Human Voices Project

Randall Cream, University of South Carolina, USA, Project Director, NEH
Michael Welge, SEASR, NCSA/University of Illinois, USA, Project Director, NSF
Paul Yachnin, McGill University, Canada, Project Director, SSHRC

The Human Voices Project proposes a unique response to the proliferation of data in humanities computing. Instead of adopting computational analysis as a means of reducing complexity, Human Voices resituates the heterogeneity of the humanities by re-activating the interconnected nature of humanities work. Our project aims to mine data not for reduction but for analytic amplification by developing the inherent interconnectedness of humanities work. Data mining applications inevitably rely on reductive algorithms that analyze texts by flattening variables, identifying patterns, and reducing the ontological complexity of concepts in order to allow orderly computation to occur. Such models yield precise and useful results and enable pattern matching on enormous scales. Human Voices rejects these approaches, since humanities data cannot be made more meaningful by reducing its complexity. The recent Netflix Prize competition illustrates this difficulty: one million dollars and a team of several Nobel laureates only generated a 10.05% efficiency increase in predictive searching—suggesting that we are in an era of diminishing returns to algorithmic searching using keywords.

Human Voices re-imagines humanities computing with the humans in it. Our approach to the challenge of the proliferation of data is to not reduce that data, but to make it more human by attending to its inherent complexity. Our project works simultaneously with three archives in the humanities: (1) the English Broadside Ballads Archive, housing 120,000 records of ballads written between 1500 and 1700; (2) the Spenser Archive, housing the works of the 16th-century poet Edmund Spenser, and (3) JSTOR, the largest archive of humanities and social science scholarship in the world, housing approximately 5 million journal articles. Using automated citation extraction to reveal the multiple "aboutness" references that distinguish humanities scholarship, Human Voices uses data mining to multiply complexity by revealing affiliation relationships that activate semantic networks inherent in the data. Our work proceeds in three phases and draws on our partners' complementary strengths: through automated citation extraction (Phase I), text segmentation into multiple overlapping units of semantic meaning using these citation units (Phase II), and semantic analysis through a process that flexibly attends to the multiplicity of layers and recursively weaves in ongoing usage of the data to refine meaning (Phase III).

Our team draws on the collaborative framework of a strong partnership with SEASR, a flexible platform for browser-based text and data analysis. Alongside the Making Publics project, a multi-institutional, multi-national team of researchers investigating public formation around works of art and intellect in the early modern period (1500-1700), Human Voices brings an innovative data mining component to this kind of humanities scholarship, allowing the underlying networks of affiliation, citation, and adaptation to emerge visually. The impact of our methodology—working with humanities material on a discrete, concept-level, thereby facilitating a massively self-proliferating network of affiliation relationships to consciously develop— is to enable a qualitatively different set of questions in the humanities to emerge. Rather than working *within* the text and then looking for connections to events *external* to those texts, Human Voices situates the primary and secondary materials of the humanities as participants in a series of overlapping conversations.

Our team is diverse, large, and multi-disciplinary, with private partnerships (participants from Google & Collexis) entering into partnerships with researchers from English, Computer Science, the Digital Humanities, History, Geography, and Art. On the US side, Randall Cream will serve as project director to a partnership on the NEH side of the project that includes a multidisciplinary team of researchers and a partnership with the private firm Collexis; the SC team will work closely with a team of researchers at the U of California at Santa Barbara, led by Project Director Patricia Fumerton. Also on the US side, Michael Welge will lead a team at the NCSA/UIUC developing SEASR. On the Canadian side, Paul Yachnin will serve as project director to a multidisciplinary team of researchers at McGill University, aided by a productive partnership with a Google software engineer.

Project Participants

Auvil, Loretta, National Center for Supercomputing Applications (NCSA), University of Illinois-UC

Buell, Duncan, Computer Science Engineering, University of South Carolina

Cream, Randall, Center for Digital Humanities, University of South Carolina, **US, Project Director, NEH**

Eberhart, Marlene, Making Publics Project, Department of English, McGill University

Folkerth, Wes, Making Publics Project, Department of English, McGill University

Fumerton, Patricia, EBBA, University of California-Santa Barbara **EBBA PI and UCSB Project Director**

Leicht, Stephen, Chief Operating Officer, Collexis, Inc, South Carolina

McAbee, Kris, EBBA, University of California-Santa Barbara

Matthews, Manton, Computer Science Engineering, University of South Carolina

Meunier, Jean-Guy, Philosophy, Université du Québec à Montréal

Miller, David Lee, Department of English, Center for Digital Humanities, and Spenser Archive, University of South Carolina

Milner, Matthew, Making Publics Project, McGill University

Nebeker, Eric, EBBA, University of California-Santa Barbara

Petric, Vlad, Software Engineer, Google, New York.

