Project Summary: Crowdsourcing for Humanities Research

Crowdsourcing offers opportunities to advance humanities research on two related fronts. First, the power of the crowd can be harnessed to assist humanities researchers in obtaining, cleaning, tagging, and analyzing the information contained in a diverse range of documents, allowing for deeper and broader interpretative work than was imaginable even a decade ago. Second, by engaging the public in humanities research through direct contact with the documents scholars use to construct arguments and interpret the past, crowdsourcing can build connections between academics and communities of interest beyond the walls of academe. In order to move forward with crowdsourcing in the humanities, scholars need to know whether this approach is appropriate for the kind of documents they work with, how to garner the interest of the crowd, and, finally, how best to create interfaces that encourage the public to participate in knowledge creation.

The project will bring together technology industry professionals, developers, and scholarly humanities researchers to apply engagement strategies proven in the game and social media industries to three distinct humanities research projects. We will explore how far these lessons can take us to aid in academic research.

In the first project, Tagging 500 Novels, we will create a user experience that invites the public into the social networks of novels, helping to explore and map relationships between characters in ways that computerized analysis cannot. In the second project, Year of the Bay, our focus will be on historical documents related to the San Francisco Bay Area and leverage local passion to derive information about maps and photographs that will inform specific historical research. Finally, the Western Railroads Project will seek to harness the knowledge of a niche community: amateur enthusiasts passionate about the history of railroads in the United States. The user interface tools we will build for the three projects will share a common technical platform, though they will each rely on different source inputs to inform the humanities research questions. Our aim is to identify and document key lessons from online user engagement strategies that can be specifically harnessed to aid humanities research.

The Center for Visualization and Spatial Analysis in the Humanities at Stanford University is the right organization to carry out this research. Our center consists of three allied projects, each exemplifying cutting-edge work in digital humanities: the Spatial History Project (initiated with generous funding from the Mellon Foundation in 2007), the Literary Lab (founded in 2010), and the Mapping the Republic of Letters project (founded in 2008). Leading faculty from the departments of History and English serve as the principle investigators of the ongoing humanities research projects herein proposed for the study of crowdsourcing. Research support is provided by 5.5 FTE staff with skills in programming, project management, and cartography. Strong institutional support from the Dean of Research, the School of Humanities and Sciences, and the Stanford Libraries undergirds our center and ensures that we will have the space and staffing to accomplish the proposed work. Additional funds from the Wallenberg Foundation (Sweden) will support the hiring of postdocs and the purchase of equipment and software vital to the projects. These resources have

been provided to the Spatial History Project and the Literary Lab as part of a larger funding package to support ongoing research in Wallenberg Hall.

Support from the Mellon Foundation will allow us to leverage the existing personnel by supporting direct staff and faculty effort and providing research funds for the research tasks contemplated in the body of this proposal. It will also provide the necessary resources for our partners in We Are What We Do (Historypin). Historypin will manage the technical implementation of the project, particularly the front end web presence and user interaction and engagement strategy. Backend data management and interoperability will be carefully coordinated with the Center for Visualization and Textual Analysis. Historypin is contributing one-third of their staffing costs for this project, and Mellon Foundation support will allow Historypin to dedicate 50% time for both a user interface designer and developer for the duration of the project, in addition to 20% time of a senior project manager.

Community engagement and crowdsourcing in humanities research

The promise of crowdsourcing in humanities research extends beyond the walls of the university. Our proposed projects provide the opportunity to encourage members of the broader public to interact directly with documents academic humanists engage with in their everyday practices. Whether this is through reading and tagging novels or uploading historical photographs and geotagging their locations, the tasks contemplated for the academic research projects outlined in section II of this narrative connect contributors to vital documents and real research. Yet, in order for crowdsourcing to become an accepted and effective means to conduct research and generate community engagement, we need to begin to answer questions that go above and beyond the purely academic. To wit:

1) What are realistic costs and expected results for a typical crowdsourcing project?
2) When does it make sense to utilize paid labor through Mechanical Turk or oDesk versus a crowdsourced community engagement strategy?
3) Are there educational benefits from community engagement strategies that add to or justify return on investment?
4) Do game mechanics improve quality or quantity of data returned, or add to successful engagement strategies?
5) Can history enthusiasts or non-academic domain experts be encouraged or trained to create or add to primary sources in a way that will enable academic citations?
6) What are the expected lifetimes of crowdsourced humanities projects?

The community engagement and digital development aspects of the project will be conducted together with We Are What We Do, the not-for-profit behavior change company, whose Historypin project serves as an active portal for "Citizen History." Historypin officially launched in July of 2011 and since then is averaging about 600,000 unique visitors a month, with over 300,000 total iPhone and Android app downloads. We intend to engage this audience for the purpose of this study, by providing captivating micro-volunteer, crowdsourcing, and gaming opportunities. Jon

Voss, who will serve as the Historypin project manager on this project, recently organized the International Linked Open Data in Libraries, Archives, and Museum Summit, funded by the Alfred P. Sloan Foundation and the NEH.

A Unique Partnership

Historypin is a project created by the non-profit organization We Are What We Do as a project to help bridge the growing divide between older and younger generations. We Are What We Do is a behavior change company based in London, San Francisco, and Sydney that creates ways for millions of people to make small, positive contributions to substantial social issues. In essence, the organization designs and creates desirable and useful products and tools which have positive behaviors built into them, aiming to incidentally but powerfully harness mass behavior.
For the purpose of this project, we refer to the "Historypin team," who will be responsible for designing and coordinating the community engagement and technology tools. Technically, team members are We Are What We Do staff dedicated to the Historypin project.

Historypin's design was based on research and testing undertaken in the east end of London, which pointed toward ways to leverage technology to engage diverse groups of users in history and generate mutually enjoyable and constructive experiences. The social and technical aims led to Google sponsorship of the project: in August 2009, Google became Historypin's non-exclusive technology partner and provided funding for the initial Historypin site. Google provided limited in-kind marketing support around the July 2011 Historypin launch in New York, and has since made several financial donations to We Are What We Do in support of the project. Historypin is hosted and built on Google App Engine, and receives discounted hosting rates. Historypin utilizes standard Google APIs for maps and Street Views and is part of the Google Maps API Premier for Non-profits[1] which allows the project the use of internal map deployments, higher geocoding limits, the option for no ads to appear on the map, and Google technical support. Historypin is wholly owned and controlled by We Are What We Do and Google has no legal, financial, technical or other ownership of the project.

Early in the Historypin beta phase, Nick Stanhope, CEO of We Are What We Do, contacted the Spatial History Lab to explore ways to utilize Historypin for humanities research. These discussions led to collaborative work, currently in early phases, with Michael Levin, a documentary filmmaker, and the East Palo Alto community. The project, conceived of as an experiment in community mapping and spatial history, utilizes the online tools and community created by Historypin combined with traditional archival sources in order to facilitate the work of citizen historians in an underprivileged community embedded in the heart of Silicon Valley. Staff and students from the Spatial History Lab work hand-in-hand with Historypin partners and a multigenerational team from East Palo Alto in an effort to map and narrate the complex history of their community, including, for example, the dispersion of High School students throughout the greater Silicon Valley region in accordance with a voluntary desegregation decree (Tinsley) that allows youth from East Palo Alto to apply to attend High School in neighboring (and very affluent) communities such as Palo Alto and Menlo Park.

Proposal Outline

In the first part of the narrative, we offer detailed proposals for three research projects with diverse documentary bases, target communities of interest, and scholarly goals. We explain the logic behind the selection of the projects and trace out the way they are designed to provide both specific information regarding a particular kind of humanities research relative to crowdsourcing and general knowledge and tools that can be shared much more broadly. Each project is described in detail with the aim to elucidate the kind of questions our projects ask and the methods we intend to employ in order to test the efficacy of different modes of data collection, interface design, and community engagement.

Second, we explain how the overarching project fits into the theme of community engagement in humanities research through crowdsourcing. We describe the collaborative relationship between the Stanford researchers and Historypin and outline the specific goals relating to the current state of knowledge regarding crowdsourcing and our understanding of the key areas in which further research is required.

In the third section we discuss dissemination of our findings, data, and opensource toolsets, as well as plan for intellectual property.

Project Narrative

I

Crowdsourcing: a tool for major humanities research projects

We propose three discrete research projects designed to highlight particular challenges and opportunities associated with crowdsourcing. The range of the projects are indicative of the methods we will employ to test engagement across a series of open toolsets for transcription, tagging, metadata improvement, and map warping in the interest of specific humanities research questions. The need for comprehensive and rigorous research along these lines is clear. Whereas a growing literature speaks to the use of crowdsourcing in academic contexts, no study directly and comprehensively addresses the range of challenges facing humanists wishing to enter into this space and harness the power of distributed effort via the internet.

Our projects were selected on two criteria. First, we considered the range of questions we wished to ask regarding crowdsourcing. In this regard, selection was made to create a portfolio of research projects covering a broad range of topics, from tagging text in novels to identifying locations and providing metadata for historic photographs of railway lines. With this approach, we expect to be able to identify both the common tools and potentialities of crowdsourcing as well as the important differences owing to the kinds of questions asked and the sources consulted. Second, the projects were selected on the basis of their strong support in terms of faculty and institutional resources. Our ambitious goals and timeline required that the projects be sufficiently developed and staffed such that clear research questions could be posed and ample resources be put in place to make the most efficient use of the Mellon Foundation's support. In addition to these criteria, we considered the degree to which each project could profit directly from collaboration with our Historypin partners. Tagging 500 Novels, Year of the Bay, and the Western Railroads Project met all of these requirements, and provided enough diversity in content and public interest to make them excellent candidates. The availability of open source libraries to draw upon for the technical design of the tools was an added benefit: transcription for Tagging 500 Novels, and photo geotagging and map rectification for the Year of the Bay and Western Railroads Project.

If we step back and survey the landscape, it is clear that the humanities, cultural heritage and academic institutions are increasingly turning to crowdsourcing and public micro-volunteering efforts to make sense of immense or untouched datasets too large to be practically utilized by individual researchers, yet which hold enormous potential to contribute to humanities research. While an array of open source tools have been created to facilitate engagement, these crowdsourcing efforts have had varying degrees of success, with mixed empirical evidence indicating elements necessary for accurate input, design requirements, public engagement, or best practices, particularly with respect to materials used in the humanities and humanistic social sciences.

The need for a more comprehensive project is clear. Crowdsourcing offers an opportunity to advance humanistic work by applying distributed human ingenuity to solve problems that computers alone cannot and by fostering broad engagement with humanities materials beyond the walls of the university. In order to pursue these opportunities effectively, humanities researchers and their collaborators in interface design and dissemination need tested working models on which to build efficient projects. We need to know what works best with what kinds of material, and we can begin to answer these questions by conducting research on tagging novels, producing maps and metadata regarding environmental change, and collecting expert information and archival material from communities of interest.

