collections known to the public realm.

Chapter 6: Conclusions and Goals

In the end, the crowdsourcing business model that Written Rummage represents is groundbreaking. The "Information Age" is voracious; libraries and private collectors are looking for means of transcribing their handwritten manuscripts to avail them to the academic community and broader public. As the void they have yet to fill becomes more apparent, collectors are seeking out convenient, cheap ways of digitizing their documents. Digitization is the focus of a number of major projects going on in the academic and internet community; likewise, crowdsourcing is recurrently being used to accomplish otherwise tedious or impossible tasks. Written Rummage utilizes avant guard technology while bridging the gap between private handwritten manuscript collectors and the academic realm in search of their resources—and it does so affordably.

Some infrastructural complications still prevent Written Rummage from being fully functional on a large scale. Automations will allow the more time-consuming aspects of management to be expedited, while (slightly) higher wages for both sorts of HITs have already proven to hasten the transcription submission and proofreading process. Immediately pressing now is formulating a tenable process measuring and ensuring the quality of "final draft" transcriptions. Once this is done, the Frederick Douglass Diary will be compiled as a digital collection of searchable manuscripts and sent back to the Library of Congress for review (hopefully). Should they be impressed, further marketing will be done to employ the Written Rummage method on behalf of other libraries and collections. Should this prevail on a small scale, more direct steps toward developing a more formal and official non-profit business model can be taken.