Sieber, Renee, Geography and Computer Science, McGill University

Stahmer, Carl, EBBA, University of California-Santa Barbara

Welge, Michael, Director, Automated Learning Group, NCSA/University of Illinois Urbana Champaign, SEASR, **US, Project Director, NSF**

Yachnin, Paul, Making Publics Project, Department of English, McGill University, **Canada, Project Director, SSHRC**

Zhou, Jun, Center for Digital Humanities, University of South Carolina

Human Voices Project

Randall Cream, University of South Carolina, Project Director, NEH
Patricia Fumerton, University of California, Santa Barbara, EBBA Project Director
Michael Welge, SEASR, NCSA/University of Illinois, Project Director, NSF
Paul Yachnin, McGill University, Project Director, SSHRC

The Human Voices project uses large-scale data mining to integrate humanities scholarship with primary texts in a flexible, re-configurable constellation of affiliation relationships. Using the common occurrence of inter-textual quotation and citation as a launching point to build dense nodes of affiliation, Human Voices resituates acts of scholarship into a multivoiced dialogue that allows multiple ambiguities of interpretation balanced by the intersubjective grounding of interpretable texts. Rather than subscribe to a notion of data mining as a practice that is inherently opposed to humanistic inquiry, Human Voices understands that the analytic methods of data mining already drive many of the familiar acts of humanities scholarship. The Human Voices project aims to displace serendipitous and accidental discovery as prime movers of humanistic interpretation. Our methodologies open up computer-assisted analytic interpretation to a broad array of scholars far removed from computationally intensive disciplines.

One key methodology within the Human Voices project segments the products of scholarship in the humanities by dividing the texts—articles in JSTOR, the largest archive of scholarship in the world—along the lines of citation and quotation. Our task—easy to describe, yet nevertheless a fairly dense computational endeavor—is to automate the identification, tagging, and segmentation of citations within the datasets. We aim to identify every sentence that cites an external work and insert the standard <cit /> tag to identify that citational act. This is a difficult enough task (see environmental scan, below) that we envision spending approximately half of the grant period refining our approach in ways that allow us to align our analysis with the vast contours of our dataset. It's easy to imagine the transformative potential of the results of such a machine-driven analysis. Articles that are now grouped by author, subject, and a few machine-extracted keywords, will instead be dynamically reconfigurable as ontological portfolios of a vast number of affiliation acts represented by citation and quotation. Instead of the paragraphs of an essay constituting an irreducible whole, they will also be able to enter into conversation with the paragraphs of the scholars and texts they cite, as well as the scholars and texts that cite them. Scholarship, thereby, becomes a human act of affiliation, and work in the humanities becomes more human centered. Or to put it more plainly, the voices of scholars occasion, through our research paradigm, a conversation that speaks across time and space to engage in sustained deliberation and dialogue.

Our project, then, contains a highly-feasible yet transformative data analysis phase, a recursive modeling phase to adapt that analytic to conform to the contours of the datasets, and, importantly, a visualization and human interface construction phase where we develop an approach to data mining that employs the uniquely human qualities of judgment, analytic insight, and creativity through a recursive research environment. Human Voices uses the **S**oftware **E**nvironment for the **A**dvancement of **S**cholarly **R**esearch (SEASR) in order to facilitate a recursive, human-centered approach to data mining that opens up data analysis to scholars who would never begin to look at code. Through a flexible platform that invites collaboration on analytic "flows" and visualization layers, SEASR allows the analytic might of data mining to operate through the browser window, opening almost any browser-readable data for analysis. Instead of asking users to take data to some external site for analysis, SEASR picks up data in the browser window and brings a fluidity to data mining by using a web services layer that interacts with remote servers through a dedicated port. Open source, community-centered, and aggressively engaged with active researchers throughout the humanities and information sciences, SEASR is a significant platform for conducting computational analysis.