In order to move forward with this research, a partnership between established humanities researchers and experts in the realm of interface design and web-based community engagement is required. Humanities scholars in traditional disciplines, such as history and English, lack the requisite skills and industry connections to develop and test high-quality crowdsourcing materials. Yet, without a connection to genuine research projects led by established scholars, crowdsourcing in general, and specific interface design and community engagement strategies in particular, are destined to remain on the periphery of humanistic research. This project brings together the vital inputs in the form of humanities questions and the necessary technical competence and global visibility required to evaluate the utility of crowdsourcing across a range of materials including text, tables, maps, and photos. In short, we believe Stanford is the right place to conduct this work (we have the people, the ongoing project momentum, the leadership and resources in place) and Historypin is the right partner to take these ideas to the public and make this happen on a large scale (in Historypin, we have the technical skills and community engagement capacity together with a proven web platform upon which to launch our projects).

1: Grammar of social networks
Tagging 500 Novels

This project, led by a research team from Stanford's Literary Lab, will extract and quantify character networks in 19th century fiction in order to investigate traditional literary hypotheses regarding the extent to which social settings (primarily urban or rural) correlate with social networks and their size, type, and density. More specifically, these hypotheses suggest that changes in novelistic form, plot, and character might be correlated to changes in real-world societal conditions, especially those involving revolution and industrialization. When it comes to character, these hypotheses tend to focus on how the number of characters and the nature of character networks in fiction appear to fluctuate depending upon the "social world" being depicted. Raymond Williams (1975) uses the term "knowable communities" to describe the networks made possible or feasible given a set of social conditions. In rural settings, for example, characters tend to organize and interact in what might be seen as "familial" structures. The networks made possible by urban settings are different: In Atlas of the European Novel, Franco Moretti (1998) argues that urban settings tend to create character networks of a less dense and more superficial nature. Moretti

writes that "the narrative system becomes complicated, unstable: the city turns into a gigantic roulette table, where helpers and antagonists mix in unpredictable combinations" (68).

Research on social networks in fiction has been recently undertaken by scholars at Columbia, and has "provided evidence that a majority of novels in this time period do not fit the suggestions provided by literary scholars" (Elson, Dames, and McKeown 2010). The Columbia research, however, is limited by its reliance upon a purely algorithmic approach to the problem of character identification and extraction. The research proposed here will begin by employing algorithms similar to those used by the Columbia research team (e.g. Named Entity Recognition and pronoun disambiguation algorithms). After this unsupervised, algorithmic tagging, human input (via crowdsourcing) will be used to confirm and correct the results. The initial algorithmic phase will correctly capture the least ambiguous entities and the second, human phase, will focus on the more difficult situations, such as the following sentence in which the identity of Mr. Jones is ambiguous: "But your father, Mr. Jones, is he living?" Here the human coder would use context to determine if Mr. Jones is the character being addressed by the speaker of the sentence or if Mr. Jones is the addressee's father. Our proposed research will also differ from prior work in venturing well beyond the standard literary corpus. Our dataset presently includes several thousand 19th-century British and North American novels.

The promise of "distant reading"—of examining thousands of texts instead of close-reading one or a few texts—has been demonstrated by Franco Moretti (1998, 2005). The computational tools and a variety of methods for realizing Moretti's notion of the "distant" have been articulated by Jockers (2012). In spite of good progress in terms of analyzing style and theme at the macroscale, some aspects of distant reading are especially challenging owing to the overwhelming difficulty of replicating algorithmically what human beings can do quite naturally. Detecting characters and their associated networks of interactions represent a substantial challenge that computation alone is not yet prepared to address; in fact, computation may never be able to fully or even adequately address this problem. Computational approaches to character identification and name disambiguation make far too many mistakes and fail to achieve the level of accuracy that is necessary for humanistic inquiry. These failures are largely associated with issues of name variance and disambiguation. If we wish to analyze spoken dialog in novels, for example, it is imperative to attribute dialog to speakers and receivers, but with nicknames (e.g. "rascal", "dodger"), abbreviated forms (e.g. Matt vs. Matthew), and the frequent use of pronouns (he, her, she, etc), it is simply not possible (currently) for a computer to correctly assign dialog to all the possible manifestations of character names and pronouns.

A combination of algorithmic techniques and crowdsourcing offers a reasonable path forward and a way of taking current research beyond its present state. Crowdsourcing is especially appropriate in this research because:

1) The material, though inherently complex for a computer, is relatively easy for humans to disambiguate.

2) The material is available in digitized form and in the public domain.
3) There is a natural constituency of interested users—millions read novels, many are fans of particular authors, such as Jane Austen—as such, the project takes advantage of existing interest and distributed expertise.

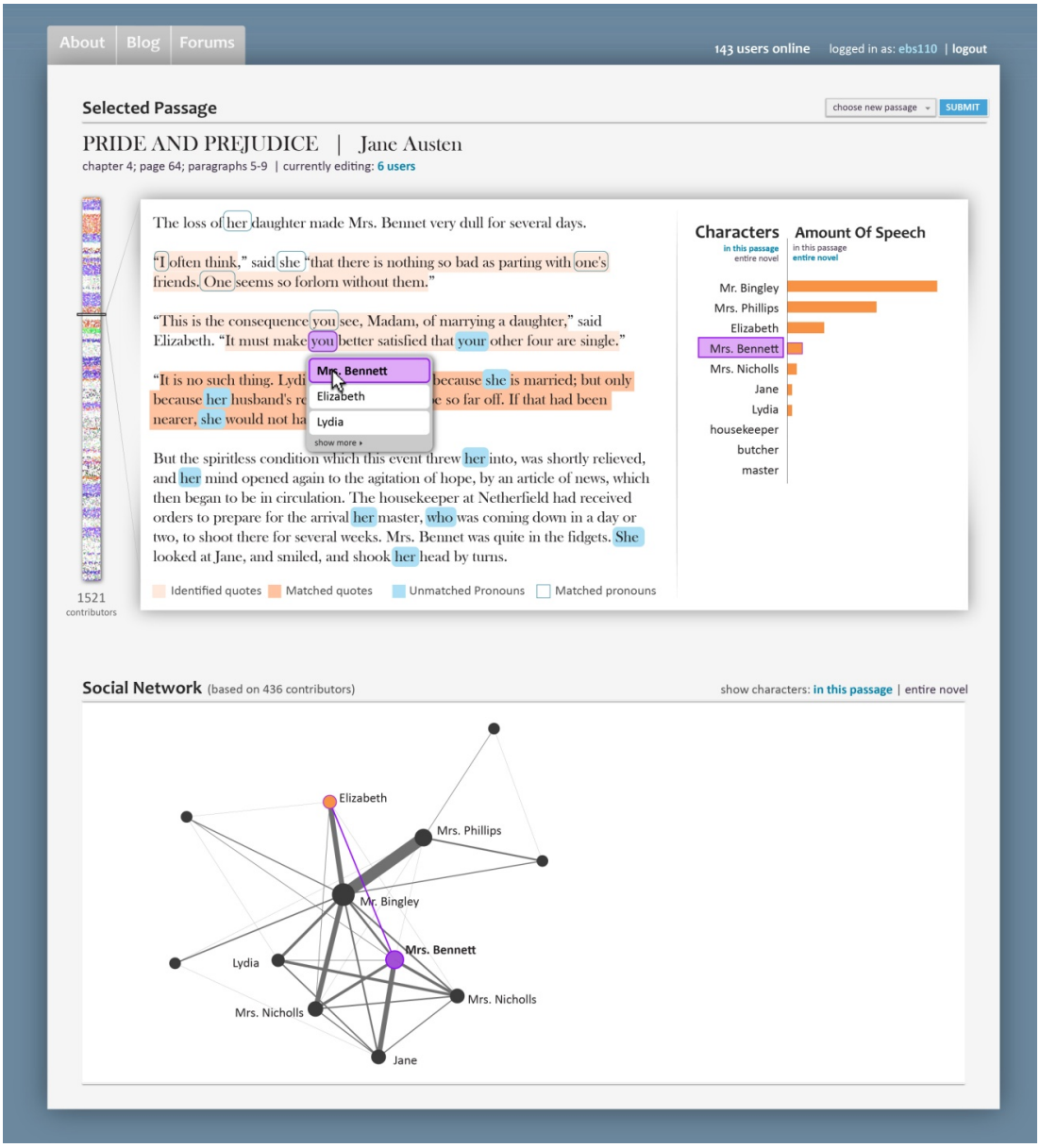
In order to provide a clear example of the kinds of research questions we hope to ask using the tagged novels, we hypothesize that network configurations based on character co-presence will differ in significant ways from networks formed by directed speech between characters in novels. Co-presence will tend to inflate the size and complexity of the network. Characters can be present in the same space and yet not interact directly. The nature of a co-presence network is, by definition, undirected. Two characters share space, but the "network" cannot tell us anything about the nature of the interaction, direction of speech, or intensity. By associating speech acts with speakers and receivers, we expect to uncover an important dimension of social interaction depicted in novels. Studying thousands of novels will allow us to distinguish the different patterns of character exchange and the different social network structures these exchanges create.

We will examine these patterns in the context of different novelistic genres (e.g. Gothic vs. Bildungsroman) and in the context of time (i.e. do the dominant network structures evolve or change over the century). Along the way we will develop a deeper understanding of the interconnection between fictional social structures and the social structures associated with the emergence of industrial capitalism and urban-centered life during the periods traced. The Stanford Literary Lab research team has already begun a similar investigation of drama (where, because of a formal style that includes speaker names for every "speech," character disambiguation is not necessary). The team compiled a corpus of over 250 plays and has used both algorithms and human labor to identify the recipients of every speech act. This research has led to some revelations about, for example, the difference between character networks in Greek drama versus Shakespearian drama as well as some observations of how the networks in Shakespearian tragedies differ markedly from Shakespearian comedies (Moretti 2011).

As a preliminary test of the concept in the novel form, we tagged the novel Memórias Póstumas de Brás Cubas, by Brazil's greatest novelist, Machado de Assis. The results of this proof-of-concept indicate that our expectations regarding the important differences between the co-presence (undirected) and speech connected (directed) networks will be confirmed. The number of nodes (characters) in the network falls from 111 in the undirected graph to 56 in the directed speech network. Additionally, the rank order of the most intensely connected characters also shifts. Whereas co-presence can be reasonably well determined computationally, directed speech and pronoun/name disambiguation requires human input to tag speech correctly. With two measures of novelistic networks, the analysis undertaken by the project will be substantially enriched, the strengths and weaknesses of the two measures will be assessed, and social network analysis of literary texts can be assayed.

The literature on crowdsourcing suggests that any humanities project wishing to draw upon this resource of collective labor and expertise will need to address at least four fundamental problems, which can be summarized in relation to a graphical representation of the space of crowdsourcing. These are: 1) the problem of limited commitment to tasks; 2) the problem of expertise relevant to tasks; 3) the problem of task type and incentives relative to quantity; 4) the problem of task type and incentives relative to quality. Research shows that the best results in crowdsourcing are obtained when tasks are simple and there is a mechanism to check for quality and to avoid garbage input—e.g. gaming the system, especially relevant in cases where there are payments for tasks (Kittur, Chi and Suh 2008). The evidence suggests that paying people more to complete tasks will increase the quantity but not the quality of user inputs (Heer and Bostock 2010; Mason and Watts 2009). Productivity, in turn, may be highly correlated with levels of attention given to the participant—hence suggesting a place for feedback mechanisms to reward effort with attention in the form of awards, game points, and the like (Huberman, Romero and Wu 2009).

Most humanities questions worth asking are at least moderately complex and require some degree of thoughtful commitment. They thus would seem more complex and to require more expertise than classic crowdsourced tasks, where task complexity, attention, time, and cognitive load are minimized. Our experimental interface design seeks to test the ways in which humanities tasks can be shifted toward an increase in the degree of commitment and level of expertise as much as possible while still retaining the ease of use and broad participation rate necessary for crowdsourcing to be truly effective. As a concrete expression of our ideas, we developed a mock-up of an interface that addresses these concerns. In the accompanying screenshot, we suggest a model for tagging pronouns to nouns, which provides crowdsourcing participants real time feedback on the changing structure of social networks in the novel as their tags are registered.



Our interface design aims to create a two-way dialogue between the task and the user – thus maximizing engagement while minimizing complexity. The interface not only highlights the pronouns in need of disambiguation, but allows the user to select the pronoun's antecedent from a precomputed list of characters—a list that it will continually re-sort—based on the current status of the data, and according to an estimate of the likelihood of each character's serving as the antecedent in this instance. Alongside the passage, further feedback is displayed: the characters mentioned in the proximity of this passage, sorted by the likelihood of their appearance in this passage; and a continually updated representation of the social network of the novel, to provide both feedback to the user making decisions on the likelihood of social interactions, as well as to foster engagement with the process of entering data.

The interface design also allows users to collaborate by focusing on different portions of the text, or on different aspects of the task. By displaying exactly which areas of the novel are still in need of disambiguation, the interface will help users locate the areas of the novel most in need of their input. Also, the individual tasks themselves build on top of each other in such a way that encourages collaboration. For example, once a novel's pronouns have been linked to their antecedents, a computer could reliably attribute most of a novel's dialogue to the characters speaking it, by linking the quoted speech to the pronoun introducing it. Users will then only be asked to attribute speech to a character when either no pronoun is associated with it, or when other confounding factors intrude – thus minimizing what we demand of our users, while also increasing the flexibility and interchangeability of the data-entry process.

Project 2: Engaging a broad public in crowdsourcing through a prominent regional event Year of the Bay

The Year of the Bay — 2013 in the San Francisco Bay Area — gives us an opportunity to test how to engage a broad public audience in crowdsourcing by tying one of our projects to a prominent regional event with compelling public and media opportunities. The Year of the Bay includes the America's Cup, one of the world's top sporting events and the opening of a new Bay Bridge span, but also events and exhibits at many of the region's historical and cultural institutions, including museums and libraries, which are interested in partnering with us in a regional effort to mobilize volunteers to rectify historical maps, digitize shapes, and identify elements on the maps; place historical photographs and documents in the landscape, and identify entities in the photographs and documents, and thus improve the metadata and usefulness of these objects; and in the process to build from the ground up an environmental history of the San Francisco Bay Area. This project will rely on diverse datasets and will result in a new crowdsourced collection of digital, geospatially and temporally located objects that can be used by scholars to address crucial contemporary research questions in environmental history and humanities and historical ecology, including:

1) Shifting baselines. Environmental historians and historical ecologists have begun to realize that our expectations for a healthy California landscape have been severely distorted by the last 100 years of history, in which modern infrastructure has created landscape features, such as perennial streams and lakes, where none existed before in the dry summers of this Mediterranean landscape. This project will help uncover historical landscape conditions on a detailed scale, so that we can better understand the past, and the possibilities of these environments in a changing climate with looming water shortages.

2) 20th-Century Modernization and Environmental History. The dominant historical narrative of the 20th century is that modern Americans turned their backs on the environment, at a time when creeks were covered over and the San Francisco Bay was nearly filled in. Brief forays into primary documents reveal more complex relationships than this dominant narrative suggests between different ethnic communities and the bay and its resources throughout history, including the last

century. Digital collections will allow us to examine these changing relationships of resource use, recreation, and aesthetics and create a multi-layered history of the changing, but enduring relationships with the bay's environments, even as these environments and communities experienced rapid modernization.

3) The Possibilities of History, the Present, and the Future. Through this crowdsourcing experiment, we will be able to assess whether exploring and practicing the work of doing history — finding and interpreting historical sources and placing them in space and time — changes participants' understanding of the possibilities of present landscapes, and features that may be hidden from view, such as buried creeks, and the potential for future environments.

Crowdsourcing is particularly appropriate in this case because:

1) A large body of material is available in digitized form, but only a small percentage of the historical documents and photographs can be geolocated using metadata. A much larger percentage can be located using human judgment and correction.

2) There is a large, enthusiastic amateur community for this work, as has been demonstrated by several informal efforts already undertaken in this area (see below).

This project will result in rich datasets allowing scholars to use these historical sources 1) to document complex changes in the landscape that have not been previously understood by the dominant narratives of environmental change, most of which have been written at a regional scale; 2) to understand human adaptation to environmental changes and how the landscape was used by different people over time (much of this complexity of quotidian human relationships to the changing land has not been captured by other sources); and 3) to interpret the ways in which changes in the landscape and changes in human relationships to the landscape were represented differently over time in the sources, particularly in the visual sources, photographs and maps. Support for this work will be provided by the Spatial History Lab through the work of its expert staff in the areas of cartography and GIS, as well as by a postdoctoral scholar in residence funded by the Wallenberg Foundation and the Bill Lane Center for the American West.

Datasets include the San Francisco Public Library's digitized photography collection, David Rumsey's digital map collection, digitized Sanborn maps, the San Francisco Estuary Institute's digitized historical ecology collections of maps and primary textual sources, and the historical photography collections of the San Francisco Public Library and the California Historical Society, among many other digital collections. The digitization of these collections has outpaced the ability of scholars to use them effectively, and the work of rectifying, digitizing shapes, and elements on maps, and placing photographs and documents in relevant locations is too time consuming for single scholars to undertake. None of this work can be automated fully at this point, but a large, enthusiastic amateur community for this work has been demonstrated by several informal efforts already undertaken in this area, including Maptcha, a project by Michael Migurski of the San

Francisco-based design firm Stamen that has empowered a crowdsourced project to quickly and easily georectify all of Rumsey's Sanborn maps of San Francisco; OldSF, in which one enthusiast geolocated 13,000 historical photographs from the San Francisco Public Library based on metadata (many more remain to be located by human judgment); and many other such projects aggregated by "Burrito Justice," the internet alias of Silicon Valley technologist John Oram, whose blog and Twitter feed, with 41,961 followers, attracts and features dozens of historical map, photograph, and document enthusiasts eager to dig into local history and create historical mashups of their own.

These datasets will be imported into the "Year of the Bay" site that Historypin will construct for this project in various ways. In some cases, such as the Rumsey map collection, we will be able to freely import the metadata and digitized objects. In other cases, such as the San Francisco Public Library, we will be able to freely import the metadata and low-resolution versions of the digitized photographs. In other cases, such as the San Francisco Estuary Institute, we will have to work more closely with the organization's staff to select classes of materials defined by different intellectual property restrictions. In some cases, with the Institute's materials we will only be able to use metadata, in some cases we will be able to use low-resolution images, and in other cases, we will be able to use high-resolution images. We will work with a variety of institutions, including those like the San Francisco Estuary Institute, which have been slow to open up their archives because of legitimate concerns that the rights to a substantial portion of the maps, images and documents in their collections, which they have used in their historical ecology research, are controlled by other institutions. SFEI, for instance, has a variety of rights to the digital copies of documents in its archives — some they own, some are out of copyright, others they only have permission to use in their own reports. We will advise institutions such as SFEI on various options from simply allowing us use metadata and low-resolution copies of images, to image licensing, and in the process demonstrate how to work with institutions all along the spectrum of openness in regards to sharing their data publicly.

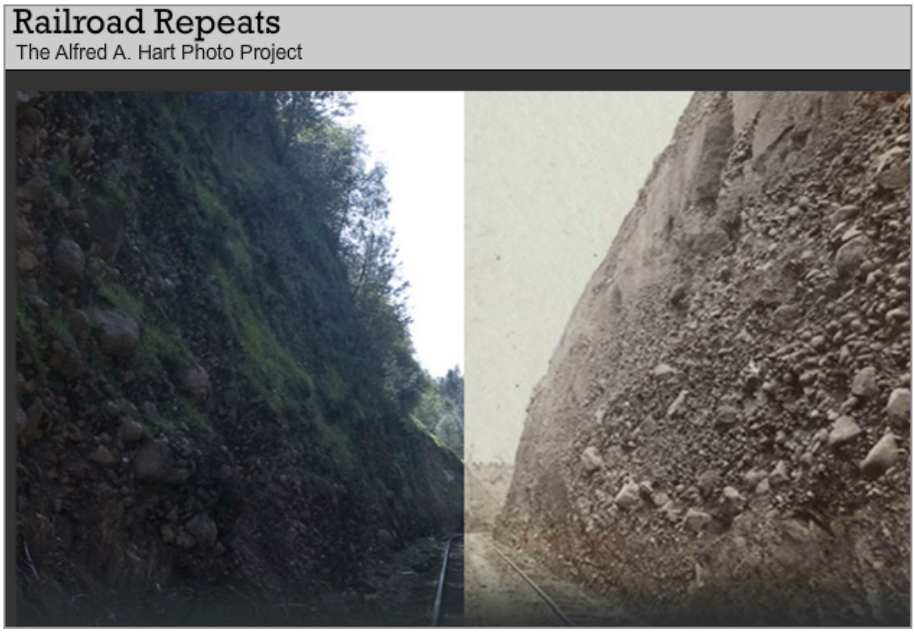
The digital toolsets that will be built by Historypin for this project will test the use of games and game-like incentives to encourage, reward and structure the experience of participants in a scholarly, historical pursuit — much like a detective's — as they geo-tag historical photographs and primary textual sources and rectify historical maps around the bay, learning as they go and contributing to the building of a larger crowdsourced environmental history of the region**.**

This project will benefit from the tremendous opportunities for publicity tie-ins with the "Year of the Bay" to reach a very broad public, particularly since the project will provide a creative way for ordinary citizens to be involved in the celebration of the bay's environment, history and cultures. The partners in this project, particularly the Spatial History Project's Jon Christensen, the San Francisco Estuary Institute, and the California Historical Society, have extensive media contacts and a track record of success in generating good media coverage for stories that connect historical research to contemporary environmental concerns and citizen engagement. The aim of this project will be to generate media coverage to fill a very wide funnel for public engagement. Historypin will

design interactions that interest, stimulate and reward the most simple forms of engagement with the project, and then steadily move players up the scale of engagement.