At its core, our project consists of three computational tasks of increasing ambition and difficulty. The most feasible element of our workplan is to automate the identification of citation references from our datasets. Although this is a non-trivial act (see environment scan, below, for some of the issues associated with citation identification and extraction), we're confident we can model this identification recursively throughout the JSTOR dataset during the 10 months of project time devoted to this process. The second phase of our workplan involves using a special subset of citations, quotations and paraphrases, to create units within the largest dataset, JSTOR. Each of these units will consist of the quotation and/or paraphrase, the citation, and associated sentences that explain and develop the referenced point. We'll construct the frameworks that allow these citational units to be cross-referenced with their cited texts and the texts that cite them, creating highly-related, densely associated nodes that represent the elements of scholarship that are most valuable, most controversial, or most noteworthy. Our longer term goal, outside the scope of this brief 15 month project, is to conduct semantic analysis on these units, identified with concepts, and allow concept mapping to span across and between essays as an assemblage of citation units.
This three-phased approach represents a responsible, extensible, innovative approach to the uniquely scaled datasets of interdisciplinary scholarship and research. It links definable, benchmarkable outcomes to high risk:reward methodologies, providing a justifiable return on investment for a project of this size and scale.

Datasets

Our approach to data analysis in the humanities takes advantage of the enormous scale afforded by advances in storage density. JSTOR, the largest dataset of our four primary data models, is a five million article archive of scholarship in the humanities, sciences, and social sciences. Our three other datasets are much smaller by comparison, but together represent a significantly dense datapoint in humanities research. The English Broadside Ballad Archive (EBBA hereafter) is a culturally significant archive of approximately 5,000 ballads from the early modern period (1500 to 1700) in the UK. The Making Publics Project (MaPs hereafter) is home to a large-scale, long-term investigation into the heterogeneous acts of "public making" in early modern Europe (1500-1700). MaPs brings together scholars based in Canada, the US, and the UK, as well as a network of over a hundred associates and correspondents, working in an array of human-centered disciplines, from Art History, Musicology, History, and Literary Studies, concerned with mining cultural archives and documents in order to trace the development of groups of association, markets, media and their impact on the formation of notions of publicity. The Spenser Archive is an archive of the works of the 16th century poet Edmund Spenser, with stunningly high resolution images of 16th and 17th century editions and highly structured markup of the texts. Together, the three early modern archives create a densely concentrated dataset for information about and from the early modern period. Our datasets weave together two very different forms of literate arts--Spenser's epic poetry is at once quite similar to (in a formal sense) but also strikingly different from the ballads of the period, allowing the Human Voices project to interrogate very specific questions of aesthetics and semantic usage. By combining these two early modern archives with the MaPs project, which seeks to foster a collaborative exchange through social networking as a means of scholarship, Human Voices is able to develop a detailed and sustained inquiry into the poetics and politics of space in the early modern period.

Working with the primary archives, even with the additional layer of metadata from scholarly interaction provided by MaPs, is insufficient ground for the sorts of questions we want to facilitate for humanities researchers. Rather, using SEASR's browser-based platform, we want to perform analysis of the primary (historical) in tandem with the extremely large secondary dataset, JSTOR. The dense datapoints of early modern material should provide a sufficient node structure to build a useful affiliation network of contemporary scholarship. JSTOR is key to this project because of its size and density. The five million

articles in JSTOR represent one hundred and fifty years of scholarship, by no means a continuous and homogenous dataset. Unlike the early modern datasets, JSTOR contains texts from a variety of disciplines, from the sciences to the arts and humanities. Early scholarship bears little formal similarity to the models under current practice in the academy. The vast contents of the JSTOR archive represent a valuable opportunity to conduct model information retrieval within a dense yet varied dataset with somewhat limited or controlled vocabularies.

In structure, the JSTOR archive's data is dissimilar in important ways from the three other archives. The metadata of JSTOR is highly structured, well schematizing the bibliographic information about each article. The contents of the articles, though, are much less structured, generally marked up only by page tags to associate text with image. Processing the data of the JSTOR archive will more closely resemble working with unstructured text than mining the richly structured, highly tagged environment of the other archives. Our analysis tools, built from scratch using Python, Java, repurposed from existing code from NORA and the Monk workbench, and existing SEASR modules using D2K and UIMA, will readily adapt to both environments.