Project 3: Cross-media engagement with expert communities
The Western Railroads Project

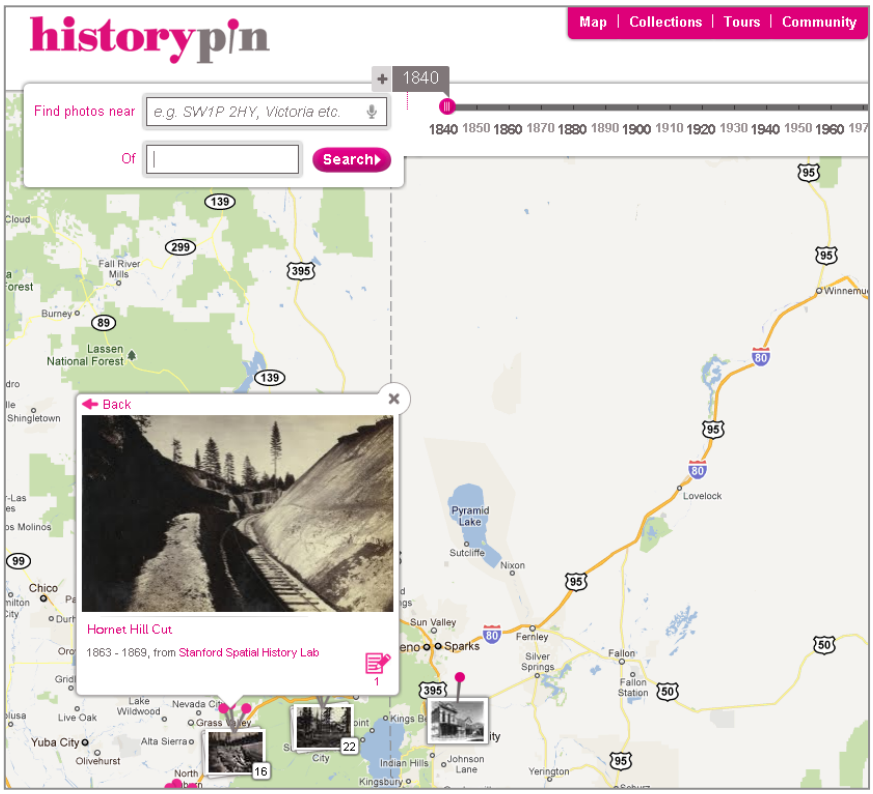
Building upon the research underpinning Richard White's *Railroaded: The Transcontinentals and the Making of Modern America* (Norton 2011), the Western Railroads Project seeks to document the construction, environmental context, and everyday historical appearance of the railroads throughout the North American West (http://www.stanford.edu/group/spatialhistory/cgi-bin/railroaded/). Richard White commissioned a preliminary proof-of-concept study in 2009 to study whether it would be possible to use repeat photography to document the changing landscapes surrounding the railroad by replicating Alfred Hart's famous images of the birth of the railroad age in the West (http://www.stanford.edu/group/spatialhistory/Visualizations/Hart/). The project involved locating the modern site of the Hart photographs and retaking the photographs from as close to the same position as possible.



The project was successful in that we have replicated virtually all the Hart photographs in California and Nevada, given them GIS identifiers, and plotted them on Google Earth, but at the same time we recognized that repeat photography is both arduous and expensive. It provides matched pairs that provide evidence of landscape change (or the lack of it) and the social development, but it tells you little directly about the process of change in the intervening century and a half. The project has

inspired us to attempt a much larger, "open-source," project to collect, tag and interpret the thousands of photographs of railroads in the United States.

Working with our partners in Historypin, the Hart collection photos have been geotagged and inserted in a web interface which receives an average of 20,000 unique user visits per day. Our ambition is to expand the Hart website in both depth and breadth. By expanding it in depth we intend to solicit photographs of the same sites that Hart photographed that fill in the century and a half between the original photograph and the repeat photograph. These will not be repeat photographs since they will not be taken from Hart's original vantage point, but since many of Hart's photographs are of cities or iconic landscape features we should be able to capture enough similar photographs to begin to record environmental and social change. By expanding the site in breadth, we intend to fill in the spaces that Hart did not photograph with other railroad photographs. Finding such photographs is only part of the problem. We need to locate them in space, and crowdsourcing could provide a solution to both problems.



Using its expertise in community engagement, Historypin will create and help implement a campaign to solicit historical photos of railroads in the western United States and, in doing so, provide an experimental interface design that will allow this project to assess the best practices associated with providing incentives for users to upload content, as well as measures to assess the quality of the user-provided metadata regarding the place, time, and context (including

environmental details) of the photos. This database will be primed with public domain historical photographs of the railroad lines from the Library of Congress and the National Archives.

Concomitantly, White's research has focused on the use of maps to reconstruct the spaces of settlement, environmental conditions, and conceptual elements of the space shaped by the railroads. Connecting photos to places provides the ability to connect two fundamental media (photos and maps) in a cohesive online interface. By doing so, this project brings together two communities of interest with proven passion for their subject: railroad buffs and map lovers. The recent donation of David Rumsey's unparalleled map collection to Stanford University provides us with a unique resource. Among Rumsey's maps are sets of county atlases that record a wealth of spatial data—land owners, sites of schools and post offices, roads, irrigation canals, now vanished landscape features – that we can only use by painstaking, and often tedious, individual labor that involves not only geo-rectifying the maps but then tagging all the features so that we can analyze them in GIS. Our ambition is to find a way to crowdsource this material. We will geo-rectify the maps but will try to get people interested in local history, railroads, and genealogy to tag and locate the map details for us and, when possible, to locate and identify historical photographs in the county of both sites and people. We hope that crowdsourcing will allow us to create a spatial database that we otherwise could not achieve on our own and open up the Rumsey map collection to spatial, social, and environmental analysis that is now impossible except on a very small scale. To get a sense of what we seek to do on a large scale, see Cameron Ormsby's county level study of land holdings in Fresno and Tulare County on the Spatial History Project web site. At Stanford we have the necessary institutional support systems, including the consulting support we have arranged with G. Salim Mohammed, Digital and Rare Maps Librarian, required for undertaking this ambitious project.

Crowdsourcing is particularly appealing in this case because this project taps into a large and dedicated community of interest with perhaps unparalleled knowledge about the specifics of railroads and local history. These railroad buffs and local historians are often dismissed as antiquarian, but they command a deep and quite specialized knowledge. Moreover, the focus on photographs draws on existing strengths in our partners from Historypin. The digital toolsets we will utilize for this project include two implementations of the crowdsourcing GUI. The first will be recruit users to assist in the geotagging of railroad photos in the US, gathered from photographs already contributed to Historypin or added during this campaign. The second implementation will recruit users to identify and tag features on select railroad maps, which can then be harvested for geospatial analysis. As with all of the projects, simple game mechanics will be tested here to track and reward contributors. Additionally, we will utilize the Historypin mobile apps to enable users to find the location of older photos and take current photographs from the same location, which are automatically uploaded to Historypin with geospatial data. The digital tools that Historypin will build for this project will enable users to upload and geolocate historical photographs in their collections or modern photographs of location along the railroad lines using mobile apps. Simple game mechanics will be tested here to track and reward contributors.

Although the basic digital data on railroad routes, land allocation along those routes, traffic across the routes, and labor on the railroads comes from either extant databases or databases that we have created through archival research, there is a great wealth of material in the hands of buffs, amateur historians, and genealogists. These materials include, but are not limited to, photographs, letters and diaries. These materials often can be placed according to location and time and thus are amenable to organization on a site such as Historypin. Historypin would become the site on which we could gather crowdsourced data for further analysis. It would be of interest in and of itself to those who upload material on it, but it would also allow us to repurpose the material for further environmental and social analysis. We might, for example, be able to visualize the changes in a particular town, say Laramie, Wyoming, over a period of years in ways of interest to casual viewers, but also useful for the analysis of urban history, the construction of social spaces, the actual use of those spaces, and their change over time. Photographs, for example, would capture the rate of landscape change, the density of houses, farms, and businesses along railroad line and also the rate of environmental change. It is very often possible to identify species and thus species change over time. It is also possible to see the changes in forest cover, the plowing of grasslands, the consequences of changes in burning patterns, and the consequences of grazing and timber harvest. Important in itself, these sources can then be linked to other sources from censuses and other material to get a much more refined picture of how rapidly environmental change takes place and how it links with social and economic change. Crowdsourcing promises to be not only a way to gather this data, but also a way to identify it. We can crowdsource questions of location and date when these, as is often the case, not known. The idea is not that we accept what we are told, but that we tap a wide realm of interest and expertise among amateur historians, genealogists, collectors and railroad buffs.

We will reach out to this community by providing short stories and compelling calls for engagement for their blogs, newsletters, and email lists. We will also develop symbols of engagement such as badges that will symbolically reward and recognize their involvement in the project. The new sources developed, geolocated, and provided with metadata in this process, will be useful for scholars as they study the environmental and social transformations wrought by the railroads—when they were successful as well as when they were not.

II

User Interface and Interaction Design: Design School at Stanford and Industry Advisory Panel

The User Interface and Interaction design elements of this project will represent a unique collaboration between the humanities scholars and Design School at Stanford University, with a significant advisory capacity from industry leaders at Google, Bolt | Peters, and other Silicon Valley firms. We're recruiting an advisory board for the project that will consist of five to seven industry leaders in user interface and user interaction design, semantic technologies, and the computer game industry. Industry advisors will meet with our team at key intervals of the study to guide the general direction of our user interface research and outreach. A total of at least six, four hour advisory meetings will be scheduled throughout the course of the research and will take place at the Stanford dSchool and Spatial History Lab. Full bios are listed in the appendix. Confirmed participants to date include:

Jamie Taylor, Minister of Information, Metaweb/Freebase at Google
George Oates, Art Director, Stamen Design, and Research Associate with the Smithsonian Libraries
Cyd Harrell, Vice President of Research, Bolt | Peters
Justin Quimby, Vice President of Product, TinySpeck
Abigail Phillips, former Senior Staff Attorney for Intellectual Property at Electronic Frontier Foundation
Ethan Watrall, Associate Director of Matrix: The Center for Humane Arts, Letters & Social Sciences Online at Michigan State University

Building on Related Work in the Field

A growing number of successes in public projects utilizing crowdsourcing or gaming in science have increased the interest in soliciting public help to sort through texts, maps, photographs and other digital assets and metadata relevant to humanities research. A number of open source transcription and mapping tools have emerged to help aid these efforts, and a wide use of various developer application programming interfaces (APIs) have also been utilized to harness public input.

Zooniverse. Certainly the gold standard in recent years for successfully harnessing the work of "Citizen Scientists," the Zooniverse suite of projects has found success in effectively combing through massive amounts of data, but just as importantly, creating a community around their work. Boasting nearly half a million users, "Citizen Scientists" have been contributing to a variety of crowdsourcing projects since 2007. Most of the projects have revolved around astronomical research, but in 2010 they branched into historical climate data with Old Weather[2]. Utilizing the

lessons learned from other successful projects, Old Weather was able to attract a faithful following of contributors, but also found that there was a lot of incidental social capital generated from user engagement in the project. Giving users access to a forum to discuss their discoveries from the logs they are transcribing, for instance, led to volunteer-initiated research. But their findings, in line with other projects across sectors like advertising and media, suggest that a core audience will make up the majority of contributors. In this case, they found that while their sites have attracted hundreds of thousands of visits, less than five per cent of visitors participate in the forums.[3]

Gaming Applied to Scientific Research. A recent headline about the Foldit initiative, a game program that enables players to analyze proteins (more fun than it sounds), read: *"Computer Gamers Solve Problem in AIDS Research."* With a headline like that[4], who wouldn't be interested in utilizing "gamification" to engage the public in their research? In this much publicized case[5], the use of a multiplayer online game was used to generate models that outperformed computer models. Further research has been conducted to examine the specifics of the gaming environment that assisted in this research, with the ultimate conclusion that "online scientific game frameworks have the potential not only to solve hard scientific problems, but also to discover and formalize effective new strategies and algorithms."[6] We've yet to see game platforms created for such in depth analysis of humanities data and intend to test whether this form of engagement will enhance the attention and quality of crowdsourced humanities research.

Analyzing the Science of Crowdsourcing. Research on the science of crowdsourcing and its use and methods for various applications have done a great deal to analyze various elements that may optimize results and engagement. The bulk of this research has been conducted within the discipline of computer sciences and while this research will inform this study, we are focused less on the scientific or statistical improvement of crowdsourcing techniques and more on the application of these techniques specifically to humanities research. Nonetheless, work in the field will inform our study, such as the 2010 research of Jeffrey Heer and Michael Bostock of Stanford University on crowdsourcing graphical perception.[7] There is far less academic research available on how the technology and game industry has successfully capitalized on crowdsourcing elements, yet this research may be even more valuable to mine data on effective tools that can be adapted for our purposes. The game industry does not tend to do a lot of academic analysis of mechanisms, instead preferring "how-to" articles about specific technical problems. Here, we plan to rely heavily on our industry advisors and glean information from sources like the *Game Programming Gems* book series and the Postmortem feature of game developer magazine, Gamasutra[8]. One of our goals of this project is to take lessons learned from the game developer world, apply it to crowdsourcing tools and user interface, and bring it to bear in the academic discourse.

Crowdsourcing in Libraries, Archives, and Museums. The popularity of Zooniverse and projects like Flickr Commons has lead to several efforts worldwide to further leverage crowdsourcing, and there has been a growing body of research to assess the elements contributing to successful projects, such as Rose Holley's work analyzing crowdsourcing efforts at the National Library of Australia[9]. Oftentimes in libraries, archives, and museums, the crowdsourcing tasks solicited have the

immediate end of transcribing and increasing discoverability, such as the New York Public Library's "What's on the Menu" project[10], though our study will be examining the possibility of crowdsourcing contributions to aid specific research questions.

Toolsets for Crowdsourcing. Several remarkable open source crowdsourcing tools have been created to help enable crowdsourcing in the humanities, which we hope to leverage and contribute to in order to create the various crowdsourcing tools needed for this study. Map Warper[11] from New York Public Library has been widely used to help rectify maps in their collections, enabling the public to help align historic maps with geographic information necessary to create tiles and overlays on geospatial information systems. To this end, we will look at four different transcription tools that are available as open source, including Scripto[12] (Center for History and New Media), FromThePage[13] (Ben Brumfield), Transcribe Bentham Transcription Desk[14] (University College London), and Scribe[15] (Zooniverse). Finally, we plan to work with a number of freely available commercial APIs to geotag and spatially locate historical photographs, such as Google Maps API and Google Street View API.

Local Community Efforts. Not all crowdsourcing successes have been carefully plotted by funded projects, and some have seen incredible success as side projects that went viral, or late night coding binges. Two recent projects in particular are worth noting, and feed into our engagement strategy and planning. In 2010, Dane Pieri, then an undergraduate at Carnegie Mellon University, built a fantastic crowdsourcing interface called Retrographer[16] for geolocating historic photographs from the City Photographer Collection hosted and curated by the University of Pittsburgh's Archives Service Center. Mr. Pieri came up with a project that struck a chord, and the graceful user interface design and low barrier to participation resulted in the successful geotagging of over 4,000 historical photographs of Pittsburgh. In the summer of 2011, the San Francisco History Center and David Rumsey Map Collection teamed up to release high resolution color images of the 1905 San Francisco Sanborn Insurance Atlas. A unique and ragtag team of cartographers, user interface designers, and local history enthusiasts put together Maptcha[17], a tool to crowdsource the alignment of the historic maps to current streets (similar to Map Warper). In a number of days, the team had come up with a great user interface and word spread quickly amongst technology savvy history enthusiasts in San Francisco. The local public radio station picked up the story and in the course of about two weeks, four hundred users had aligned over 600 of the 700 pages. Finding and tapping into these kinds of local or topical communities is a critical aspect to successful crowdsourcing endeavors.

Crowdsourcing Research Projects (Methods and Interfaces)

Our design and development team will work closely with our humanities researchers to carefully consider the academic research questions, identify the various appropriate methods of human judgments, and design interactive and intuitive web interfaces for crowdsourcing. At the same time, the public engagement strategy and implementation will face the challenge of attracting enough contributors to make meaningful progress and also building community around the three projects.

Publicity and Community Engagement

We will create a Citizen History portal on Historypin which will highlight the three projects available for user participation, all with distinct narratives and design aimed to attract users with different motives. This strategy allows us to bring attention to the overall project while playing on the distinctive character of each of the projects. Carefully planned publicity around this portal, prominent links from the Historypin front page, and social networking queues will help attract existing Historypin users and new visitors to the project. Historypin will build announcements of the Citizen History project into press releases and media outlets managed by our New York publicity firm, Sunshine and Sachs.

Each of the projects will also have specific social networking and outreach strategies designed around them, building community and providing transparent process and progress of the academic research the crowdsourcing efforts are assisting. In addition, each project will leverage its existing resources and connections, utilizing the publicity and advertising associated with major book publishing houses (ie. for the Western Railroads Project, we may leverage the Railroaded website, a companion to Richard White's best-selling book published by Norton) as well as the publicity afforded to projects through Stanford University's media relations (alumni magazine, campus website, media placements generated by Stanford media relations via our contact, Shelly Goldman). The social networking and outreach strategies will be defined for each of the projects. Tagging 500 Novels will target online book groups and focus around a handful of select books initially. The Year of the Bay will seek to connect to influential bloggers in the Bay Area who focus, at least in part, on historical subjects. Finally, the Western Railroads Project will be directed primarily toward railroad enthusiasts and the range of technical abilities within that community.

In addition to publicity designed to reach a broad general audience, the primary focus of our outreach campaigns will be niche audiences specific to the content. Each of the three crowdsourcing initiatives will include an engagement sequence focused on these niche audiences. While each target audience is different, the steps of the engagement sequence are the same.

1. Identify the target audience. These may be train enthusiasts for the Western Railroads Project, Bay Area history buffs for Year of the Bay, or avid readers for Tagging 500 Novels.
2. Find existing online communities. Active online communities exist for each of the three initiatives, with various barriers to entry.
3. Find champions. Outreach to each of the communities will be most successful if championed by participants that hold sway in the group. Engaging them in the project early on will help leverage their voice and support.
4. Feedback loop. Early engagement of crowdsourcing audiences will focus on simple tasks with immediate feedback to encourage a sense of accomplishment, and instantly visible progress on the project. While tasks grow in complexity, our assumption is that smaller numbers of participants will advance to complete the tasks.

    a. The use of a global account, similar to Microsoft's Xbox Live Gamer Tag system, gives users a direct and measurable sense of progress. Every time they complete a task, they are awarded points.
5. Encourage sharing. Making it easy to encourage sharing at regular intervals helps spread word of the project through peer groups and reach increasing numbers.
    a. By allowing players to push their accomplishments to Facebook, Twitter, and other social media, the project will draw in new users who are interested in the specific subject material.
    b. A key element of this sharing is that it does not become a source of spam for a user's social circle. Every time a user chooses to share their progress and accomplishments, it will be a meaningful exchange.
6. Engage, promote, and reward power contributors. Studies on public crowdsourcing all show that a small number of users do the vast majority of work. Once we're able to reach and involve those users, it's imperative to continue to engage them.
    a. The global account system gives power contributors a way to compare their contributions against other users, thereby allowing a soft form of competition. A simple leaderboard, showing the users with the highest point totals, will generate additional interest and retention among power users.

Based on the success of "citizen science" and various other projects (Zooniverse, Maptcha, etc), and existing literature on crowdsourcing efforts (Romeo and Blaser, 2011; Holley, 2010), it will be critical to 1) be clear that the work is contributing to real research efforts; 2) provide immediate feedback that encourages additional input; 3) ensure that professional researchers are interacting with volunteers; and 4) share analyzed data and ongoing findings with the community as often as is possible.