Working closely with the rich and diverse archives that constitute our dataset—EBBA, the Spenser Archive, and JSTOR—the MaPs and other project participants can draw on their long history of looking at inter-textual and inter-scholarly discourse of the early modern period. Our robust collaborative relationship with JSTOR especially gives Human Voices on the one hand an unprecedented ease of access to the data for modeling and testing and at the same time a vested interest in developing tools that are portable to a more restricted model and therefore available for general usage. While project testing will use local copies of all of the data, finished modules will use API and a webservices model to make calls against remote data, allowing user authentication, session denomination, and other interface and security layers to be deployed. The Human Voices project recognizes the tensions between responsible access and the economics of preservation, and is bound by its agreement to respect the intellectual property contained within the archives of the dataset.

Environmental Scan

The Human Voices project innovates not through the creation of never-before seen algorithms but by repurposing inherent capabilities from established areas of computing in new and exciting ways to produce transformative analytic results in the humanities. Our first phase, automated citation identification in academic scholarship, is a necessary pre-component to the analysis provided by our methods. Citation extraction is notoriously difficult to model due to the inherent ambiguities afforded by proper names of authors and journals. Just recently, in *D-Lib* 15.3 (2009), J.H. Canos et al propose adopting a unique identifier (DOI) in order to facilitate disambiguating citation references in scientific literature. Given the growth of extremely large datasets in scholarship (particularly in the sciences), ambiguity is a persistent and irreducible component of citation identification. Our chosen dataset of secondary scholarship, JSTOR, however, already employs a metadata field with a unique identifier tag, allowing precise references to build and inhere. We'll take advantage of this precision in our code, structuring relationships not between authors and texts but between unique identifier fields. The unique identifier tag of the JSTOR dataset allows us to access unique, 1:1 relationships of citations and disambiguate extremely common journal names and surnames with ease.

Methodologically, Human Voices draws from several existing digital humanities projects that work in areas of citation extraction and identification. Perhaps the best known of these is the open source citation project, *OpCit*. An extremely ambitious project whose agenda far exceed the technical capabilities of its time, *OpCit* worked from the 1990s until 2002 to produce a standard for identifying, using, and referring to citations in the emerging world of text markup and text analysis. Their team, funded through the JISC and the NSF and deeply influential, aimed to produce an open standard for archives, intellectual rights

holders, and content aggregators to facilitate interoperable search and extraction. Although dormant for a few years, the intellectual agenda, open source code, and methodologies of *OpCit* remain relevant for researchers such as the Human Voices project, Gregory Crane at Tufts, CiteSeer, and the OAI initiative.

Quotation identification is a subset of citation identification that is particularly significant because quotations create a logic of semantic association between units. As Ernst and Crane (2008) point out, there is little existing work that automates matching and identifying quotation and allusion between primary texts of a historical nature in cases where authors don't already structurally identify such affiliations. Ernst and Crane propose an algorithm for identifying patterns in texts as likely candidates for hidden instances of quotation and reference; we would welcome the chance to adapt their algorithms to our project to work with the historical archives of the early modern period. Due to the vagaries of our datasets, though, this algorithmic approach would largely function as a quality control assurance, since the historical archives we use (EBBA and Spenser) consist of highly structured texts where citation, paraphrase, allusion, and quotation are well understood and already tagged by experts in the discipline. Nevertheless, Ernst and Crane demonstrate the feasibility of an automated approach to textual citation beyond mere parenthetical reference and metadata extraction.

Identifying and extracting citations is one thing; using the presence of citations to generate concept maps for domains of knowledge is a much more difficult affair. One difficulty is the non-standardization of references between scholarly papers and venues. Andrew McCallum has worked extensively in this area, conducting feasibility tests that identified the difficulties of canonicalization, abbreviation, and other ambiguities in text citations (2007). Goldstone (2004), Bradshaw (2002, 2001), and Hammond (2002, 2001) have developed approaches to citation that link citations to indexing based on keywords. None of these approaches, however, uses citation relationships to create detailed models of domain knowledge. Instead, existing approaches seem to suggest that citations themselves tell a user surprisingly little about the propositional content of an article, since there are disciplinary biases towards citing specific authors and recent texts.

More recently, Fuzzy Association Concept Mapping (FACM) has been suggested as a promising model for developing concepts in texts processed by Natural Language Processing abstraction. Wang, Lee, Cheung and Kwok (2008) develop an information organization model that recursively asks the user to interact with automated concept maps, refining the map through human interaction. Similarly, Ritchie, Robertson, and Teufel (2008) suggest indexing articles using the sentences that surround citation points. These researchers indicate the potential of the surrounding text of what we term "citation units" as of particular semantic value. However, by focusing on keyword extraction, their methodology fails to overcome the limits of current approaches to keyword mapping and searching.

Our environmental scan reveals a sufficient number of research teams working in close parallel to our methods to suggest that our approach is worthy of pursuit. No published study aligns with our goals, though, in providing a reconfigurable network of affiliation relationships to visualize concept maps in the discipline of the humanities. We are cautiously ambitious that our approach will yield results of interest to these researchers and others pursuing similar lines of inquiry.

Project Members

Growing out of several projects in the digital humanities—MaPs, EBBA, Spenser Archive—the Human Voices project insists on a productive simultaneity between humanistic inquiry and computational data mining. The broad range of team members represents a sustained focus on delivering useful tools and methodologies that can serve researchers from a variety of fields in the humanities, social sciences, and sciences. Our project team has little duplication, instead bringing together stakeholders into a jointly beneficial mutual endeavor.

The team at the National Center for Supercomputing Applications and UIUC, led by Michael Welge, brings a demonstrated interest in algorithmic data mining to yield results that are useful to humans. Their SEASR platform, an open-source and community centered environment for scholarly research that uses a web services model as a browser plugin, is under widespread adoption in the humanities for its powerful flexibility in text and data analysis. The **a**utomated **l**earning **g**roup at NCSA, led by Loretta Auvil, is insistently collaborative, routinely partnering with external researchers to incubate and sustain leading-edge techniques in data mining. NCSA has access to staff with diverse expertise in areas critical to the development and support of projects such as ours, from visualization experts to researchers in applied math. The strong collaboration of the NCSA team is a critical component of the success of Human Voices.

The team at the University of South Carolina, led by Randall Cream, consists of a diverse set of expertise and a significant private-public partnership that represents a substantial investment in the Human Voices project. The Center for Digital Humanities at South Carolina is committed to sustaining and developing useful tools in the humanities; Human Voices represents an exciting opportunity to collaboratively coordinate efforts that exceed the scope and scale of research currently possible at the CDH. Randall Cream and Jun Zhou have combined on a wide array of projects in their brief time together—from a Humanities Gaming Institute and software to support digital editions to building useful tools to enable textual scholarship. Their work on Human Voices reflects a continuation of their ongoing interest in data mining as a stage in modeling cognitive behavior. The CDH's work with the Spenser Archive, led by David Miller, provides a mutually beneficial relationship for both the archive and the data mining initiative. Duncan Buell and Manton Matthews lead the Computer Science Engineering program at the U of SC, with current research interests in data mining, pattern recognition, and information retrieval. These CS researchers also bring an external partnership into the team with the software firm Collexis. Collexis builds software that allows predictive and anticipatory results building in the biomedical sciences, generating recursion effortlessly as a component linking inquiry, writing, and researching. We're proud of a strong and beneficial partnership that offers us a proven platform to model our results through, and the chance to learn from one of the leading vendors in the marketplace.

The team at the University of California-Santa Barbara, led by Patricia Fumerton, brings both an amazingly diverse and valuable archive, EBBA, and demonstrated experience in successfully data mining humanities texts. Eric Nebeker and Kris McAbee bring project management experience and the ability to develop interfaces for digital work in the humanities that humanities scholars actually use. Their EBBA project, multiply funded through the NEH, routinely wins awards and generates scholarship interrogating the space for digital inquiry in traditional humanities disciplines. Another key participant in the UCSB team is Carl Stahmer, a researcher in data mining in the humanities whose work includes the very successful NORA project. The UCSB has considerable experience successfully navigating the difficulties of large-scale distributed projects, and brings that awareness to Human Voices.