Iterative Design and User Feedback

User interface and user interaction design frequently turns our initial assumptions about website use and navigation on their heads. Looking at the website from the perspective of the users, we first work to identify our prospective user communities and try to understand what they are getting out of the microsites, and review and test this data on an ongoing basis. We will use optional and occasional surveys to learn more about our user demographics and motivations. Utilizing user interaction design research, we will focus on building interfaces that are intuitive, provide rewards and incentives, and allow participants to share their contributions with friends on their own social networks. Additionally, our interface designs will allow for tests of the accuracy of crowdsourced inputs and will provide built-in means to assess data quality—which may include forced-agreement schemes, double-key entry, and machine learning techniques.

Engagement Strategy

Creating and documenting a successful engagement strategy is a critical component of this study. As noted above, we will seek participants able to move toward increasingly complex tasks. Each of

the three humanities projects will focus on a different kind of user, with varied interests, and the strategy of engagement will necessarily be slightly different. However, the process of moving them through a funnel of engagement will be the same. Initially, we will cast as wide a net as possible, and design initial tasks that can be completed quickly and easily to draw in a user. As users complete tasks and reach clear goals, they may then have an option to take on increasingly complex tasks.

Technical Development and Data Architecture

Crowdsourcing for Humanities Research represents an unprecedented collaboration between technology industry professionals, developers, and scholarly humanities researchers to find the best possible ways to engage the general public in crowdsourcing projects.

From a technological standpoint, the project will be focused on developing a reusable modular Graphical User Interface (GUI) to a web application that enables users to accomplish simple tasks such as refining and editing metadata, transcribing text, geolocating photos with Google Maps APIs, and identifying places, people or objects in photos. The GUI will be used for each of the three projects, though the graphic appearance can be different and the GUI will utilize different kinds of inputs (i.e. photos, text, maps, etc) for the different projects. Working closely with the Industry Advisors on the design of prototypes, the GUI will explore the use of gaming incentives like leaderboards and rewards, and leverage game industry technical design. The web application will create or largely draw upon open source libraries for transcription, geotagging, and other metadata refinements. (see "Toolsets for Crowdsourcing," above). The codebase for the GUI will be made available on Github with a GNU General Public License and design templates will be made available with a Creative Commons By Attribution license.

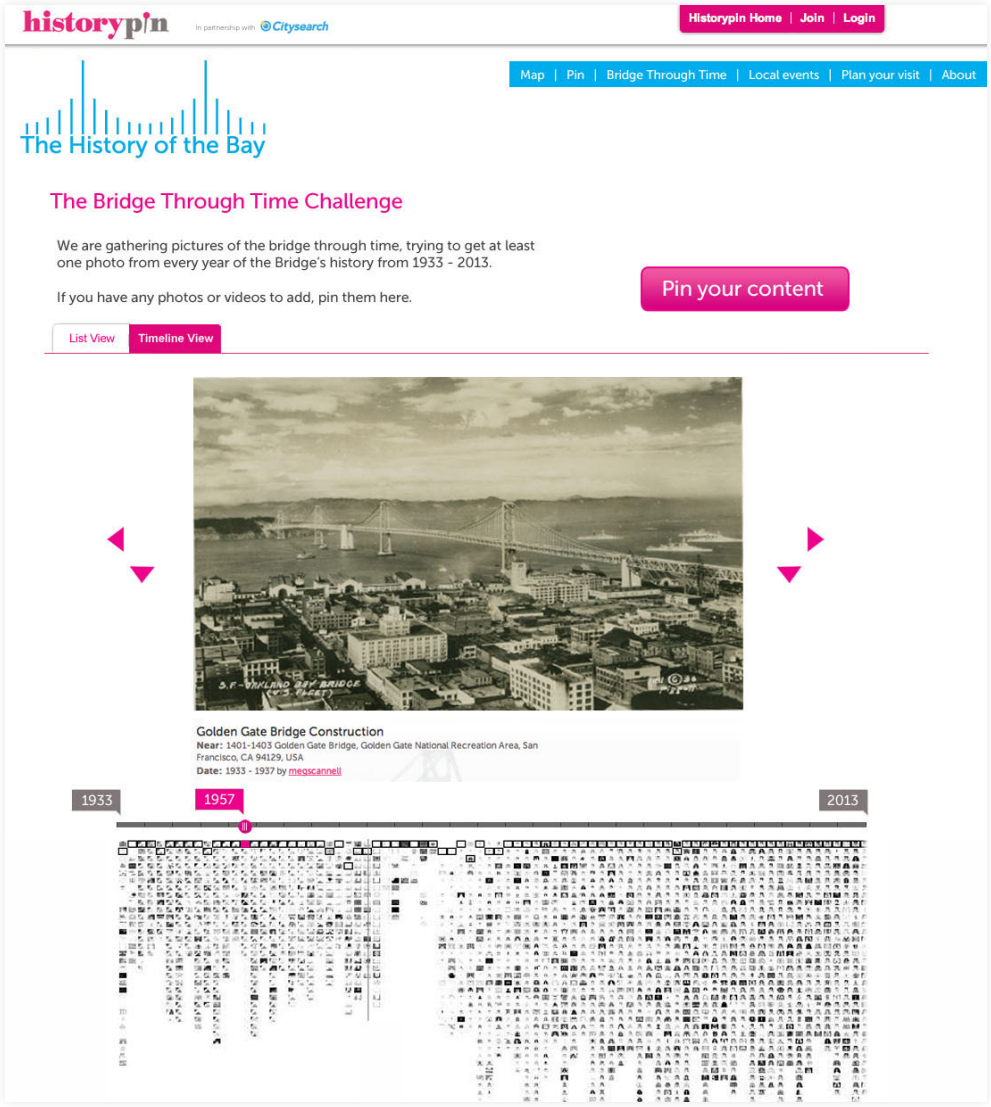
Additionally, existing Historypin tools and the learnings from the Historypin development, such as the mobile apps, and website interfaces or best practices, may also be utilized to support each of the projects. The Historypin team will be responsible for technical development, including website and toolset design, testing, and development. The Historypin platform utilizes the Python programming language and is built on Google App Engine, which will supply the common platform for each of the toolsets. The general technical approach, outline of data architecture, and a framework for technical development is outlined below.

How the Interfaces Apply to Each of the Projects
Each of the three projects will share the same technical design, overall technology approach and underlying data architecture. However, each of the projects will have unique graphic design and user interfaces that will rely on different source inputs and code libraries.

Tagging 500 Novels will revolve around text transcription and visualization of social networks within the novels. The interface may explore elements of contextual reading and one of the main user interface questions here will be how much text is too much or too little to allow for detailed reading and meaningful input. How much text will people read between contributions?

The Western Railroads Project aims to reach an audience of knowledgeable enthusiasts, so will likely test a rather rapid progression from easy to more difficult tasks. The primary source inputs will be railroad maps, in addition to geolocating photographs specific to railroads. Here again, timed specific community challenges may be an effective way to document and measure urban growth around railroads.

The Year of the Bay interface will revolve around a particular campaign focused on the Bay Area, so here, specific, timed challenges may play a prominent role. For instance, the figure on page 27 shows how an interface soliciting the public to contribute pictures of the Bay Bridge or Golden Gate Bridge for each year that they have been open, or correctly geotagging photos for each year from various contributing collections, might be configured. There may also be a georectifying challenge that draws users into the many maps of the Bay spanning cartographic records. The feedback cycle will be critical in this interface, showing increasingly populated timelines or map showing pinned photos, in addition to the leaderboards and progression bars.



Overall Technology Approach

Services-based: The core idea behind the *Crowdsourcing for Humanities Research* infrastructure is that of the web service. Whenever possible, components of functionality will be made into web services, with clearly defined interfaces. This approach gives the framework three key strengths: simplicity, isolation, and scalability.

Simplicity: A major risk in developing a framework for a variety of potential projects is that of over-development. Rather than building a solution for a single purpose, development time extends to allow for the creation of an over-engineered core codebase which can handle any eventual demand placed upon it. By focusing on small and modular services, it allows for quick and rapid deployment of individual chunks of functionality. This focus on rapid deployment also means that

development can occur in short bursts which result in noticeable functional improvements, rather than multi-month development cycles that result in messy code and untestable features.

Isolation: One of the key aspects of building a service-based infrastructure is the inherent isolation of the various components. If the server for project A crashes, that crash will not affect any of the other projects, nor the shared systems. Individual servers can be brought down for maintenance without affecting the other systems.

The other nice thing about the isolation of the different services is that they can be written in entirely different programming languages on radically different hardware. Since the only thing that matters to the other services is the programming interface and the format which data is transferred, a PHP server on Amazon hardware can talk to a node.js server running on a MacBook in a dorm room with no problems.

Scalable: A services model also allows for as needed scaling based on user load. Since all of the systems communicate through standardized interfaces. This means that the hardware and code behind that interface can be easily swapped without requiring any code changes by the other system. This decoupling means that data storage can start as a simple mysql database and evolve over time into a multi-tiered sharded data store utilizing load balancers.

Metrics and Analytics

One of the key changes in game development in the last two years is the introduction of metrics and analytics. Following user interaction and behavior allows us to discover important elements that will feed back into user interface design improvements. There are two aspects to metrics: collection and analysis.

Collection of usage logs from each project will enable us to measure usage. Every time a participant takes a significant action, such as inviting a friend, posting to twitter through the application, uploading a photo, etc, the project server will log their user ID along with the salient points of the action they took. Every link that we post to Facebook, Twitter, and elsewhere can embed a tracking number. This tracking number allows us to determine the source of new users, enabling acquisition tracking and analysis.

By creating this trail of virtual breadcrumbs, we can reconstruct the key events in a user's activity over the course of a day, week or month. While the individual data about actions being logged is rather trivial, it is through the comprehensive study of all these actions that second and third level analysis can be performed.

Analysis of the metrics can be done offline and will not affect the live projects. There are many interesting metrics that can be pulled from these metrics. Some examples include:
- Daily users
- Time of day when most users are online
- Average length of session

- Where new users are coming from (Facebook, Google, Twitter, Ads, Forums)
- What type of content users are interacting with, and for how long

As new projects are added to the suite of products, their metrics can go through the same metrics collection system, allowing for analysis to be run over all the projects in the *Crowdsourcing for Humanities Research* suite.

Once data has entered the metrics system, it can be normalized and exported to a variety of other systems. So this system can support both a set of researchers who wish to use Excel on comma separated text files as well as a group of analysts that want to run queries against a SQL database.

For specifically scaling the metrics system, since it is provided as a service to the projects, the back end can be rapidly scaled to handle increased load. Initially, the system will simply deposit write-requests to a buffer that a process will pull from and write to an Amazon data store. With increased load, the queuing mechanism will be switched to ZeroMQ and storage will transfer to a no-SQL solution, to allow for flexible data formats from a variety of product services.

Specific Analysis of Metrics

Crowd productivity will be measured in two key ways, engagement and retention. Engagement is the study of how active users are in the project while retention is the measurement of the duration that users are taking actions in the project.