The team at McGill University, led by Paul Yachnin, brings a diverse set of experiences in yielding revolutionary results using digital methods in the humanities. The McGill researchers belong to the SSHRC MCRI 'Making Publics' (MaPs), an international research project examining the formation of informal groups of association and collaboration in Early Modern Europe. As part of its mandate MaPs has undertaken exploration of large-scale international collaboration within the humanities around a given research thematic. Its partners bring considerable experience to bear in regards not only to conducting humanities collaboration, but also the study of ideas and the formation of groups around works of art, literature, science, and academia. Rather than bringing an archive of scholarship to the project, our partners offer their experience in building and designing interactive recursive environments for scholarship. Matthew Milner has led the move as part of his interest in social networking models for humanities research and the federation of thematic-focused archival resources. Renee Sieber, working in

Geography, the School of the Environment, and Computer Science, has led several projects that demarcate the line between the social sciences and the humanities, applying computational methods to humanities interpretations to answer urgent questions about affiliation and space. Jean Guy Meunier, from Universite du Quebec a Montreal, has engaged in a variety of semantic approaches to text mining, using artificial intelligence to build fuzzy recursive systems that self-refine. These core researchers, deeply experienced on multi-site collaboration with the MaPs project, also bring a noteworthy external collaboration into the project. Vlad Petric, a software engineer at Google, collaborates with the McGill team on search design, information extraction, and using metadata to build interactive user applications.

These four teams are separately skilled with each having a unique ability to combine computational methods in the humanities with original humanistic inquiry. Together, the four teams represent a concentration of excellence that far exceeds the scope of the Digging into Data project. By building this impressive team, we hope to demonstrate our commitment to delivering results within the timeframe of the grant (January 2010 to March 2011) and beginning a project that grows beyond the contours of the initial investment.

Project Scope & Duration

The large scale of the Human Voices project necessitates a distributed project that coordinates multiple simultaneous tasks in order to allow efficiency and collaboration to quickly develop amongst the partners. The aims of the Human Voices project, to deliver a flexible, modular data mining application that scholars working in the humanities will actually want to use, require coordination on several simultaneous goals. We envision our project spanning the entire grant period, January 2010 to March 2011.

First of all, the project must write the code that performs the analytic routines in Phase I (citation identification), Phase II (citation unit segmentation by associating contextual sentences with the citational relationships), and Phase III (aggregation and association of citational units to generate concept maps and facilitate semantic analysis of disciplinary concepts). This sort of process is inherently recursive—it uses the output of one stage as an input for another stage of analysis that passes over the same data, creating a workflow that is largely sequential. The UCSB team will take the lead in adapting and porting the open-source code generated for the NORA and Monk projects to our more specific purposes. With Carl Stahmer's firsthand experience on the NORA project, we feel confident that the UCSB team can adapt the material to new purposes in a timely fashion. The SEASR team at the NCSA/UIUC will work to ensure that the SEASR platform can reliably sustain the computational load of the data mining proposed by Human Voices, supporting the growth and development of the project in the collaborative environment of SEASR. The team at SC will write new modules for SEASR, using a variety of languages including python, Java, UIMA, and other frameworks, refining and developing computational methods for citation identification, extraction, and disambiguation. The team at McGill will develop an interactive and visualization layer for each phase of the work, allowing scholars to interact with the processes and visualize the data to build textual maps. Loretta Auvil at NCSA, Randall Cream at SC, and Kris McAbee at UCSB will work to coordinate visualization and interaction layers with the underlying data mining.

Given the dependencies involved, we recognize that the first phase of our project (automated citation identification and extraction) constitutes a bottleneck of sorts on the project. We anticipate this crucial stage of code development to last about ten months for each team, with the application that provides the visualization and interaction layers trailing by approximately three months. To create efficiencies, we'll work through the well-structured early modern archives that already contain citation tags, essentially seeding our environment with a few dense relationships. This dual-stage work over the first phase should allow project staff the time to recursively refine the algorithms to address the peculiarities of the JSTOR data model efficiently. On the other hand, the second phase, segmenting the texts using this extracted citation data, is far more transformative in terms of human-computer interaction, but much easier

computationally. During the remaining five months of the grant period, we anticipate completing the second stage or making enormous headway to completion. To make sure the collaboration stays current over the course of the grant period, Human Voices project will meet twice during the fifteen months, facilitating a means of interaction that can help constitute an effective international partnership that can survive the vagaries of a 15 month project.

From our timelines, the ambition of our third phase—semantic analysis and concept mapping between texts—becomes apparent. We anticipate the work of Human Voices project to continue well beyond the fifteen month frame. Specifically, the semantic analysis and concept mapping modules are computationally much more intensive than analysis of citations and segmentation along instances of quotation. Our project teams will likely require the entire grant period just to develop the first two phases of the project, with the associated visualization and interaction layers that make such tasks worthwhile in the humanities. Rather than concentrating on performing as much data mining as possible within the grant window, we'll work to ensure that mature tools are disseminated, usability is built into these tools, and the project works from the outset with an eye toward structuring and maintaining a collaborative environment for research.

Our teams bring disparate goals into collaboration in order to tie us together in an effective partnership. Some teams are interested in maturing a software model that has already proven itself competent and successful; others, such as JSTOR, are interested in delivering innovative new tools that have persuasive utility to analyze culturally significant material. In essence, this DiD challenge will serve as a startup incentive to form a stable coalition, balancing a host of concerns against an array of possibilities. We anticipate a long-term collaboration, and have built in two public-private relationships (Vlad Petric, software engineer at Google, and Stephen Leicht, COO at Collexis) that hold potential for our project after the grant period is complete. Human Voices will seek a variety of funding models to sustain its effort after the grant period, including internal funds, entrepreneurial partnerships, and government sponsored funding opportunities.

Anticipated Outcomes

The Human Voices project undertakes three primary outcomes from its fifteen month grant period. First of all, the Human Voices team seeks to develop and sustain a collaborative environment for data mining as a human-centered act, an act which develops out of the participatory and interconnected nature of human culture. Our project invests heavily in the collaborative environment of SEASR, developing interactive components to ensure it remains relevant for humanities researchers. Secondly, we want to return significant value to the archives which contribute their data to the challenge, by developing secure, mature, and distributable tools that maximize the utility and significance of the humanities data within their archives. And thirdly, our project conducts innovative and worthwhile research into data mining in the humanities, creating a model for automated citation identification and extraction that should assist researchers in a variety of fields and disciplines.

Publicity is an important element of attending to the responsibilities of the public trust reflected in competitions such as Digging into Data, and Human Voices will actively work to ensure that there is a large impact of the substantial governmental investment in our project. With such a large and distributed project, word of mouth and social networking at conferences and discipline specific events is an effective strategy for disseminating project results. Of course, our project will invest in building a white paper for the results of our data mining research, and heartily participate in the Digging into Data conference to discuss approaches to large scale data in the humanities.

The real impact, though, of Human Voices will be a sustained collaborative partnership that crosses national borders and disciplinary barriers to construct a focused and sustained team for data mining in the

humanities. We've worked diligently to recruit private partners to share in this public investment, suggesting that our project enjoys a relevance not only to university researchers but also to archivists, information workers, and anyone interested in mechanisms which bring the massive data from the past into useful focus for individuals far removed in space and time from the creation of that data.

Budget Overview

Human Voices consists of four teams of researchers across North America (McGill, UCSB, NCSA, U of SC) and two significant private partnerships with leading-edge companies in information science (Google and Collexis). Our fifteen month project invests almost the entirety of its funds—100,000 dollars from each of SSHRC, NEH, and NSF—in labor, the limiting element in most projects in digital humanities. Our approach to the Digging into Data challenge is to treat this competition as a lengthy start-up project, where public monies entice an investment by private corporations and experienced researchers. Given the scales of the problems we tackle, and the dimensions of the data we have chosen to work with, we do not enter into such a partnership lightly. Instead, our commitment to this endeavour is reflected in the fact that each of the four teams commits its leaders almost entirely pro bono, choosing instead to fund graduate students, young researchers, and junior career personnel. We have joined together to form a working group with an understanding that the best results may resist us for much of the grant period. An investment in Human Voices is an investment in the persistent forms of collaboration that facilitate research programs such as our own.