Engagement:

* DAU - daily active users
* MAU - monthly active users
* Daily actions per user - how many 'actions' are users taking in the project on a daily basis
* Action distribution graph - Some users will perform hundreds of actions a day, many more will only do a single action. Seeing the distribution of those numbers on a daily/weekly basis can help see how engaged user are with the project.

Retention

* Cohort weekly retention - for a group of people who signup, how many stick around week after week
* Virality - for each user, how many viral channels do they use? Are they posting to Facebook or Twitter about the project? Are they inviting other people to join? For the people they invite to join, are they sticking around?
* Player lifetime - How long do players stick around once they start?

Acquisition channel effectiveness:

All users come from a source (Facebook, referrals, Twitter, email campaigns, specific posts on community forums, etc). By tracking the source of each user, we can judge the productivity of

various acquisition mechanisms. It may be that users that come from Twitter tend to invite a lot of new people to the project, but users that come from specific topic-specific forums are actually far more engaged and active with the project for a longer duration.

By measuring the different types of effectiveness of given acquisition channels, it will allow each of the projects to better target future acquisition efforts.

Means of analysis

All of the above analysis can be performed with simple formulas in Excel. The difficult part in performing the analysis is on the data acquisition and collation. Once the datasets are prepared, then analysis can be done in a relatively simple process. All this data is logged by the account system and then exportable into a variety of formats. The vast majority of this analysis will be done by Stanford faculty and Historypin staff, with leadership from the Industry Advisory Panel and limited support from independent contractors. Output from this research will be fed to and inform independent evaluators in their reports, surveys and baseline measurements of engagement of participants, as well as third party user interaction research (as described in the Evaluation section on page 49).

Underlying Data Architecture

Google App Engine is a cloud computing platform as a service (PaaS) for developing and hosting web applications in Google-managed data centers. It virtualizes applications across multiple servers. App Engine offers automatic scaling for web applications - as the number of requests increases for an application, App Engine automatically allocates more resources for the web application to handle the additional demand.

The platform we will develop on is Python with a combination of the Google App Engine and Django frameworks. The data store will be scalable using Google App Engine's Datastore. The Datastore has a SQL-like syntax called "GQL" designed to be efficient in a virtualized hosting environment. Switching from a relational database to the Datastore requires a paradigm shift for developers when modeling their data. Unlike a relational database, the Datastore API is not relational in the SQL sense, and is an important architectural distinction.

To provide a consistent feel across all of the projects, any gaming element implemented will span all projects. The approach mirrors how Xbox Live players accumulate gamer points by playing individual games. For this effort, each project will be the equivalent of an Xbox game. As the user completes tasks, they are awarded points. These points are accumulated in an account that spans all the projects. As users complete tasks, they will be granted points for their global account. This global account will be visible in the chat forums and other online communities.

To support this cross-project “identity”, a central account server will handle authentication and recording the accumulation of points. Each project will have access to an authentication API and a point API. As a user complete tasks for project A, project A’s server will inform the shared account

system that the user has earned X points. The next time the user logs into any project, they will see their point total increased by X points.

Each project will have a limit of the number of points it can award. Each project determines how it would like to award these points. The reason for a limit on the maximum number of points per project is to encourage users to find other projects where they can earn additional points. The account server will authenticate the points allocated to users and prevent non-authorized sources from awarding points.

Framework for Technical Development:

Transcription

Infrastructure

- Likely to utilize less scalable platforms such as PHP/MySQL for preliminary transcription development here, with the potential of using the learnings whilst developing scalable tools on App Engine (python/Big Table).
- We will choose one opensource toolset to work from and modify as necessary to fit both the user interface and technical requirements of the specific application. We cannot commit to the specific tools until we have hired a developer and had time to examine the codebase of the available options in order to identify the best suitable for the various projects.
- Cloud-based servers and databases will allow for the project to start small, without spending huge resources on infrastructure, yet provide easy avenues for rapid growth.

Functionality

- Points awarded upon successfully providing verifiable information.
- Discussion functionality on specific items.
- Allow users to incrementally improve meta-data. Users can build upon each other's knowledge.
    - Each contribution will be logged and points awarded as appropriate
- A versioning system will be implemented (or explored for that matter) to allow incorrect information to be flagged and rolled back upon.
    - Each award of points or other changes to a user's account will be logged, indicating the source of the change and the details of the change. This will allow administrators to roll back any specific change.
    - Should the need arise, these logs can be used to rebuild the entirety of the user database.
    - The log system can also be mined to perform analysis of how users are engaging with the various projects, providing the raw data for a metrics and analytics system.
- A notification system when something a user is involved in is updated or commented upon.
- A hierarchical scheme to provide differing permission levels to users based on status.

- Administrative interface to allow monitoring of progress and other critical stats to dev team and humanities researchers.

Output

- Likely to include forks from existing transcription toolsets

The transcription tools and technical infrastructure will be used primarily with the 500 Novels project. However, the basic toolkit will be designed in such a way as to encompass tasks that might be developed for the other two projects. These might include the creation of connections across corporate boards (railroad project) or the tagging of co-presence of places and words regarding environmental conditions for semantic analysis (Year of the Bay).

Geotagging

Infrastructure

- Adapting code frequently on availability of new developments in open APIs especially Google Maps (potentially Google's Natural language processing resources as APIs become available)
- Possibly rapid prototyping using less scalable platforms such as PHP/MySQL. Eventually using the learnings whilst developing scalable tools on App Engine (python/Big Table).

Functionality

- Possibility for open call for submissions from those who aren't necessarily experts in the field to provide some of the source material. Users don't need to join the crowdsourcing effort to provide their original content but will be made aware that the option of improving their content is available.
- Points awarded upon successfully providing verifiable information.
- Discussion functionality on specific items.
- Allow users to incrementally improve meta-data. Users can build upon each other's knowledge.
    - Each contribution will be logged and points awarded as appropriate
- A versioning system will be implemented (or explored for that matter) to allow incorrect information to be flagged and rolled back upon.
- A notification system when something a user is involved in is updated or commented upon.
- A hierarchical scheme to provide differing permission levels to users based on status.
- Administrative interface to allow monitoring of progress and other critical stats to dev team and humanities researchers.

Output

- Toolsets that can improve any media based archival content that requires description (time, place, description etc). Fields can be added or removed as appropriate to the subject matter (by modifying the code). A strong reusable element will be the geolocation of information, including Street View for orientation.

Geotagging will be utilized extensively in both the Year of the Bay and railroads projects. There is also a possibility of extending geotagging to the place names indicated in the 500 novels project.

User Acquisition

User acquisition through Social hooks: Facebook and Twitter are incredibly powerful tools for reaching a broad group of likeminded people in a short amount of time. Part of the engineering effort will be devoted to creating entry points into the applications from both Facebook and Twitter. While integration requires minimal engineering effort, the tracking url flags associated with each link need to be thought through and tracked. By providing the users cut-and-paste ready blocks of html, *Crowdsourcing for Humanities Research* makes it easy for involved users to contact people not just on Facebook and Twitter, but also subject matter-dedicated web forums.

Data Quality

Since the vast majority of the project's data content will come from users, it does raise the issue of how we maintain a high quality of content in the project. There are two sources of bad data, unintentional bad data and intentional malicious users.

Unintentional bad data comes from users who may not be subject matter experts who in their excitement to participate start pushing bad data into the project. User interface design will focus on first time prompts to orient participants to the project and also start with easier tasks and moving to more complex tasks. Additionally, user contributions of content will be processed through the existing Historypin moderation process, which runs all contributions through an admin interface that is monitored by Historypin staff for inappropriate, spam, or advertising content. Additionally, data contributions can be verified with multiple human judgments, or with administrator approval.

We do not anticipate intentional malicious users to be a problem, based on the data of similar crowdsourcing projects in the humanities (Romeo & Blaser, 2011). In the game industry, this problem is known as "griefing," in which players are motivated to play a game not from gameplay, but instead in annoying and harassing other players. There are methods for grief prevention, but we do not plan to allocate resources to this type of programming unless required.

To conclude this section, we emphasize that our plans for technical development and interface design are meant to be flexible and responsive to feedback. We expect to fine-tune our approach and develop additional tools and interfaces as we learn by doing.

III
Findings and Dissemination: From Crowdsourced Inputs to Scholarly Outputs

Results from the collection and analysis of crowdsourced information will be presented in traditional academic forums, such as peer-reviewed journals, edited books, and scholarly conferences, as well as in online products. Some of the specific publication plans for the projects are included in the project description section of this narrative, such as the online repository of 500 tagged novels; other publications will, of course, evolve from the ongoing give-and-take of the research process over the course of the grant period. In addition, findings will be presented in the form of feedback to crowdsourced contributors—providing up-to-the-minute information regarding the kinds of interpretive analysis conducted on the data, preliminary results and interpretations, and summary statistics and visualizations of data allowing the public to chart the progress of the research as it unfolds.

Wherever possible, findings published in the traditional fora of academia will include explicit analysis of the methods of crowdsourcing employed in the research and critical evaluations of the efficiency and accuracy of the processes involved in finding the crowd, engaging its interest, and harvesting its assistance. In this way, we hope to render these processes as transparent as possible, allowing other researchers to peer under the hood, as it were, and see for themselves how the information was collected and processed. This mode of dissemination will serve as further inducement for scholarship that is open, distributed, collaborative, and transparent—hallmarks we believe of the future of humanities scholarship inasmuch as it partakes of the affordances of new digital media and the distributed labor and expertise available via the internet.

Deliverables
Peer-reviewed scholarly articles, including a multi-author article reviewing the comparative findings of this experiment in crowd-sourcing across three different projects, including the principal investigators in each of the projects, with Zephyr Frank as the lead author; at least one article from each of the projects on the findings relevant to that particular scholarly field in an appropriate journal, with lead authors Franco Moretti, Richard White, and Jon Christensen.
A synthetic, comprehensive, detailed, multi-author white paper with lead authors Jon Voss and Zephyr Frank on the technical and social best-practices in crowdsourcing methods developed and tested in this project, including sufficient information and resources so that readers could replicate the process in their own projects.

We've budgeted for a dissemination strategy that supports publication and distribution of findings in print and online, as well as travel stipends to present findings at prominent conferences (two international and four in the United States). Domestic meetings may include the Modern Language Association, the American History Association, Museums and the Web, Museum Computer

Network, American Library Association Annual Meeting; international meetings may include the International Digital Humanities Conference.

Our goal is to submit a paper to a peer-reviewed journal for each of the three projects, as well as self-publishing a white paper of our complete findings, which will be available in limited print editions and on the web, both licensed with CC-BY licensing.

Beyond this, we expect to produce databases of project inputs and outputs freely available to the public via simple interfaces, including the 500 novels database; a database of nationwide railroad photographs and maps; and a regional database of environmental history sources for the Bay Area.

Findings and Dissemination: Interfaces and Community Engagement

By the end of this project, we will have a well-documented analysis of micro-volunteering, crowdsourcing, and game mechanisms used to collect and refine data, engage various demographics, and contribute to the body of humanities research. This research and analysis will be critical in informing future opportunities and prioritizing the most likely to succeed projects for this type of strategy.

Documenting and Sharing the Process
Throughout the project, researchers will maintain a public blog documenting every phase of the work, examining research questions and documenting our methodology and decision points as the project progresses. Using written entries as well as video documentation, we will demonstrate free or low-cost tools that we use to build community, engage participants through social networking, conduct user interaction testing, and manage and harvest data.

Documenting and Sharing the Tools
We intend to leverage open source crowdsourcing tools wherever possible (highlighted in Related Work), sharing our contributions to the code-base with the same open license. The primary technical tool being created for *Crowdsourcing for Humanities Research* is a Graphical User Interface which will facilitate various discreet crowdsourcing tasks. The codebase for the GUI will be made available on Github with a GNU General Public License and design templates will be made available with a Creative Commons By Attribution license.

Publishing and Sharing the Historical Data
The various datasets that will be utilized for this project may have digital assets and metadata with different licenses. As described in detail below, we may utilize images with protected copyright status within the Historypin domain, however, we will require that all metadata be contributed to the project with an open license, namely CC-BY, CC0, or Public Domain. We will make the metadata available under a CC-BY license in a variety of ways, including at least csv files on the project blog. The datasets include photographs and maps of the San Francisco Bay Area; novels; and railroad maps, photos, and data extracted from the maps and photos.

Publishing and Disseminating the Research Data
We intend to collect as much data as possible about how users actually use the tools created for *Crowdsourcing for Humanities Research*, which may include information like how long they spent on the sites, where were they accessing it from geographically, etc. This data will be published at the end of the project under a CC0 license.

Publishing and Disseminating the Research Findings
We've budgeted for an aggressive dissemination plan to share the discoveries and tools for crowdsourcing in the humanities. We will focus on quantitative evidence, as well as step by step how-to's and access to free or inexpensive tools to be used in the course of a project.

Reporting
Our project has been conceived to provide for high levels of accountability and assessment. In the two preceding sections on dissemination of project results, we contemplate the publication of white papers, tools, and databases as public reporting conducted throughout the period of the grant. Beyond this, our project will provide annual reports to the Foundation, with clear accounting of the work accomplished, challenges faced and specific uses to which funds have been put as detailed in the Foundation's financial reporting template. Our team is experienced in grant management and has a solid track record of efficient use of Mellon Foundation support in the past (White 2007-11).

The reports to the Foundation will include an executive summary (2 pages) of the work conducted over the previous 12-month cycle of funding. This will be followed by detailed reports regarding each of the three core projects contemplated in this proposal—each of these reports will reflect the rate of progress on the academic research itself as well as on the study of interfaces, methods, and crowdsourcing more generally. As per the Foundation's guidelines and according to the templates for narrative and financial reporting, annual reports will include detailed narrative and financial sections covering each of the expected areas of reporting including personnel, deliverables, publications, and plans for future work. Each report will be accompanied by a detailed budget, with data and comments regarding expenditures-to-date, expenditures during the yearly reporting period, and projected expenditures, broken down in each case by line item and individual project. The first report will be provided on September 1, 2013; the second on September 1, 2014; and the third on September 1, 2015.

Intellectual Property
Technology

We intend to leverage open source crowdsourcing tools wherever possible (highlighted in Related Work), sharing our contributions to the code-base with the same open license. Whenever we create new crowdsourcing tools, those will be published with open source licensing (GNU) on Github.

Content
Copyright of images, digitized documents, and literary texts contributed to Historypin for inclusion in this project will be determined by the contributing institutions. While we will be working largely with datasets already in the public domain, the terms and conditions of the Historypin website permit the use of all contributed assets within the confines of the Historypin website and mobile apps, so we are not limited to using images in the public domain. Each project leader will take responsibility for working carefully with institutional partners and Historypin to ensure that their contributions are properly credited, do not infringe copyright, and carry the appropriate license for further use, if appropriate. When individuals contribute materials, we will make it easy for contributors to specify the license they want the material to carry. We will also provide mechanisms for unauthorized, unlicensed material to be tagged and removed from sites (see Appendix 4).

Historypin holds no copyright on contributed content and institutions and individuals alike can choose from the following options:
All rights reserved (full copyright)
No known copyright restrictions
Public Domain Mark
CC0
CC BY
CC BY-SA
CC BY-NC-SA
CC BY-ND
CC BY-NC-ND

We differentiate between assets (such as digital surrogates, scans, and photographs) and metadata, both of which can have separate licenses. For purposes of this project, all metadata contributed must have an open license ascribed to it (CC0 or PDM) in order to be accepted. By March, 2012, basic metadata for content on Historypin is scheduled to be available as CC0 through JSON feeds, and later an API, though users will be able to opt out of this.

The project will work with content contributors, both individual and institutional, to educate them about their various licensing options and the importance and potential benefits of open data, including increased discoverability, greater use and reuse, increased web traffic, and raised awareness of collections and content. Industry Advisor Abigail Phillips will provide additional advice and support on IP issues.

In the case of user-contributed content, users will warrant ownership, permission, or license to post all contributions and will be subject to the published take-down policy of the Terms and Conditions posted on the Historypin site.

References for Narrative, Part 1

Elson, David K., Nicholas Dames, Kathleen R. McKeown. 2010. "Extracting Social Networks from Literary Fiction," *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (ACL 2010). Uppsala, Sweden. <http://www.cs.columbia.edu/~delson/pubs/ACL2010-ElsonDamesMcKeown.pdf>.

Frank, Zephyr. "Layers, Flows and Intersections: Jeronymo José de Mello and Artisan Life in Rio de Janeiro, 1840s-1880s," *Journal of Social History* 41:2 (Winter 2007), 307-328.

Heer, Jeffrey and Michael Bostock. 2010. "Crowdsourcing graphical perception: using mechanical turk to assess visualization design," *Proceedings of the 28th international conference on Human factors in computing systems* (CHI '10). ACM, New York, NY, USA, 203-212.

Huberman, Bernardo, Daniel Romero, and Wu Fang, "Crowdsourcing, attention and productivity," *Journal of Information Science*, 35 (6) 2009, 758–765.

Jockers, Matthew. "A comparative study of machine learning methods for authorship attribution," *Literary and Linguistic Computing,* Volume 25, Issue 2 (2010), 215-223.

_____. *Macroanalysis: Methods for Digital Literary History*. University of Illinois Press, under contract (expected publication in 2012).

_____. "Reassessing authorship of the Book of Mormon using delta and nearest shrunken centroid classification," *Literary and Linguistic Computing*, Volume 23, Issue 4 (2008), 465-491.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. "Crowdsourcing user studies with Mechanical Turk," *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (CHI 2008). ACM, New York, NY, USA, 453-456.

W. A. Mason and D. J. Watts. Financial incentives and the "performance of crowds". In KDD-HCOMP, 2009

Moretti, Franco. *Atlas of the European Novel, 1800-1900* (London: Verso, 1998).

_____. *Graphs, Maps, Trees: Abstract Models for a Literary History* (London: Verso, 2005).

_____. "Network Theory, Plot Analysis," *New Left Review*, Issue 68 (March-April 2011), 80-102.

White, R. *Railroaded: The Transcontinentals and the Making of Modern America* (New York: Norton, 2011).

Endnotes for Narrative, Part 2

[1] http://earth.google.com/outreach/program_details.html. Consulted February 11, 2012

[2] http://www.oldweather.org. Consulted November 10, 2011

[3] Romeo, F., and L. Blaser. Bringing Citizen Scientists and Historians Together. In J. Trant and D. Bearman (eds). *Museums and the Web 2011: Proceedings*. Toronto: Archives & Museum Informatics. Published March 31, 2011. Consulted November 10, 2011. http://conference.archimuse.com/mw2011/papers/bringing_citizen_scientists_historians_together

[4] http://blogs.discovermagazine.com/notrocketscience/2011/09/18/computer-gamers-solve-problem-in-aids-research-that-puzzled-scientists-for-years/. Consulted November 10, 2011.

[5] Firas Khatib, Frank DiMaio, Foldit Contenders Group, Foldit Void Crushers Group, Seth Cooper, Maciej Kazmierczyk, Miroslaw Gilski, Szymon Krzywda, Helena Zabranska, Iva Pichova, James Thompson, Zoran Popović, Mariusz Jaskolski & David Baker. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature Structural & Molecular Biology* **18**, 1175–1177 (2011) doi:10.1038/nsmb.2119. Published online 18 September 2011. Consulted November 10, 2011. http://www.nature.com/nsmb/journal/v18/n10/full/nsmb.2119.html

[6] Firas Khatib, Seth Cooper, Michael D. Tyka, Kefan Xu, Ilya Makedon, Zoran Popovic, David Baker, and Foldit Players. Algorithm discovery by protein folding game players. PNAS 2011 : 1115898108v1-5. Published November 7, 2011. Consulted November 10, 2011. http://www.pnas.org/content/early/2011/11/02/1115898108

[7] Jeffrey Heer and Michael Bostock. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. *ACM Human Factors in Computing Systems (CHI), 203–212, 2010.* http://vis.stanford.edu/papers/crowdsourcing-graphical-perception

[8] Gamasutra. http://www.gamasutra.com/features/postmortem/. Consulted February 3, 2012.

[9] Rose Holley. Crowdsourcing: How and Why Should Libraries Do It? *D-Lib Magazine, March/April 2010 Volume 16, Number 3/4. doi:10.1045/march2010-holley.* http://www.dlib.org/dlib/march10/holley/03holley.html

[10] http://menus.nypl.org/. Consulted November 10, 2011.

[11] http://mapwarper.net. Consulted November 10, 2011.

[12] http://scripto.org/. Consulted November 10, 2011.

[13] https://github.com/benwbrum/fromthepage/wiki. Consulted November 10, 2011.

[14] http://code.google.com/p/tb-transcription-desk/ and http://www.ucl.ac.uk/transcribe-bentham/. Consulted November 10, 2011.

[15] https://github.com/zooniverse/Scribe and http://oldweather.org. Consulted November 10, 2011.

[16] http://retrographer.org. Consulted November 10, 2011.

[17] http://sanborn.maptcha.org/. Consulted November 10, 2011.